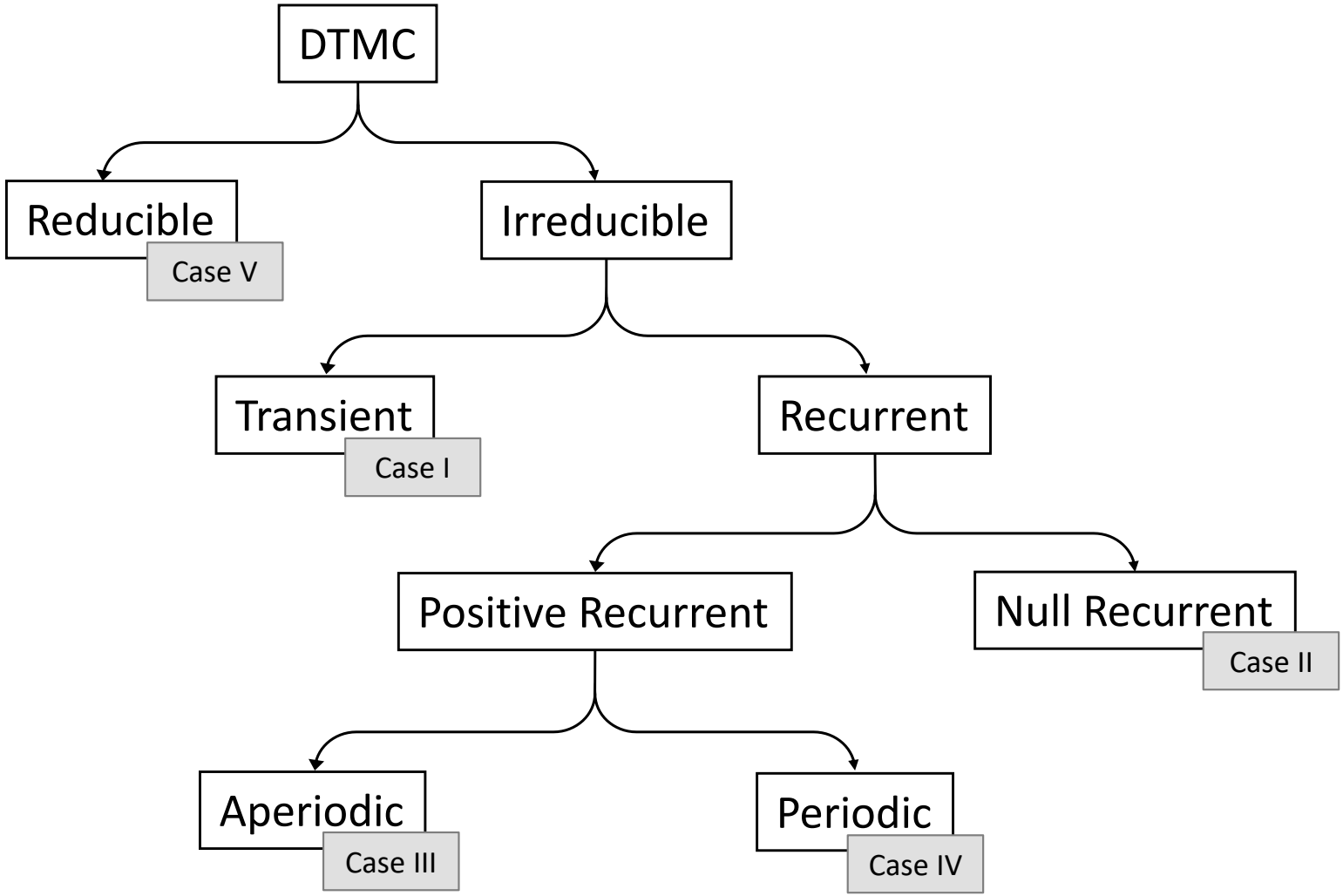
CE 273
Markov Decision Processes

Lecture 14

Optimality Conditions & Classification of Average Cost MDPs

Previously on Markov Decision Processes



Previously on Markov Decision Processes

From the examples, we can see that

- $\lim_{n\to\infty} P^{(n)}$ doesn't always exist
- $\lim_{n\to\infty} \frac{M^{(n)}}{n+1}$ however always exists and equals $\lim_{n\to\infty} P^{(n)}$ when the later exists. (Why is this intuitively true?)

Case	$\lim_{n\to\infty} P^{(n)}$	$\lim_{n\to\infty} \frac{M^{(n)}}{n+1}$	Identical Rows	Row Sum = 1
I	✓	✓	✓	X
II	✓	✓	✓	X
III	✓	✓	✓	✓
IV	X	✓	✓	✓
V	✓	✓	X	✓

Previously on Markov Decision Processes

Now consider the cases where P_μ^* is stochastic (Cases III, IV, V).

Recall from the analysis of total cost MDPs, $\sum_{k=0}^{N-1} P_\mu^k g_\mu$ represents the cost accumulated after N stages. (We started with the zero cost vector and used the T_μ operator.)

Thus, the average cost of policy μ is

$$J_\mu = \lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P_\mu^k g_\mu = \left(\lim_{N\to\infty} \frac{1}{N} \sum_{k=0}^{N-1} P_\mu^k \right) g_\mu = P_\mu^* g_\mu$$

Definition

The gain J_μ of a policy μ is defined as

$$J_\mu = P_\mu^* g_\mu$$

Previously on Markov Decision Processes

Definition

The bias h_μ of a policy μ is defined as

$$h_\mu = H_\mu g_\mu$$

where $H_\mu = (I - P_\mu + P_\mu^*)^{-1} - P_\mu^*$ and is called the fundamental matrix.

In addition, suppose the associated Markov chain is aperiodic, i.e., if $P_\mu^* = \lim_{N\to\infty} P_\mu^N$ (Case III), then we can interpret h_μ as

$$h_\mu = \lim_{N\to\infty} \sum_{k=0}^{N} P_\mu^k (g_\mu - J_\mu)$$

a relative cost vector, i.e., the difference of the total cost of μ and the total cost if one-stage costs were set to J_μ.

Previously on Markov Decision Processes

Theorem

For any transition matrix P and $\alpha \in (0,1)$,

$$(I - \alpha P)^{-1} = (1-\alpha)^{-1} P^* + H + O(|1-\alpha|)$$

where $O(|1-\alpha|)$ is an α-dependent matrix such that $\lim_{\alpha \to 1} O(|1-\alpha|) = 0$ and P^* and H are given by

$$P^* = \lim_{N \to \infty} \frac{1}{N} \sum_{k=0}^{N-1} P^k$$

$$H = (I - P + P^*)^{-1} - P^*$$

Previously on Markov Decision Processes

Recall that $J_\mu = P_\mu^* g_\mu$ and $h_\mu = H_\mu g_\mu$. Multiplying both sides of the Laurent series expansion with g_μ,

Theorem (Laurent Series Expansion)

For a given stationary policy μ with transition matrix P_μ and $\alpha \in (0,1)$,

$$J_{\alpha,\mu} = (1-\alpha)^{-1} J_\mu + h_\mu + O(|1-\alpha|)$$

where $O(|1-\alpha|)$ is an α-dependent matrix such that $\lim_{\alpha \to 1} O(|1-\alpha|) = 0$ and J_μ an h_μ represent gain and bias of the policy μ respectively.

Hence, we can write

$$J_\mu = (1-\alpha) J_{\alpha,\mu} - (1-\alpha) h_\mu + O(|1-\alpha|^2)$$

Thus, we expect that a policy minimizing $J_{\alpha,\mu}$ for α close to 1 will also minimize the average cost J_μ!

Lecture Outline

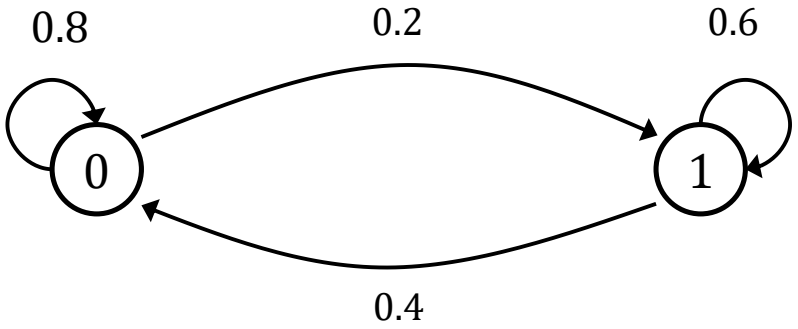
1. Blackwell Optimality
2. Optimality Equations
3. Unichain and Multichain MDPs

Blackwell Optimality

Blackwell Optimality

Gain and Bias Revisited

Suppose a policy μ induces the following Markov chain on a two-state MDP.



$$P_\mu = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \quad g_\mu = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Find the gain-bias pair (J_μ, h_μ) for the above policy.

$$P_\mu^* = \begin{bmatrix} 2/3 & 1/3 \\ 2/3 & 1/3 \end{bmatrix} \quad H_\mu = \begin{bmatrix} 0.55 & -0.55 \\ -1.11 & 1.11 \end{bmatrix}$$

Thus, the gain and bias are

$$J_\mu = \begin{bmatrix} 4/3 \\ 4/3 \end{bmatrix} \quad h_\mu = \begin{bmatrix} -0.55 \\ 1.11 \end{bmatrix}$$

Blackwell Optimality

Gain and Bias Revisited

Proposition

For any transition matrix P and fundamental matrix H

$$P^* = PP^* = P^*P = P^*P^*$$
$$P^*H = HP^* = 0$$
$$P^* + H = I + PH$$

Proof.

Proving the first equality is not difficult but a bit involved and we will skip it in the interest of time.

From the definition of the fundamental matrix,

$$\begin{aligned} H &= (I - P + P^*)^{-1} - P^* \\ \Rightarrow (I - P + P^*)H &= I - (I - P + P^*)P^* \\ \Rightarrow (I - P + P^*)H &= I - P^* \\ \Rightarrow P^*(I - P + P^*)H &= P^*(I - P^*) \\ \Rightarrow P^*H &= 0 \end{aligned}$$

Proof.

Proof of $HP^* = 0$ is similar.

From one of the above equations,

$$\begin{aligned}
&(I - P + P^*)H = I - P^* \\
\Rightarrow &H - PH = I - P^* \\
\Rightarrow &P^* + H = I + PH
\end{aligned}$$

■

Blackwell Optimality

Gain and Bias Revisited

Using the above proposition, and since $J_\mu = P_\mu^* g_\mu$ and $h_\mu = H_\mu g_\mu$,

Proposition (Policy Evaluation)

The gain and bias vectors of a policy μ, J_μ and h_μ satisfy,

$$J_\mu = P_\mu J_\mu$$
$$J_\mu + h_\mu = g_\mu + P_\mu h_\mu$$

Proof.

From the previous proposition, $P_\mu^* = P_\mu P_\mu^*$. Right-multiplying both sides by g_μ, $J_\mu = P_\mu J_\mu$.

Also, from the previous proposition, $P_\mu^* + H_\mu = I + P_\mu H_\mu$. Again, right-multiplying both sides by g_μ,

$$J_\mu + h_\mu = g_\mu + P_\mu h_\mu$$

■

Blackwell Optimality

Gain and Bias Revisited

In the last class, assuming aperiodicity, we interpreted the bias as the relative cost or the difference in the total cost of μ and the total cost if the one-stage costs were J_μ. Let's see why.

First, since $P_\mu^* H_\mu = 0$, $P_\mu^* H_\mu g_\mu = 0 \Rightarrow P_\mu^* h_\mu = 0$. Second, using a set of equations from the previous proposition, $J_\mu + h_\mu = g_\mu + P_\mu h_\mu$,

$$\Rightarrow g_\mu - J_\mu = h_\mu - P_\mu h_\mu$$

$$\Rightarrow \sum_{k=0}^{N} P_\mu^k (g_\mu - J_\mu) = \sum_{k=0}^{N} P_\mu^k (h_\mu - P_\mu h_\mu) = h_\mu - P_\mu^N h_\mu$$

Taking limits on both sides as $N \to \infty$, $P_\mu^N h_\mu \to P_\mu^* h_\mu = 0$. (Why?) Therefore,

$$h_\mu = \lim_{N \to \infty} \sum_{k=0}^{N} P_\mu^k (g_\mu - J_\mu)$$

Blackwell Optimality

Gain and Bias Revisited

Unlike discounted and total cost MDPs, where we could solve a system of equations for a given policy (and use this in the policy iteration algorithm), we cannot simply solve

$$J = P_\mu J$$
$$J + h = g_\mu + P_\mu h$$

to get the average cost of policy μ. (Why?)

If (J_μ, h_μ) solves the above system, then $(J_\mu, h_\mu + \text{constant})$ also satisfies the above system. Hence, there are an infinite number of solutions. We will call these policy evaluation equations for easy referencing.

In general, it can be shown that all solutions to the above system are of the form $(J_\mu, h_\mu + d)$, where $d = P_\mu d$.

Blackwell Optimality

Definition and Existence

We saw earlier that the average cost of an MDP for a given policy can be expressed in terms of its discounted cost when the discount factor is close to 1.

$$J_\mu = \lim_{\alpha \to 1} (1 - \alpha) J_{\alpha,\mu}$$

Thus, if we can find an optimal policy that solves the α-discounted problem for $\alpha \approx 1$, we expect it to be optimal to the average cost problem.

But what if one policy is optimal for $\alpha = 0.99$ and a different one is optimal for $\alpha = 0.9999$? How close to 1 should we go?

Definition

A stationary policy μ is said to be Blackwell optimal if it is optimal for all α-discounted problems with $\alpha \in (\bar{\alpha}, 1)$, where $0 < \bar{\alpha} < 1$

Blackwell Optimality

Definition and Existence

Proposition

A Blackwell optimal policy always exists.

A Blackwell optimal policy is optimal to the average cost problem when we restrict our attention to stationary policies. It also happens to be optimal over all non-stationary policies as well!

Optimality Equations

Optimality Equations

Properties of Blackwell Optimal Policies

Proposition

1. *All Blackwell optimal policies have the same gain and bias*
2. *Let (J^*, h^*) be the gain-bias pair of a Blackwell optimal policy, then*

$$J^*(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) J^*(j) \, \forall \, i = 1, \ldots, n$$

Let $\bar{U}(i)$ be the set of controls that attain the minimum in the above equation.

$$J^*(i) + h^*(i) = \min_{u \in \bar{U}(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h^*(j) \right\} \forall \, i = 1, \ldots, n$$

If μ^* is Blackwell optimal, it attains the minimum in the RHS of the above two equations.

Optimality Equations

Properties of Blackwell Optimal Policies

Proof.

Proof of 1: Suppose μ and μ' are Blackwell optimal. Then, $J_{\mu,\alpha} = J_{\mu',\alpha}$ for all $\max\{\bar{\alpha}, \bar{\alpha}'\} < \alpha < 1$.

Since, $J_\mu = \lim_{\alpha\to 1}(1-\alpha)J_{\alpha,\mu}$ and $J_{\mu'} = \lim_{\alpha\to 1}(1-\alpha)J_{\alpha,\mu'}$, it follows that $J_\mu = J_{\mu'}$

From the Laurent series expansion, of $J_{\alpha,\mu}$ and $J_{\alpha,\mu'}$, setting $\alpha \to 1$ again, we get $h_\mu = h_{\mu'}$.

Optimality Equations

Properties of Blackwell Optimal Policies

Proof.

Proof of 2: Let μ^* be a Blackwell optimal policy. For every stationary policy μ and $\alpha \in (\bar{\alpha}, 1)$,

$$TJ_{\alpha,\mu^*} \leq T_\mu J_{\alpha,\mu^*}$$
$$g_{\mu^*} + \alpha P_{\mu^*} J_{\alpha,\mu^*} \leq g_\mu + \alpha P_\mu J_{\alpha,\mu^*}$$

Using the Laurent series expansion of J_{α,μ^*} and the above inequality,

$$0 \leq g_\mu - g_{\mu^*} + \alpha(P_\mu - P_{\mu^*})((1-\alpha)^{-1}J^* + h^* + O(|1-\alpha|))$$
$$0 \leq (1-\alpha)(g_\mu - g_{\mu^*}) + \alpha(P_\mu - P_{\mu^*})(J^* + (1-\alpha)h^* + O((1-\alpha)^2))$$

Taking limits as $\alpha \to 1$, $P_{\mu^*}J^* \leq P_\mu J^* \Rightarrow J^* \leq P_\mu J^*$. Hence, we can write

$$J^*(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) J^*(j) \, \forall \, i = 1, \ldots, n$$

Optimality Equations

Properties of Blackwell Optimal Policies

Proof.

Select any μ such that $P_{\mu^*} J^* = P_\mu J^* \Rightarrow J^* = P_\mu J^*$, i.e., we are looking at a policy comprising of controls that attain the minimum in the above expression. The earlier inequality can thus be written as

$$0 \leq g_\mu - g_{\mu^*} + \alpha(P_\mu - P_{\mu^*})(h^* + O(|1-\alpha|))$$

Taking limits as $\alpha \to 1$,

$$g_{\mu^*} + P_{\mu^*} h^* \leq g_\mu + P_\mu h^*$$

Thus, μ^* minimizes $g_\mu + P_\mu h^*$ over all μ which satisfy $J^* = P_\mu J^*$. (This looks a lot like $Th^* \leq T_\mu h^*$.) From the policy evaluation equations,

$$J^* + h^* = g_{\mu^*} + P_{\mu^*} h^*$$

Therefore, $J^* + h^* \leq g_\mu + P_\mu h^*$. In other words,

$$J^*(i) + h^*(i) = \min_{u \in \bar{U}(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h^*(j) \right\} \forall\, i = 1, \ldots, n$$

■

Optimality Equations

Sufficient Conditions

The earlier proposition and discussion established that a Blackwell optimal policy is optimal to the average cost problem.

Further, optimal policies were found to satisfy some equations which are the necessary conditions for optimality. It can also be shown that they are sufficient.

Proposition

If J' and h' satisfy the following pair of optimality equations

$$J(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) J(j) \, \forall \, i = 1, \ldots, n$$

$$J(i) + h(i) = \min_{u \in \bar{U}(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h(j) \right\} \forall \, i = 1, \ldots, n$$

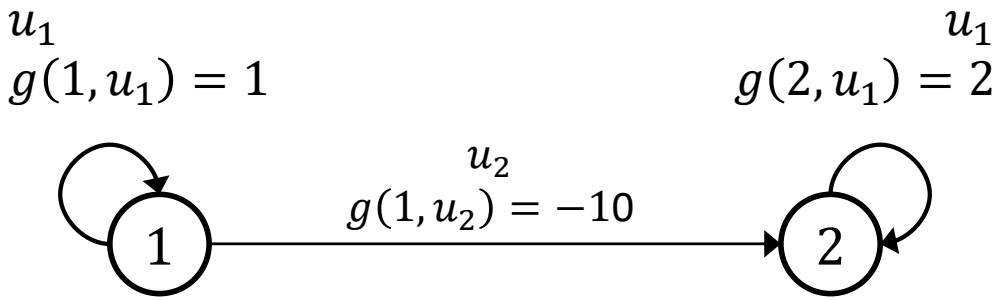
where $\bar{U}(i)$ is the set of controls that attain the minimum in the above equation. Then, $J' = J^*$ is the optimal average cost vector.

Further, if a stationary policy μ attains the minimum in the above equations, then it is the optimal policy μ^*.

Optimality Equations

Example

Consider the following two-state MDP. Find the optimal average cost.



- Guess the optimal solution
- Find the gain and bias of the optimal policy
- Check the necessary conditions
- Solve the sufficient conditions

Lecture Outline

Unichain and Multichain MDPs

Unichain and Multichain MDPs

Equal Costs

These optimality equations are analogous to the Bellman equations for discounted MDPs but solving it is a two-stage problem. It holds true irrespective of the structure of the Markov chains for different policies.

However, for instances in which the underlying Markov chains has a certain structure, the optimal average cost is equal for all states! This was true in the opening example.

We will discuss conditions required for this property soon but let us study the consequence of having equal costs. What happens to

$$J^*(i) = \min_{u \in U(i)} \sum_{j=1}^{n} p_{ij}(u) J^*(j) \, \forall \, i = 1, \ldots, n$$

These equations are superfluous. Every $u \in U(i)$ satisfies it and hence $\bar{U}(i) = U(i)$. Denoting $J^*(i) = \lambda$, we can thus write the second set of equations as

$$\lambda^* + h^*(i) = \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h^*(j) \right\} \forall \, i = 1, \ldots, n$$

We will use the following shorthand notation for the T and T_μ operator, but will apply them on the bias h instead of J.

Formally, define

$$(Th)(i) = \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h(j) \right\} \forall\, i = 1, \ldots, n$$

For a stationary policy μ, define

$$(T_\mu h)(i) = g(i, \mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i)) h(j) \,\forall\, i = 1, \ldots, n$$

Unichain and Multichain MDPs

Optimality Conditions with Equal Costs

In summary, if the average cost is independent of the initial state, the following proposition is true

Proposition

If a scalar λ and a vector h satisfy

$$\lambda + h(i) = \min_{u \in U(i)} \left\{ g(i,u) + \sum_{j=1}^{n} p_{ij}(u) h(j) \right\} \ \forall\, i = 1, \ldots n$$

then λ is the optimal average cost $J^*(i)$ for all i, i.e.,

$$\lambda = \min_{\mu} J_\mu(i) = J^*(i) \ \forall \ i = 1, \ldots, n$$

Further, if μ^* attains the minimum in the first expression, then $J_{\mu^*}(i) = \lambda \, \forall \, i$.

In shorthand, the first equation can be rewritten as $\lambda e + h = Th$. Think of this as being analogous to $J^* = TJ^*$ in the discounted world.

Unichain and Multichain MDPs

Evaluating a Policy with Equal Costs

Likewise, we can also evaluate the cost of a stationary policy which has equal average costs starting from any state using the following result

Proposition

Given a stationary policy μ, if a scalar λ_μ and a vector h satisfy

$$\lambda_\mu + h(i) = g(i, \mu(i)) + \sum_{j=1}^{n} p_{ij}(\mu(i))h(j) \;\; \forall\, i = 1, \ldots n$$

then $\lambda_\mu = J_\mu(i)$ for all i

Using the T notation, this takes the form $\lambda_\mu e + h = T_\mu h$. Think of this as being analogous to $J_\mu = T_\mu J_\mu$ in the discounted world.

Unichain and Multichain MDPs

Classification of MDPs

Consider the easy case in which **every** stationary policy has a single recurrent class.

Does the average cost of such policy μ have equal costs?

Now what if the state space can be divided into $C \cup \mathcal{C}$, where C is a recurrent class and $\mathcal{C}$ is the set of transient states for **every** policy?

MDPs which satisfy this property are called **Unichain MDPs** and the simplified optimality equations can be used in this case.

Unichain and Multichain MDPs

Classification of MDPs

MDPs in which at least one policy results in two or more closed communicating classes and a transient class (possibly empty) are called **Multichain MDPs**.

The equal costs property does not hold in this case, but it still holds within each closed communicating class.

Your Moment of Zen