Solving Growth Rates and Average Optimality in Risk-Sensitive Markov Decision Chains

Karel Sladký

Institute of Information Theory and Automation
of the Academy of Sciences of the Czech Republic

Congreso de Matematica Aplicada in Valle de Toluca, 4.-6.11.2009

Outline

1. Motivation and Objective
2. Notations and Preliminaries
   - 2.1. Discrete-Time Markov Chains
   - 2.2. Continuous-Time Markov Chains
3. Risk-Sensitive Optimality and Nonnegative Matrices
   - 3.1. Discrete-Time Case
   - 3.2. Continuous-Time Case
4. Products of Nonnegative Matrices
5. Asymptotic Behavior of Expected Utilities
   - 5.1. Discrete-Time Case
   - 5.2. Continuous-Time Case
6. Value and Policy Iteration Algorithms

1. Motivation and Objective

The usual optimization criteria examined in the literature on stochastic dynamic programming, such as

a total discounted or
mean (average) reward (resp. cost) structures,

may be quite insufficient to characterize the problem from the point of a decision maker.
To this end it may be preferable if not necessary to select more sophisticated criteria that

also reflect the variability-risk
features of the problem.

Perhaps the best known approaches stem from the classical work of Markowitz on mean variance selection rules oriented primarily on the portfolio selection problem.

On this other hand risky decisions can be also eliminated using exponential utility functions.

In this talk we focus attention on the so-called

risk-sensitive optimality criteria,

i.e., when expectation of the stream of rewards (or of costs) generated by Markov chain is evaluated by an

exponential utility function.

The topic of risk-sensitive optimality criteria in Markov decision processes initiated in 1972 in the seminal paper by Howard and Matheson (Manag. Sci., vol. 23, pp. 356–369) and followed by many researchers, e.g.

Jaquette Manag. Sci. (1976)
Hernández-Hernández, Marcus SCL (1996)
Bielecki, Hernández-Hernández, Pliska MMOR (1999)
Borkar, Meyn MOR (2002)
Cavazos-Cadena MMOR (2002), (2003), SCL (2008)
Cavazos-Cadena, Fernández-Gaucherand MMOR (1999), IEEE AC (2000)
Cavazos-Cadena, Montes de Oca MOR (2003), JAP (2005)
Cavazos-Cadena, Hernández-Hernández AAP (2005), SCL (2009)
Sladký Kybernetika (2008)

In the most of above papers the analysis was restricted to Markov processes with a single class of recurrent states and no transient states. Moreover, the analysis was restricted only to discrete-time Markov decision chain, only little attention, if any was devoted to the continuous-time Markov chains.

The aim of this talk is to extend the analysis to the case of general case of unichain models, i.e. Markov chains with a single class of recurrent states and non-empty set of transient states and indicate how the method can be even extended to reducible (multichain) Markov processes both in discrete- and continuous-time setting.

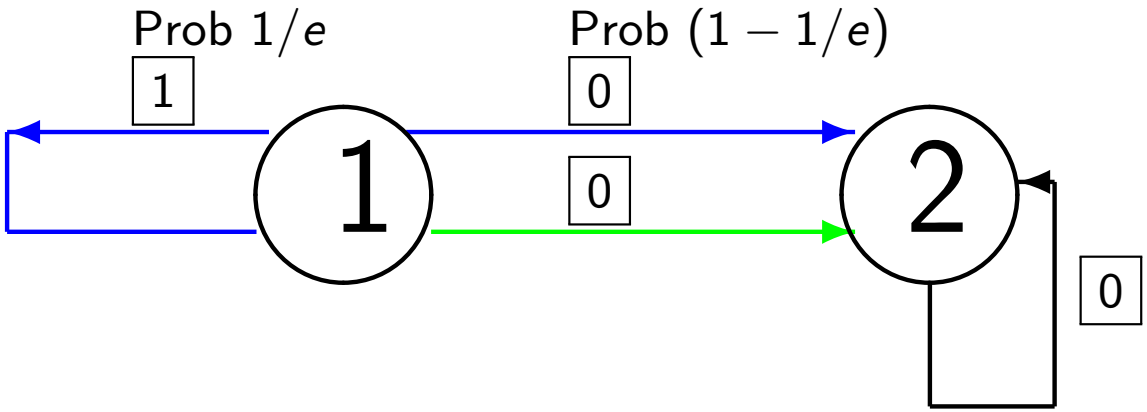
Our analysis of the risk-sensitive optimality in Markov decision chains can be based on more complicated models of stochastic dynamic programming where in discrete-time models transition probability matrices are replaced by general nonnegative matrices and in continuous-time case transition rate matrices are replaced by matrices with nonnegative off-diagonal entries.

Motivational Example 1

Consider Markov decision chain with two states 1,2, possible actions 1,2 only in state 1 and the following transition and reward structure:

$$p_{11}(1) = \frac{1}{e}, \quad p_{12}(1) = 1 - \frac{1}{e}, \quad p_{11}(2) = 0, \quad p_{12}(2) = 1$$
$$p_{22} = 1, \quad r_{11} = 1, \quad r_{12} = r_{22} = 0$$

depicted in the following diagram:



Obviously, regardless the selected action in state 1, state 1 is transient, state 2 is absorbing, i.e. we consider unichain model.

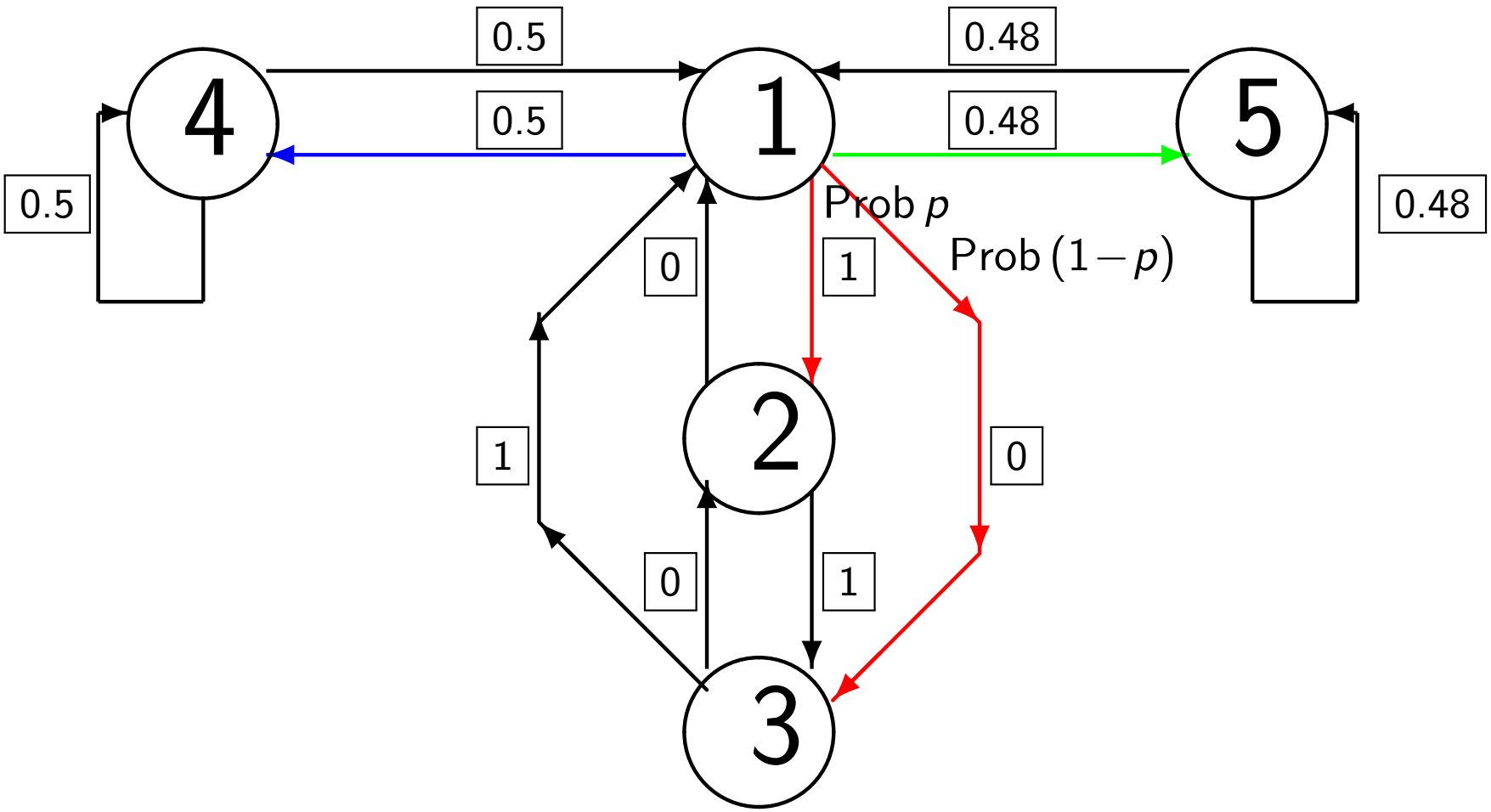
If the process starts in state 1 and action 1 (blue line) is taken, total reward received after n transitions is equal to $\sum_{k=1}^{n} \left(\frac{1}{e}\right)^k = \frac{e^2}{e-1}(1 - \frac{1}{e^n})$; hence is uniformly bounded by $\frac{e^2}{e-1}$; if action 2 (green line) is taken or if the chain starts in state 2 total reward equals 0. Hence in both cases long run mean reward is equal to zero.

On the other hand if the stream of received rewards is evaluated by an exponential utility function $e^{\gamma x}$ with risk sensitive coefficient $\gamma = 1$ then if the chain starts in state 1 and action 1 is followed then the expected value of utility function assigned to total reward received in the n following transitions is equal to n and the corresponding long run mean value equals 1. On the contrary if action 2 is followed or if the chain starts in state 2 total reward equals 1 and the corresponding long run mean value equals 0.

Similar example can be easily produced also for continuous-time case.

Motivational Example 2

Consider the depicted Markov reward chain with 5 states and only three possible actions in state 1 (in the remaining states no option is possible); transition rewards are included.



Observe that if the process starts in state 1 and the blue decision is selected we obtain a constant sequence of one-stage rewards and after k transitions the total reward $R_1(k) = 0.5\,k$.

Similarly, if in state 1 the green decision is selected we again obtain a constant sequence of one-stage rewards and the total reward $R_1(k) = 0.48\,k$.

On the contrary if in state 1 the red decision is taken the chain visits only the states $1, 2, 3$ and the sequence of received rewards obeys a binomial distribution with parameter p. In particular of $p = 0.5$ the total reward $R_1(k) = 0.5\,k$, however its variance $V_1(k) = k\,0.5\,(1 - 0.5) = 0.25\,k$.

Question: Should we prefer constant increase of the reward to a time-varying reward with a little higher average value?

We show how the above mentioned generalization of stochastic dynamic programming can help for better understanding of risk-sensitive optimality in Markov decision processes and extensions of above mentioned results, in particular, to unichain models with transient state and multichain Markov processes.

Recall that in the Howard's and Matheson's seminal paper the underlying Markov chain is assumed to be irreducible and aperiodic. Then on multiplying transition probabilities by nonnegative numbers the resulting matrix is again irreducible and its Perron eigenvector is strictly positive.

We show that existence of strictly positive right Perron eigenvector of the resulting matrix (not necessarily irreducible if the underlying Markov chain has also transient states or more recurrent classes) guarantees similar behavior as for the irreducible and aperiodic case. Moreover, analogous algorithmic procedures of value and policy iteration types for finding growth rate of expected utilities and the corresponding certainty equivalents will be provided.

2. Notation and Preliminaries

We shall consider discrete- and continuous-time Markov decision chains

$X^{\mathrm{d}} = \{X_n,\ n = 0, 1, \ldots\}$ and $X^{\mathrm{c}} = \{X(t), t \geq 0\}$

with finite state spaces

$\mathcal{I} = \{1, 2, \ldots, N\}$ and a finite set

$\mathcal{F}_i = \{1, 2, \ldots, K_i\}$ of actions in state $i \in \mathcal{I}$.

Supposing that in state $i \in \mathcal{I}$ action $k \in \mathcal{F}_i$ is selected, then
in discrete-time case
state j is reached in the next transition with a given probability p_{ij}^k and one-stage transition reward r_{ij} (or transition cost c_{ij}) is accrued,
in continuous-time case
state j is reached with a given transition rate $q(i, j|k)$, reward rate $r(i)$ (or cost rate $c(i)$) has been obtained in state i and transition reward $r(i, j)$ (or transition cost $c(i, j)$) is accrued.

We suppose that the stream of generated rewards or costs is evaluated by an exponential utility function, say $u^\gamma(\cdot)$, i.e. a utility function with constant risk sensitivity $\gamma \in \mathbb{R}$. Then the utility assigned to the (random) reward ξ is given by

$$u^\gamma(\xi) := \begin{cases} \operatorname{sign}(\gamma)\, \exp(\gamma\xi), & \text{if } \gamma \neq 0 \\ \xi & \text{for } \gamma = 0. \end{cases} \tag{1}$$

If ξ is a (bounded) random variable then for the corresponding certainty equivalent of the (random) variable ξ, say $Z^\gamma(\xi)$, since

$$u^\gamma(Z^\gamma(\xi)) = \mathbf{E}[\operatorname{sign}(\gamma)\, \exp(\gamma\xi)]$$

($\mathbf{E}$ is reserved for expectation), we immediately get

$$Z^\gamma(\xi) = \begin{cases} \frac{1}{\gamma}\, \ln\{\mathbf{E}\,[\exp(\gamma\xi)]\}, & \text{if } \gamma \neq 0 \\ \mathbf{E}[\xi] & \text{for } \gamma = 0. \end{cases} \tag{2}$$

A (Markovian) policy, say π, controlling the chain is a rule how to select actions in each state.
For the *discrete-time models* policy $\pi = (f^0, f^1, \ldots)$ where for $n = 0, 1, 2, \ldots$ $f^n \in \mathcal{F} \equiv \mathcal{F}_1 \times \ldots \times \mathcal{F}_N$ and $f_i^n \in \mathcal{F}_i$ is the decision at the nth transition when the chain X^{c} is in state i.

For the *continuous-time case* policy $\pi = f(t)$, is a piecewise constant, right continuous vector function where $f(t) \in \mathcal{F}$ and $f_i(t) \in \mathcal{F}_i$ is the decision (or action) taken at time t if the process $X(t)$ is in state i. Then for each π we can identify time points $0 < t_1 \ldots < t_i < \ldots$ at which the policy switches; we denote by $f^i \in \mathcal{F}$ the decision rule taken in the time interval $(t_{i-1}, t_i]$.

Policy π is stationary if it takes at all times the same decision rule, i.e. selects actions only with respect to the current state.
Stationary policy is fully identified by decision vector f selecting the transition probability matrix $\mathbf{P}(f)$ of X^{c} or transition rate matrix $\mathbf{Q}(f)$ with elements $q(j|i, f_i)$ of X^{d}, and hence by the Kolmogorov equation also the transition probability matrix $\mathbf{P}(f)^{\pi}(0, t)$ along with $\mathbf{p}^{\pi}(t)$ probability distribution at time t.

For the more detailed analysis it is required to consider the discrete- and continuous-time case separately.

Discrete-time Markov Chains

$\xi_n = \sum_{k=0}^{n-1} r_{X_k, X_{k+1}}$ .... stream of transition rewards

received in the n next transitions,

$\xi^{(m,n)}$ ... total (random) reward obtained from the mth up to the nth transition (obviously, $\xi_n = r_{X_0, X_1} + \xi^{(1,n)}$).

If $\gamma \neq 0$ then $u^\gamma(\xi_n) := \operatorname{sign}(\gamma)\, \mathrm{e}^{\gamma \sum_{k=0}^{n-1} r_{X_k, X_{k+1}}}$

is the (random) utility assigned to ξ_n, and

$Z^\gamma(\xi_n) = \frac{1}{\gamma} \ln \{\mathbf{E}[\mathrm{e}^{\gamma \sum_{k=0}^{n-1} r_{X_k, X_{k+1}}}]\}$ its certainty equivalent.

If $\gamma = 0$ then $u^\gamma(\xi_n) = \sum_{k=0}^{n-1} r_{X_k, X_{k+1}}$, and $Z^\gamma(\xi_n) = \mathbf{E}[\sum_{k=0}^{n-1} r_{X_k, X_{k+1}}]$.

Supposing that the chain starts in state $X_0 = i$ and policy $\pi = (f^n)$ is followed, then for the expected utility in the n next transitions and the corresponding certainty equivalent we have ($\mathbf{E}_i^\pi$ denotes expectation if policy π is followed and $X_0 = i$)

$$
\begin{aligned}
U_i^\pi(\gamma, 0, n) &:= \mathbf{E}_i^\pi[\exp(\gamma \sum_{k=0}^{n-1} r_{X_k, X_{k+1}})] \\
\bar{U}_i^\pi(\gamma, 0, n) &:= (\operatorname{sign} \gamma)\mathbf{E}_i^\pi[\exp(\gamma \sum_{k=0}^{n-1} r_{X_k, X_{k+1}})] \\
Z_i^\pi(\gamma, 0, n) &:= \frac{1}{\gamma} \ln \{\mathbf{E}_i^\pi[\exp(\gamma \sum_{k=0}^{n-1} r_{X_k, X_{k+1}})]\} \\
&= \frac{1}{\gamma} U_i^\pi(\gamma, 0, n).
\end{aligned}
$$

In what follows we shall often abbreviate

$$U_i^\pi(\gamma, 0, n) \quad \text{by} \quad U_i^\pi(\gamma, n), \quad \text{and}$$

$$Z_i^\pi(\gamma, 0, n) \quad \text{by} \quad Z_i^\pi(\gamma, n).$$

Moreover,

$\mathbf{U}^\pi(\gamma, n)$ is the vector of absolute values of expected utilities whose ith element equals $U_i^\pi(\gamma, n)$.

Similarly,

$\mathbf{Z}^\pi(\gamma, n)$ is reserved for the vector of certainty equivalents whose ith element equals $Z_i^\pi(\gamma, n)$, and

$J_i^\pi(\gamma) := \liminf_{n\to\infty} \frac{1}{n} Z_i^\pi(\gamma, n)$ is the mean value of $Z_i^\pi(\gamma, n)$ over time.

Continuous-time Markov Chains

Let for any piecewise constant policy $\pi = f(t)$

$$\xi^{\pi}_{X(0)}(t) = \int_0^t r(X(\tau))\mathrm{d}\tau + \sum_{k=0}^{N(t)} r(X(\tau^-), X(\tau^+))$$

be the total (random) reward obtained up to time t,
where $X(t)$ denotes the state at time t, $X(\tau^-)$, $X(\tau^+)$ is the state just prior and after the kth jump, and $N(t)$ is the number of jumps up to time t.
Similarly

$$\xi^{\pi}_{X(t')}(t', t) = \int_{t'}^t r(X(\tau))\mathrm{d}\tau + \sum_{k=N(t')}^{N(t)} r(X(\tau^-), X(\tau^+))$$

is the total (random) reward obtained in the time interval $[t', t)$.

Then

$$u^\gamma(\xi^\pi_{X(0)}(t)) \quad \text{and} \quad u^\gamma(\xi^\pi_{X(0)}(t)\xi^\pi_{X(t')}(t',t))$$

is the (random) value of the exponential utility assigned to $\xi^\pi_{X(0)}(t)$ and to $\xi^\pi_{X(t')}(t',t)$ respectively.
Moreover,

$$U^\pi_i(\gamma,t) := \mathrm{E}\{|u^\gamma(\xi^\pi_i(t))|\}, \quad U^\pi_i(\gamma,t',t) := \mathrm{E}\{|u^\gamma(\xi^\pi_i(t',t))|\}$$

is the absolute value of the expected utility assigned to $\xi^\pi_{X(0)}(t)$ for $X_0 = i$ and to $\xi^\pi_{X(t')}(t',t)$ for $X(t') = i$ respectively, and

$$Z^\pi_i(\gamma,t) := \frac{1}{\gamma}\, U^\pi_i(\gamma,t)$$

is the certainty equivalent corresponding to $U^\pi_i(\gamma,t)$.

For what follows it will be convenient to introduce more compact notations. To this end

$\mathbf{U}^{\pi}(\gamma, t)$ is the vector of absolute values of expected utilities whose ith element equals $U_i^{\pi}(\gamma, n)$.

Similarly,

$\mathbf{Z}^{\pi}(\gamma, t)$ is reserved for the vector of certainty equivalents whose ith element equals $Z_i^{\pi}(\gamma, t)$, and

$J_i^{\pi}(\gamma) := \liminf_{t \to \infty} \frac{1}{t} Z_i^{\pi}(\gamma, t)$ is the mean value of $Z_i^{\pi}(\gamma, t)$ over time.

3. Risk-Sensitive Optimality and Nonnegative Matrices

Discrete-time Markov Chains

Conditioning on X_1 we have

$u^\gamma(\xi_n) = \mathbf{E}[\mathrm{e}^{\gamma\, r_{X_0,X_1}} \cdot u^\gamma(\xi^{(1,n)})|X_1 = j]$.

Hence we immediately get for $\bar{p}_{ij}^{f_i} := p_{ij}^{f_i} \cdot \mathrm{e}^{\gamma r_{ij}}$

$$\begin{aligned} \bar{U}_i^\pi(\gamma, 0, n) &:= \operatorname{sign}(\gamma)\mathbf{E}_i^\pi[\exp(\gamma \sum_{k=0}^{n-1} r_{X_k,X_{k+1}})] \\ &= \sum_{j\in\mathcal{I}} \bar{p}_{ij}^{f_i^0}\ \bar{U}_j^\pi(\gamma, 1, n) \\ U_i^\pi(\gamma, 0, n) &= \sum_{j\in\mathcal{I}} \bar{p}_{ij}^{f_i^0}\ U_j^\pi(\gamma, 1, n) \end{aligned} \tag{3}$$

with $U_i^\pi(\gamma, n, n) = 1$.

In vector notation we can write

$$\mathbf{U}^{\pi}(\gamma, 0, n) = \bar{\mathbf{P}}^{(\gamma)}(f^0) \cdot \mathbf{U}^{\pi}(\gamma, 1, n) \tag{4}$$

where the ij-th entry of $\bar{\mathbf{P}}^{(\gamma)}(f)$ equals $\bar{p}_{ij}^{f_i} = p_{ij}^{f_i} \cdot \mathrm{e}^{\gamma r_{ij}}$
and $\mathbf{U}^{\pi}(\gamma, n, n) = \mathbf{e}$ (unit column vector).

Observe that the set of matrices $\{\bar{\mathbf{P}}^{(\gamma)}(f), f \in \mathcal{F}\}$
fulfils the "product property."

Iterating (4) we immediately get

Result 1. If policy $\pi = (f^n)$ is followed then

$$\mathbf{U}^{\pi}(\gamma, n) = \bar{\mathbf{P}}^{(\gamma)}(f^0) \cdot \bar{\mathbf{P}}^{(\gamma)}(f^1) \cdot \ldots \cdot \bar{\mathbf{P}}^{(\gamma)}(f^{n-1}) \cdot \mathbf{e}. \tag{5}$$

In particular, for $\gamma = 0$ we have

$$\mathbf{U}^{\pi}(0, n) = \mathbf{P}(f^0) \cdot \mathbf{P}(f^1) \cdot \ldots \cdot \mathbf{P}(f^{n-1}) \cdot \mathbf{e} = \mathbf{e}.$$

Continuous-time Markov Chains

Now we present continuous-time analog of (5) for the continuous-time Markov chain X^{c}.

Result 2. The expected utility $U_i^{\pi}(\gamma, t)$ for any $(i = 1, \ldots, N)$ and $t \in [0, t^*]$ where $U_i^{\pi}(\gamma, t^*) = 1$ fulfills the following set of differential equations

$$\begin{aligned} \frac{\mathrm{d}U_i^{\pi}(\gamma, t)}{\mathrm{d}t} &= [q_{ii}(f_i(t) + \gamma r(i)]U_i^{\pi}(\gamma, t) \\ &\quad + \sum_{j=1, j\neq i}^{N} q_{ij}(f_i(t))\mathrm{e}^{\gamma r(i)} \cdot U_j^{\pi}(\gamma, t) \end{aligned}$$

that can be also written in matrix form as

$$\frac{\mathrm{d}\mathbf{U}^{\pi}(\gamma, t)}{\mathrm{d}t} = \bar{\mathbf{Q}}^{(\gamma)}(f(t)) \cdot \mathbf{U}^{\pi}(\gamma, t), \quad \text{with} \quad \mathbf{U}^{\pi}(\gamma, t^*) = \mathbf{e} \qquad (6)$$

where $\bar{\mathbf{Q}}^{(\gamma)}(f) = [\bar{q}_{ij}^{\gamma}(f_i)]$ is an $N \times N$ matrix with nonnegative off-diagonal elements

$$\bar{q}_{ij}^{\gamma}(f_i) = \begin{cases} q_{ii}(f_i) + \gamma \cdot r(i) & \text{for} \quad i = j \\ q_{ij}(f_i) \cdot \mathrm{e}^{\gamma\, r(i,j)} & \text{for} \quad i \neq j \end{cases}$$

Observe that if $\gamma = 0$ then $\mathbf{U}^{\pi}(0,t) = \mathbf{P}^{\pi}(0,t)$ and (6) takes on the standard form of the Kolmogorov equation for calculating probability distribution of the Markov chain X^{c}.

To verify (6), since $u^{\gamma}(\cdot)$ is separable and multiplicative on taking expectations and conditioning on $X(\Delta)$ we immediately conclude that

$$U_i^{\pi}(\gamma, t+\Delta) =$$

$$= \sum_{j=1}^{N} P_{ij}^{\pi}(\Delta) \cdot [\mathrm{e}^{\gamma r(i)\Delta}\delta_{ij} + \mathrm{e}^{\gamma r(i,j)}(1-\delta_{ij})] \cdot U_j^{\pi}(\gamma, \Delta, t+\Delta).$$

Since for policy $\pi = f(t)$

$$P_{ij}^{\pi}(\Delta) = \begin{cases} 1 + q_{ii}(f_i(0))\Delta + o(\Delta^2) & \text{for} \quad i = j \\ q_{ij}(f_i(0))\Delta + o(\Delta^2) & \text{for} \quad i \neq j \end{cases}$$

on letting $\Delta \to 0+$ we conclude that

$$\begin{aligned} U_i^{\pi}(\gamma, t+\Delta) &= (1 + q_{ii}(f_i(0))\Delta)\mathrm{e}^{\gamma r(i)\Delta} \cdot U_i^{\pi}(\gamma, \Delta, t+\Delta) \\ &\quad + \sum_{j=1, j\neq i}^{N} q_{ij}(f_i(0))\Delta \mathrm{e}^{\gamma r(ij)} \cdot U_j^{\pi}(\gamma, \Delta, t+\Delta) + o(\Delta^2), \\ \mathrm{e}^{\gamma r(i)\Delta} &= 1 + \gamma r(i)\Delta + o(\Delta^2). \end{aligned}$$

Observe that if $\pi = f(t)$ is a piecewise constant policy controlling

the chain with switching time points
$t_0 = 0 < t_1 < t_2 < \ldots < t_i < \ldots < t_{n-1} < t < t_n$ such that
$f(t) = f^i$ for $t \in (t_{i-1}, t]$, $i = 1, 2, \ldots$ then

$$\mathbf{U}^\gamma(\pi, t) = \prod_{i=1}^{n-1} \exp[\bar{\mathbf{Q}}^{(\gamma)}(f^i)(t_i - t_{i-1})] \cdot \exp[\bar{\mathbf{Q}}^{(\gamma)}(f^n)(t - t_{n-1})] \cdot \mathbf{e}$$

Hence $\mathbf{U}^\pi(\gamma, t)$ is a linear combination of exponential functions with the exponents being the eigenvalues of the matrix $\bar{\mathbf{Q}}^{(\gamma)}(f) = [\bar{q}_{ij}^\gamma(f_i)]$ and the real part of the eigenvalues determines the growth of the elements of $U^\pi(\gamma, t)$.

Moreover, for the certainty equivalent we have
$Z_i^\gamma(\pi, t) = \gamma^{-1} \ln U_i^\gamma(\pi, t)$ and
$J_i^\pi(\gamma) = \limsup_{t\to\infty} t^{-1} Z_i^\gamma(\pi, t)$ is the mean value of $Z_i^\gamma(\pi, t)$.

4. Products of Nonnegative Matrices

We employ the following useful properties of the matrix family $\{\bar{\mathbf{P}}^{(\gamma)}(f) : f \in \mathcal{F}\}$, $\rho(f)$ is the spectral radius of $\bar{\mathbf{P}}^{(\gamma)}(f)$:

Result 3. If every $\bar{\mathbf{P}}^{(\gamma)}(f)$ is irreducible, then there exists $f^* \in \mathcal{F}$, and $\mathbf{v}(f^*) > \mathbf{0}$ (i.e. $\mathbf{v}(f^*)$ strictly positive) such that for any $f \in \mathcal{F}$

$$\bar{\mathbf{P}}^{(\gamma)}(f) \cdot \mathbf{v}(f^*) \leq \bar{\mathbf{P}}^{(\gamma)}(f^*) \cdot \mathbf{v}(f^*) = \rho(f^*)\,\mathbf{v}(f^*), \quad \rho(f^*) \geq \rho(f). \tag{7}$$

Moreover, (7) can be fulfilled even for reducible matrices, on condition that $\bar{\mathbf{P}}^{(\gamma)}(f^*)$ can be decomposed as

$$\bar{\mathbf{P}}^{(\gamma)}(f^*) = \begin{bmatrix} \bar{\mathbf{P}}^{(\gamma)}_{(00)}(f^*) & \bar{\mathbf{P}}^{(\gamma)}_{(01)}(f^*) & \dots & \bar{\mathbf{P}}^{(\gamma)}_{(0r)}(f^*) \\ \mathbf{0} & \bar{\mathbf{P}}^{(\gamma)}_{(11)}(f^*) & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \bar{\mathbf{P}}^{(\gamma)}_{(rr)}(f^*) \end{bmatrix} \tag{8}$$

such that:

a) For the spectral radius of every irreducible class of $\bar{\mathbf{P}}^{(\gamma)}_{(ii)}(f^*)$ (with $i = 1, \ldots, r$) it holds

$$\rho_i(f^*) = \rho(f^*), \text{ and}$$

b) the spectral radius of (possibly reducible) $\bar{\mathbf{P}}^{(\gamma)}_{(00)}(f^*)$ is less than $\rho(f^*)$, and some $\bar{\mathbf{P}}^{(\gamma)}_{(0j)}(f^*)$ is nonvanishing.

Using the terminology of Markov chain theory, conditions a) and b) can be formulated as:
Each irreducible class of $\bar{\mathbf{P}}^{(\gamma)}_{(00)}(f^*)$ has access to some diagonal class $\bar{\mathbf{P}}^{(\gamma)}_{(ii)}(f^*)$ with $i = 1, \ldots, r$; accessibility is considered in accordance with accessibility of the underlying Markov chain.

Remark. If every $\bar{\mathbf{P}}^{(\gamma)}(f)$ is irreducible, then there also exists $\hat{f} \in \mathcal{F}$, such that $\mathbf{v}(\hat{f}) > \mathbf{0}$ and, for any $f \in \mathcal{F}$, $\rho(\hat{f}) \leq \rho(f)$

$$\bar{\mathbf{P}}^{(\gamma)}(f) \cdot \mathbf{v}(\hat{f}) \geq \bar{\mathbf{P}}^{(\gamma)}(\hat{f}) \cdot \mathbf{v}(f^*) = \rho(\hat{f})\,\mathbf{v}(\hat{f}).$$

Remark. Assume that in the matrix $\mathbf{P}(f)$ contains a single class of recurrent states. Since $\bar{p}_{ij}^{f_i} := p_{ij}^{f_i} \cdot \mathrm{e}^{\gamma r_{ij}}$, i.e. elements of the matrix $\bar{\mathbf{P}}^{(\gamma)}(f)$, are continuous function of the risk aversion coefficient γ and for γ sufficiently close to null the matrix $\bar{\mathbf{P}}^{(\gamma)}(f)$ has a strictly positive right Perron eigenvector. Hence it may happen that for sufficiently large $\gamma > 0$ no strictly positive right Perron eigenvector of $\bar{\mathbf{P}}^{(\gamma)}(f)$ exists.

However, if (after suitable permutations of rows and corresponding columns) the submatrix corresponding to transient states can be written as a upper triangular matrix with null elements on the main diagonal, spectral radius of the submatrix of transient state equals zero, regardless the its elements and the recurrent state of $\mathbf{P}(f)$ must be reached in a number if transitions not exceeding the number of transient states.

Up to now we have assumed existence of strictly positive right Perron eigenvectors of the matrix $\bar{\mathbf{P}}^{(\gamma)}(f)$. In general it holds the following:

Fact 1. For a given matrix $\{\bar{\mathbf{P}}^{(\gamma)}(f), f \in \mathcal{F}\}$ on suitably permuting rows and corresponding columns the matrix $\bar{\mathbf{P}}^{(\gamma)}(f)$ can be written in the following block triangular form:

$$\bar{\mathbf{P}}^{(\gamma)}(f) = \begin{bmatrix} \bar{\mathbf{P}}_{11}^{(\gamma)}(f) & \bar{\mathbf{P}}_{12}^{(\gamma)}(f) & \dots & \bar{\mathbf{P}}_{1s}^{(\gamma)}(f) \\ \mathbf{0} & \bar{\mathbf{P}}_{22}^{(\gamma)}(f) & \dots & \bar{\mathbf{P}}_{2s}^{(\gamma)}(f) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \bar{\mathbf{P}}_{ss}^{(\gamma)}(f) \end{bmatrix} \tag{9}$$

where the diagonal blocks $\bar{\mathbf{P}}_{ii}^{(\gamma)}(f)$ with spectral radii $\rho_i(f)$ are the "biggest" submatrices of $\bar{\mathbf{P}}^{(\gamma)}(f)$ having strictly positive right eigenvectors corresponding to $\rho_i(f)$, i.e.

$$\bar{\mathbf{P}}_{ii}^{(\gamma)}(f) \cdot \mathbf{v}_i(f) = \rho_i(f) \cdot \mathbf{v}_i(f), \quad \text{where}$$

$$\rho_i(f) \geq \rho_{i+1}(f) \quad \text{for} \quad i = 1, \ldots, s.$$

Observe that each diagonal block $\bar{\mathbf{P}}_{ii}^{(\gamma)}(f)$ in (9) may be reducible and if $\bar{\mathbf{P}}_{ii}^{(\gamma)}(f^*)$ is reducible, then it can be decomposed according to (8).

It is not difficult to verify the assertion of Fact 1. Simply identify the basic class(es) of $\bar{\mathbf{P}}^{(\gamma)}(f)$, i.e. the irreducible classes with maximal spectral radius, and identify those non-basic classes having access to the basic class. In such a way we can construct the diagonal class $\bar{\mathbf{P}}_{11}^{(\gamma)}(f)$ and repeat the same procedure for the remaining classes of the matrix $\bar{\mathbf{P}}^{(\gamma)}(f)$.

For what follows the extension of the above results of Fact 1 on the whole family of matrices $\{\bar{\mathbf{P}}^{(\gamma)}(f), f \in \mathcal{F}\}$ is very important.

Result 4. For the set of nonnegative matrices there exists suitable labelling of states such that:

Every $\bar{\mathbf{P}}^{(\gamma)}(f)$ with $f \in \mathcal{F}$ is block triangular, i.e.

$$\bar{\mathbf{P}}^{(\gamma)}(f) = \begin{bmatrix} \bar{\mathbf{P}}_{11}^{(\gamma)}(f) & \bar{\mathbf{P}}_{12}^{(\gamma)}(f) & \dots & \bar{\mathbf{P}}_{1s}^{(\gamma)}(f) \\ \mathbf{0} & \bar{\mathbf{P}}_{22}^{(\gamma)}(f) & \dots & \bar{\mathbf{P}}_{2s}^{(\gamma)}(f) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \bar{\mathbf{P}}_{ss}^{(\gamma)}(f) \end{bmatrix} \tag{10}$$

where all $\bar{\mathbf{P}}_{ii}^{(\gamma)}(f)$ have fixed dimensions, and are the "biggest" submatrices of $\bar{\mathbf{P}}^{(\gamma)}(f)$ having strictly positive right eigenvectors corresponding to the maximal possible spectral radii of the corresponding submatrices, i.e.

there exists $\bar{\mathbf{P}}^{(\gamma)}(f^*)$ along with $\mathbf{v}_i(f^*) > \mathbf{0}$ (i. e. strictly positive) such that for all $i = 1, 2, \dots, s$

$$\rho_i(f^*) \geq \rho_i(f); \qquad \rho_i(f^*) \geq \rho_{i+1}(f^*) \tag{11}$$

$$\begin{aligned} \bar{\mathbf{P}}_{ii}^{(\gamma)}(f) \cdot \mathbf{v}_i(f^*) &\leq \bar{\mathbf{P}}_{ii}^{(\gamma)}(f^*) \cdot \mathbf{v}_i(f^*) \\ &= \rho_i(f^*)\, \mathbf{v}_i(f^*) \end{aligned} \tag{12}$$

Observe that each diagonal block $\bar{\mathbf{P}}_{ii}^{(\gamma)}(f)$ in (10) may be reducible and if $\bar{\mathbf{P}}_{ii}^{(\gamma)}(f)$ is reducible, then it can be decomposed according to (8).

Remark. The proof of Result 3 can be performed by policy iteration. However, to verify Result 4 it is possible to show that for every $f \in \mathcal{F}$ $\bar{\mathbf{P}}^{(\gamma)}(f)$ can be decomposed in a block-triangular form according to (9). Then on combining policy iterations for each diagonal block along with accessibility to diagonal blocks with higher spectral radius we can finish the proof of Result 4.

We make the following assumption:

Assumption GA. A strict inequalities hold in the second part of (16), i.e.:

$$\rho_1(f^*) > \rho_2(f^*) > \ldots > \rho_s(f^*) \tag{13}$$

Remark. Observe the case $\rho_i(f) = \rho_{i+1}(f)$ can be easily excluded, since we may assume that, if necessary, after small perturbations of some values $p_{ij}^{f_i}$ and r_{ij}, we arrive at $\rho_i(f) > \rho_{i+1}(f)$ and condition (13) will be fulfilled.

Moreover, notice that the total reward or total β-discounted reward of Markov decision chains with the linear utility function can be also expressed as a product of nonnegative matrices:

$$\bar{\mathbf{P}}^{(0)}(f) = \begin{bmatrix} \mathbf{P}(f) & \mathbf{r}(f) \\ \mathbf{0} & 1 \end{bmatrix} \quad \text{or} \quad \bar{\mathbf{P}}^{(0)}(f) = \begin{bmatrix} \beta\mathbf{P}(f) & \mathbf{r}(f) \\ \mathbf{0} & 1 \end{bmatrix}$$

and for the undiscounted case no strictly positive right Perron eigenvector of $\bar{\mathbf{P}}^{(0)}(f)$ exists.

Extension to Products of Matrices with Nonnegative Off-Diagonal Entries

Considering any matrix with nonnegative off-diagonal entries, say $\bar{\mathbf{Q}}^{(\gamma)}(f)$, then for $\alpha > 0$ sufficiently large the resulting matrix

$\bar{\mathbf{P}}^{(\gamma)}(f) := (\bar{\mathbf{Q}}^{(\gamma)}(f) + \alpha \mathbf{I})$ is nonnegative ($\mathbf{I}$... identity matrix). Moreover, if $\lambda(f)$ is an eigenvalue of $\bar{\mathbf{Q}}^{(\gamma)}(f)$ then

$(\lambda(f) + \alpha)$ is an eigenvalue of $\mathbf{P}^{(\gamma)}(f)$ and the corresponding eigenvectors of $\bar{\mathbf{P}}^{(\gamma)}(f)$ and $\bar{\mathbf{Q}}^{(\gamma)}(f)$ are identical. In particular, for the spectral radius of $\bar{\mathbf{P}}^{(\gamma)}(f)$ and the maximum real eigenvalue $\sigma(f)$ of $\bar{\mathbf{Q}}^{(\gamma)}(f)$ we have

$\rho(f) = (\sigma(f) + \alpha)$ and $\mathbf{v}(f)$ is the corresponding Perron eigenvector for both $\bar{\mathbf{P}}^{(\gamma)}(f)$ and $\bar{\mathbf{Q}}^{(\gamma)}(f)$.

Moreover, from the previous results for nonnegative matrices we immediately conclude the following facts:

Result 5. If every $\bar{\mathbf{Q}}^{(\gamma)}(f)$ is has strictly positive right Perron eigenvector (e.g. if it is irreducible), then there exists $f^* \in \mathcal{F}$, and $\mathbf{v}(f^*) > \mathbf{0}$ such that for any $f \in \mathcal{F}$

$$\bar{\mathbf{Q}}^{(\gamma)}(f) \cdot \mathbf{v}(f^*) \leq \bar{\mathbf{Q}}^{(\gamma)}(f^*) \cdot \mathbf{v}(f^*) = \sigma(f^*)\mathbf{v}(f^*). \tag{14}$$

Result 6. For the set $\{\bar{\mathbf{Q}}^{(\gamma)}(f), f \in \mathcal{F}\}$ of matrices with nonnegative off-diagonal entries there exists suitable labelling of states such that:
Every $\bar{\mathbf{Q}}^{(\gamma)}(f)$ with $f \in \mathcal{F}$ is block triangular, i.e.

$$\bar{\mathbf{Q}}^{(\gamma)}(f) = \begin{bmatrix} \bar{\mathbf{Q}}_{11}^{(\gamma)}(f) & \bar{\mathbf{Q}}_{12}^{(\gamma)}(f) & \dots & \bar{\mathbf{Q}}_{1s}^{(\gamma)}(f) \\ \mathbf{0} & \bar{\mathbf{Q}}_{22}^{(\gamma)}(f) & \dots & \bar{\mathbf{Q}}_{2s}^{(\gamma)}(f) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \bar{\mathbf{Q}}_{ss}^{(\gamma)}(f) \end{bmatrix} \tag{15}$$

where all $\bar{\mathbf{Q}}_{ii}^{(\gamma)}(f)$ have fixed dimensions, and are the "biggest" submatrices of $\bar{\mathbf{Q}}^{(\gamma)}(f)$ having strictly positive right eigenvectors corresponding to the maximal possible spectral radii of the corresponding submatrices, i.e. there exists $\bar{\mathbf{Q}}^{(\gamma)}(f^*)$ along with $\mathbf{v}_i(f^*) > \mathbf{0}$ (i. e. strictly positive) such that for all $i = 1, \ldots, s$

$$\sigma_i(f^*) \geq \sigma_i(f); \qquad \sigma_i(f^*) \geq \sigma_{i+1}(f^*) \tag{16}$$

$$\bar{\mathbf{Q}}_{ii}^{(\gamma)}(f) \cdot \mathbf{v}_i(f^*) \leq \bar{\mathbf{Q}}_{ii}^{(\gamma)}(f^*) \cdot \mathbf{v}_i(f^*) = \sigma_i(f^*)\, \mathbf{v}_i(f^*) \tag{17}$$

Observe that each diagonal block $\bar{\mathbf{Q}}_{ii}^{(\gamma)}(f)$ in (15) may be reducible and if on suitably permuting rows and corresponding columns of $\bar{\mathbf{Q}}^{(\gamma)}(f)$ it is possible to decompose $\bar{\mathbf{Q}}^{(\gamma)}(f)$ in the following block-triangular form:

$$\bar{\mathbf{Q}}^{(\gamma)}(f) = \begin{bmatrix} \bar{\mathbf{Q}}_{(\mathrm{NN})}^{(\gamma)}(f) & \bar{\mathbf{Q}}_{(\mathrm{NB})}^{(\gamma)}(f) \\ 0 & \bar{\mathbf{Q}}_{(\mathrm{BB})}^{(\gamma)}(f) \end{bmatrix} \tag{18}$$

where $\bar{\mathbf{Q}}^{(\gamma)}_{(\mathrm{NN})}(f)$ and $\bar{\mathbf{Q}}^{(\gamma)}_{(\mathrm{BB})}(f)$ (with maximum real eigenvalues $\sigma^{(\gamma)}_{(\mathrm{N})}(f)$ and $\sigma^{(\gamma)}_{(\mathrm{B})}(f)$) are (in general reducible) matrices such that:

- $\sigma_{(\mathrm{N})}(f) < \sigma(f)$,
- $\sigma_{(\mathrm{B})}(f) = \sigma^{(\gamma)}(f)$ and $\bar{\mathbf{Q}}^{(\gamma)}_{(\mathrm{BB})}(f)$ is diagonal with irreducible diagonal blocks $\bar{\mathbf{Q}}^{(\gamma)}_{(ii)}(f)$ (for $i = 1, \ldots, r$), such that the real eigenvalue $\sigma_i(f)$ of every $\bar{\mathbf{Q}}^{(\gamma)}_{(ii)}(f)$ is equal to $\sigma(f)$,
- maximum real eigenvalue of each irreducible class of $\bar{\mathbf{Q}}^{(\gamma)}_{(\mathrm{NN})}(f)$ is less than $\sigma^{(\gamma)}(f)$, and each class has access to $\bar{\mathbf{Q}}^{(\gamma)}_{(\mathrm{BB})}(f)$.

Observe that the above decomposition well correspond to the canonical decomposition of a continuous-time multichain transition rate matrix.

5. Asymptotic Behaviour of Expected Utilities

Discrete-time Case

Recall that in vector notation we can write

$$\mathbf{U}^{\pi}(\gamma, 0, n) = \bar{\mathbf{P}}^{(\gamma)}(f^0) \cdot \mathbf{U}^{\pi}(\gamma, 1, n) \tag{19}$$

where the ij-th entry of the $N \times N$ matrix $\bar{\mathbf{P}}^{(\gamma)}(f)$ is equal to $\bar{p}_{ij}^{f_i} = p_{ij}^{f_i} \cdot \mathrm{e}^{\gamma r_{ij}}$, and $\mathbf{U}^{\pi}(\gamma, n, n) = \mathbf{e}$.

Iterating (19) we get if policy $\pi = (f^n)$ is followed

$$\mathbf{U}^{\pi}(\gamma, n) = \bar{\mathbf{P}}^{(\gamma)}(f^0) \cdot \bar{\mathbf{P}}^{(\gamma)}(f^1) \cdot \ldots \cdot \bar{\mathbf{P}}^{(\gamma)}(f^{n-1}) \cdot \mathbf{e}. \tag{20}$$

Observe that in general the matrix $\bar{\mathbf{P}}^{(\gamma)}(f)$ can be decomposed according to (17) and there exists decision vector $f^* \in \mathcal{F}$ such that for all $i = 1, 2, \ldots, s$

$$\rho_i(f^*) \geq \rho_i(f); \qquad \rho_i(f^*) \geq \rho_{i+1}(f^*)$$

$$\begin{aligned} \rho_i(f^*)\mathbf{v}_i(f^*) &= \bar{\mathbf{P}}_{ii}^{(\gamma)}(f^*) \cdot \mathbf{v}_i(f^*) \\ &\geq \bar{\mathbf{P}}_{ii}^{(\gamma)}(f) \cdot \mathbf{v}_i(f^*) \end{aligned}$$

Now we focus attention on the case with $s = 1$, i.e. we set $\bar{\mathbf{P}}^{(\gamma)}(f) = \bar{\mathbf{P}}_{11}^{(\gamma)}(f)$ and assume that the underlying Markov chain contains one communicating class of recurrent states and some transient states, i.e.

$$\mathbf{P}(f) = \begin{bmatrix} \mathbf{P}_{TT}(f) & \mathbf{P}_{TR}(f) \\ \mathbf{0} & \mathbf{P}_{RR}(f) \end{bmatrix}$$

Since for elements of $\bar{\mathbf{P}}^{(\gamma)}(f)$ we have

$$\bar{p}_{ij}^{f} := p_{ij}^{f} \cdot e^{\gamma r_{ij}}$$

at least γ sufficiently close to 0 it holds

$$\rho(\bar{\mathbf{P}}_{TT}^{(\gamma)}(f)) < \rho(\bar{\mathbf{P}}_{RR}^{(\gamma)}(f)) \quad \text{for any} \quad f \in \mathcal{F}.$$

Under the above condition the basic class of $\bar{\mathbf{P}}^{(\gamma)}(f^*)$ corresponds to the communicating class of $\mathbf{P}(f)$, and there exists $\rho(f^*) = \rho^*$ and $\mathbf{v}(f^*) > \mathbf{0}$ (strictly positive) such that (cf. (14), (15))

$$\begin{aligned} \bar{\mathbf{P}}^{(\gamma)}(f) \cdot \mathbf{v}(f^*) &\leq \rho(f^*) \cdot \mathbf{v}(f^*) = \bar{\mathbf{P}}^{(\gamma)}(f^*) \cdot \mathbf{v}(f^*), \\ \rho(f) &\leq \rho(f^*) \equiv \rho^* \quad \text{for all } f \in \mathcal{F}. \end{aligned}$$

Iterating (4) and using (5) we can immediately conclude that for any policy $\pi = (f^n)$

$$\begin{aligned} \prod_{k=0}^{n-1} \bar{\mathbf{P}}^{(\gamma)}(f^n) \cdot \mathbf{v}(f^*) &\leq (\bar{\mathbf{P}}^{(\gamma)}(f^*))^n \cdot \mathbf{v}(f^*) \\ &= (\rho^*)^n \cdot \mathbf{v}(f^*) \end{aligned} \tag{21}$$

and hence the asymptotic behaviour of $\mathbf{U}^\pi(\gamma, n)$ (or of $\mathbf{U}^\pi(\gamma, m, n)$ if m is fixed) heavily depends on $\rho(f^*)$.

Moreover, for any stationary policy $\pi \sim (f)$ we have

$$\rho(f)\,\mathbf{v}(f) = \bar{\mathbf{P}}^{(\gamma)}(f) \cdot \mathbf{v}(f) \tag{22}$$

i.e. for unknowns $g(f)$, $w_i(f)$ $(i = 1, \ldots, N)$
defined by

$$v_i(\cdot) = \mathrm{e}^{\gamma w_i(\cdot)}, \quad \rho(f) = \mathrm{e}^{\gamma g(f)}$$

from (22) we get the following set of equations

$$\mathrm{e}^{\gamma(g(f)+w_i(f))} = \sum_{j\in\mathcal{I}} p_{ij}^{f_i} \cdot \mathrm{e}^{\gamma(r_{ij}+w_j(f))} \tag{23}$$

Keeping this notations (21) can be written

$$\begin{aligned} \mathrm{e}^{\gamma(g(f)+w_i(f^*))} &\leq \mathrm{e}^{\gamma(g(f^*)+w_i(f^*))} \\ &= \sum_{j\in\mathcal{I}} p_{ij}^{f_i^*} \cdot \mathrm{e}^{\gamma(r_{ij}+w_j(f^*))} \end{aligned}$$

$$\text{for} \quad i = 1, \ldots, N \tag{24}$$

and the set of equations (with respect to $g(f)$, $w_i(f)$'s)

$$\mathrm{e}^{\gamma(g(f)+w_i(f))} = \max_{f\in\mathcal{F}}\Big\{\sum_{j\in\mathcal{I}} p_{ij}^{f_i}\cdot \mathrm{e}^{\gamma(r_{ij}+w_j(f))}\Big\} \tag{25}$$

can be called as

γ-average reward optimality equation.

In the multiplicative form (used before) we write

$$\rho(f)\cdot v_i(f) = \max_{f\in\mathcal{F}}\Big\{\sum_{j\in\mathcal{I}} p_{ij}^{f_i}\cdot \mathrm{e}^{\gamma r_{ij}}\cdot v_j(f)\Big\}$$

$$\text{for} \quad i=1,\ldots,N \tag{26}$$

Observe that the solution to (26) is unique up to a multiplicative constant, say K.

Since $v_i(\cdot) = \mathrm{e}^{\gamma w_i(\cdot)}$ and $w_i(\cdot)$'s must be unique up to an additive constant, say $\bar{c}$, where

$$K = \mathrm{e}^{\gamma \bar{c}} \Longleftrightarrow \bar{K} = \frac{1}{\gamma} \cdot \ln K.$$

Now we shall consider the case with $s = 2$.

We again assume that the Markov chain contains a single class of recurrent states and some transient states, i.e.

$$\mathbf{P}(f) = \begin{bmatrix} \mathbf{P}_{TT}(f) & \mathbf{P}_{TR}(f) \\ \mathbf{0} & \mathbf{P}_{RR}(f) \end{bmatrix}$$

but for the selected value of γ the basic class of $\bar{\mathbf{P}}^{(\gamma)}(f)$ (with elements $\bar{p}_{ij}^{f} := p_{ij}^{f} \cdot \mathrm{e}^{\gamma r_{ij}}$)

is contained in the set of transient states of $\mathbf{P}(f)$,
and is unique.
Then the resulting matrix can be decomposed as

$$\bar{\mathbf{P}}^{(\gamma)}(f) = \begin{bmatrix} \bar{\mathbf{P}}_{11}^{(\gamma)}(f) & \bar{\mathbf{P}}_{12}^{(\gamma)}(f) \\ \mathbf{0} & \bar{\mathbf{P}}_{22}^{(\gamma)}(f) \end{bmatrix} \tag{27}$$

with $\rho_1(f) > \rho_2(f)$ and

$\mathbf{v}_1(f) > \mathbf{0}, \quad \mathbf{v}_2(f) > \mathbf{0}$

such that

$\rho_1(f)\mathbf{v}_1(f) = \bar{\mathbf{P}}_{11}^{(\gamma)}(f) \cdot \mathbf{v}_1(f)$, and

$\rho_2(f)\mathbf{v}_1(f) = \bar{\mathbf{P}}_{22}^{(\gamma)}(f) \cdot \mathbf{v}_2(f)$.

Then $\varepsilon(f) := \rho_2(f)/\rho_1(f) < 1$.

Since

$$(\bar{\mathbf{P}}^{(\gamma)}(f))^n = \begin{bmatrix} (\bar{\mathbf{P}}_{11}^{(\gamma)}(f))^n & \sum_{k+\ell=n-1} (\bar{\mathbf{P}}_{11}^{(\gamma)}(f))^k \bar{\mathbf{P}}_{12}^{(\gamma)}(f) (\bar{\mathbf{P}}_{22}^{(\gamma)}(f))^\ell \\ \mathbf{0} & (\bar{\mathbf{P}}_{22}^{(\gamma)}(f))^n \end{bmatrix} \tag{28}$$

we can conclude that for some $\bar{\mathbf{P}}_{12}^{(\gamma)} \geq \bar{\mathbf{P}}_{12}^{(\gamma)}(f)$
such that $\bar{\mathbf{P}}_{12}^{(\gamma)} \cdot \mathbf{v}_2(f) = \alpha \cdot \mathbf{v}_1(f)$
it holds

$$\begin{aligned} & \sum_{k+\ell=n-1} (\bar{\mathbf{P}}_{11}^{(\gamma)}(f))^k \cdot \bar{\mathbf{P}}_{12}^{(\gamma)}(f) \cdot (\bar{\mathbf{P}}_{22}^{(\gamma)}(f))^\ell \cdot \mathbf{v}_2(f) \\ \leq \quad & \alpha \cdot (\rho_1(f))^k \cdot (\rho_2(f))^\ell \mathbf{v}_1(f) \\ \leq \quad & \alpha \cdot (\rho_1(f))^{n-1} \cdot \frac{1}{1-\varepsilon(f)} \cdot \mathbf{v}_1(f) \end{aligned}$$

Hence for suitably selected $\mathbf{v}_1(f)$, $\mathbf{v}_2(f)$

$$\begin{bmatrix} \mathbf{U}_1^{\pi}(\gamma, n) \\ \mathbf{U}_2^{\pi}(\gamma, n) \end{bmatrix} \leq \begin{bmatrix} \bar{\mathbf{P}}_{11}^{(\gamma)}(f) & \bar{\mathbf{P}}_{12}^{(\gamma)}(f) \\ \mathbf{0} & \bar{\mathbf{P}}_{22}^{(\gamma)}(f) \end{bmatrix}^n \cdot \begin{bmatrix} \mathbf{v}_1(f) \\ \mathbf{v}_2(f) \end{bmatrix}$$

$$\leq \begin{bmatrix} (\rho_1(f))^n \left\{ \rho_1(f) + \alpha \frac{1}{1-\varepsilon(f)} \cdot \right\} \cdot \mathbf{v}_1(f) \\ (\rho_2(f))^n \cdot \mathbf{v}_2(f) \end{bmatrix}$$

and the maximal growth rate of $\mathbf{U}_1^{\pi}(\gamma, n)$, $\mathbf{U}_2^{\pi}(\gamma, n)$ is bounded by $\rho_1(f)$, $\rho_2(f)$ respectively.

Continuous-time Case

Recall that in the continuous-time setting the vector of expected utilities $\mathbf{U}^{\pi}(\gamma, t)$ satisfies for $t \in [0, t^*)$ the following differential equation

$$\frac{\mathrm{d}\mathbf{U}^{\pi}(\gamma, t)}{\mathrm{d}t} = \bar{\mathbf{Q}}^{(\gamma)}(f(t)) \cdot \mathbf{U}^{\pi}(\gamma, t), \quad \text{with} \quad \mathbf{U}^{\pi}(\gamma, t^*) = \mathbf{e} \quad (29)$$

where $\bar{\mathbf{Q}}^{(\gamma)}(f) = [\bar{q}_{ij}^{\gamma}(f_i)]$ is an $N \times N$ matrix with nonnegative off-diagonal elements

$$\bar{q}_{ij}^{\gamma}(f_i) = \begin{cases} q_{ii}(f_i) + \gamma \cdot r(i) & \text{for} \quad i = j \\ q_{ij}(f_i) \cdot \mathrm{e}^{\gamma\, r(i,j)} & \text{for} \quad i \neq j \end{cases}$$

Observe that if $\pi = f(t)$ is a piecewise constant policy controlling the chain with switching time points
$t_0 = 0 < t_1 < t_2 < \ldots < t_i < \ldots < t_{n-1} < t^* < t_n$ such that
$f(t) = f^i$ for $t \in (t_{i-1}, t_i]$, $i = 1, 2, \ldots$ then

$$\mathbf{U}^{\gamma}(\pi, t) = \prod_{i=1}^{n-1} \exp[\bar{\mathbf{Q}}^{(\gamma)}(f^i)(t_i - t_{i-1})] \cdot \exp[\bar{\mathbf{Q}}^{(\gamma)}(f^n)(t - t_{n-1})] \cdot \mathbf{e}$$

Hence $\mathbf{U}^{\pi}(\gamma, t)$ is a linear combination of exponential functions with the exponents being the eigenvalues of the matrix $\bar{\mathbf{Q}}^{(\gamma)}(f) = [\bar{q}_{ij}^{\gamma}(f_i)]$ and the real part of the eigenvalues determines the growth of the elements of $U^{\pi}(\gamma, t)$.

Similarly to the discrete-time case, if for every $f \in \mathcal{F}$ the matrix $\bar{\mathbf{Q}}^{(\gamma)}(f)$ is irreducible, or at least every $\bar{\mathbf{Q}}^{(\gamma)}(f)$ has a strictly positive eigenvector corresponding to $\sigma^{(\gamma)}(f)$, there exists $\hat{\sigma}^{(\gamma)} \equiv \sigma^{(\gamma)}(\hat{f})$ and $\widehat{\nu}^{(\gamma)} \equiv \nu^{(\gamma)}(f^*) > 0$ (i.e. strictly positive eigenvector corresponding to $\sigma^{(\gamma)}$ such that

$$\hat{\sigma}^{(\gamma)}\ \widehat{\nu}^{(\gamma)} = \max_{f \in \mathcal{F}} \{\bar{\mathbf{Q}}^{(\gamma)}(f) \cdot \widehat{\nu}^{(\gamma)}\} = \bar{\mathbf{Q}}^{(\gamma)}(f^*) \cdot \widehat{\nu}^{(\gamma)}. \qquad (30)$$

Moreover, if condition (30) is fulfilled then for numbers $\alpha_2^{(\gamma)} > \alpha_1^{(\gamma)} > 0$ selected such that $\alpha_2^{(\gamma)}\nu(f^*) > e > \alpha_1^{(\gamma)}\nu(f^*)$ we conclude that

$$\begin{aligned} U^{(\gamma)}(\pi, t) &= \prod_{i=1}^{n-1} \exp[\bar{\mathbf{Q}}^{(\gamma)}(f^i)(t_i - t_{i-1})] \cdot \\ &\qquad \exp[\bar{\mathbf{Q}}^{(\gamma)}(f^n)(t - t_{n-1})] \cdot e \\ &\leq \alpha_2^{(\gamma)} \exp[\bar{\mathbf{Q}}^{(\gamma)}(f^*)t]\, \nu(f^*). \end{aligned}$$

However,

$$U^{\pi^*}(\gamma, t) = \exp[\bar{\mathbf{Q}}^{(\gamma)}(f^*)]e \geq \alpha_1^{(\gamma)} \exp[\bar{\mathbf{Q}}^{(\gamma)}(f^*)t]\, \nu(f^*)$$

So we have arrived at the following fact.

If condition (30) holds then for a given γ there exist numbers $\alpha_2^{(\gamma)} > \alpha_1^{(\gamma)} > 0$ such that for stationary policy $\pi^* \sim f^*$ and for arbitrary piecewise constant policy $\pi = f(t)$ it holds

$$\alpha_1^{(\gamma)}\, v^{(\gamma)}(f^*) \leq \exp[-\hat{\sigma} t]\, U^{\pi^*}(\gamma, t) \tag{31}$$

$$\alpha_2^{(\gamma)}\, v^{(\gamma)}(f^*) \geq \exp[-\hat{\sigma} t]\, U^{\pi}(\gamma, t). \tag{32}$$

This may be rephrased in words as:

Under condition (30) if policy $\pi = f(t)$ maximizing $U^{\pi}(\gamma, t)$ is followed the *growth rate* of each element of $U^{\pi}(\gamma, t)$ is the same and equals $\hat{\sigma}$. Moreover, stationary policy $\pi^* \sim f^*$ also maximizes the growth rate.

Summarizing these facts we arrive at

Result 7. If condition (30) holds then for any policy $\pi = f(t)$ the asymptotical mean value (i.e. maximum average reward) $J_i^{\pi}(\gamma)$ is bounded from below by $\gamma^{-1}\,\hat{\sigma}$.
Moreover, stationary policy $\hat{\pi} \sim \hat{f}$ yields the maximal asymptotical mean value $J_i^{\pi}(\gamma)$ that is independent of the starting state $i \in \mathcal{I}$ and equal to $\gamma^{-1}\,\hat{\sigma}$.

Up to now our analysis of the continuous-time case was based on maximizing real eigenvalue of a set of matrices with nonnegative off-diagonal entries (cf. (29)), in particular

$$\sigma^{(\gamma)}(f)\;\nu^{(\gamma)}(f) \;=\; \bar{\mathbf{Q}}^{(\gamma)}(f)\cdot\nu^{(\gamma)}(f) \tag{33}$$

$$\hat{\sigma}^{(\gamma)}\;\widehat{\nu}^{(\gamma)} \;=\; \max_{f\in\mathcal{F}}\{\bar{\mathbf{Q}}^{(\gamma)}(f)\cdot\widehat{\nu}^{(\gamma)}\} \tag{34}$$

However, in the discrete-time case we have shown that finding maximal $\rho^{(\gamma)}(f)$ is the same as finding solution of the well-known Poissonian equations.

Similarly, for the continuous-time case let us introduce $\bar{g}(f)$, $\bar{w}_i(f)$ $(i = 1, \ldots, N)$ such that

$$\nu_i(\cdot) = \mathrm{e}^{\gamma \bar{w}_i(\cdot)}, \quad \sigma(\cdot) = \mathrm{e}^{\gamma \bar{g}(\cdot)}$$

Then from (33) we get the following set of equations

$$\begin{aligned} \mathrm{e}^{\gamma[\bar{g}(f)+\bar{w}_i(f)]} &= \sum_{j \in \mathcal{I}, j \neq i} q_{ij}(f_i) \cdot \mathrm{e}^{\gamma[r(i,j)\bar{w}_j(f)]} \\ &+ \left[q_{ii}(f_i) + \gamma r(i)\right] \mathrm{e}^{\gamma \bar{w}_i(f)} \end{aligned} \tag{35}$$

Similarly, from (34) we have for $i = 1, \ldots, N$

$$\begin{aligned} \mathrm{e}^{\gamma(\bar{g}(f)+\bar{w}_i(f))} &\leq \mathrm{e}^{\gamma(\bar{g}(f^*)+\bar{w}_i(f^*)} = \sum_{j \in \mathcal{I}\, j \neq i} q_{ij}(f_i^*) \cdot \mathrm{e}^{\gamma(r(i,j)+\bar{w}_j(f^*)} \\ &+ \left[q_{ii}(f_i^*) + \gamma r(i)\right] \mathrm{e}^{\gamma \bar{w}_i(f^*)} \end{aligned} \tag{36}$$

Hence the set of equations (with respect to $\bar{g}(f)$, $\bar{w}_i(f)$'s)

$$\begin{aligned} \mathrm{e}^{\gamma(\bar{g}(f)+\bar{w}_i(f))} &= \max_{f\in\mathcal{F}}\{\sum_{j\in\mathcal{I}\, j\neq i} q_{ij}(f_i)\cdot \mathrm{e}^{\gamma(r(i,j)+\bar{w}_j(f)} \\ &+ [q_{ii}(f_i)+\gamma r(i)]\,\mathrm{e}^{\gamma\bar{w}_i(f)} \end{aligned} \tag{37}$$

can be called

continuous-time γ-average reward optimality equation.
Observe that the solution to (36) is unique up to a multiplicative constant, say K, and the values $\bar{w}_i(\cdot)$'s in (37) must be unique up to an additive constant, say $\bar{c}$, where $K = \mathrm{e}^{\gamma\bar{c}}$.

Similarly to nonnegative matrices, considering a (reducible) matrix $\bar{\mathbf{Q}}^{(\gamma)}(f)$ with nonnegative off-diagonal entries, there exists $f^* \in \mathcal{F}$ and a block-triangular decomposition of $\bar{\mathbf{Q}}^{(\gamma)}(f^*)$ such that every matrix $\bar{\mathbf{Q}}^{(\gamma)}(f)$ has a block-triangular structure with some specific properties summarized as

Result 8. There exists $f^* \in \mathcal{F}$ and a suitable labelling of states inducing the partition of the state space $\mathcal{I}$, say $\hat{\mathcal{I}} \equiv \bigcup_{i=1}^{s} \mathcal{I}_i(f^*)$, called the *basic partition*, such that:

Keeping the partition in accordance of $\hat{\mathcal{I}}$ then each $\bar{\mathbf{Q}}^{(\gamma)}(f)$ is block triangular, i.e.

$$\bar{\mathbf{Q}}^{(\gamma)}(f) = \begin{bmatrix} \bar{\mathbf{Q}}_{11}^{(\gamma)}(f) & \bar{\mathbf{Q}}_{12}^{(\gamma)}(f) & \dots & \bar{\mathbf{Q}}_{1s}^{(\gamma)}(f) \\ 0 & \bar{\mathbf{Q}}_{22}^{(\gamma)}(f) & \dots & \bar{\mathbf{Q}}_{2s}^{(\gamma)}(f) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \bar{\mathbf{Q}}_{ss}^{(\gamma)}(f) \end{bmatrix}, \qquad \forall f \in \mathcal{F} \tag{38}$$

where all $\bar{\mathbf{Q}}_{ii}^{(\gamma)}(f)$ have fixed dimensions equal to $\operatorname{card} \mathcal{I}_i(f^*)$, and for $i = 1, \dots, s$ $\bar{\mathbf{Q}}_{ii}^{(\gamma)}(f^*)$'s are the "biggest" submatrices of $\bar{\mathbf{Q}}^{(\gamma)}(f)$ having strictly positive right eigenvectors corresponding to the maximum real eigenvalues of the corresponding submatrices, i.e.

there exists $\bar{\mathbf{Q}}_{ii}^{(\gamma)}(f^*)$ along with $v_i^{(\gamma)}(f^*) > 0$ such that for any $f \in \mathcal{F}$ and all $i = 1, 2, \ldots, s$

$$\sigma_i^{(\gamma)}(f^*) \geq \sigma_i^{(\gamma)}(f); \qquad \sigma_i^{(\gamma)}(f^*) \geq \sigma_{i+1}^{(\gamma)}(f^*) \tag{39}$$

$$\begin{aligned} \bar{\mathbf{Q}}_{ii}^{(\gamma)}(f) \cdot v_i^{(\gamma)}(f^*) \;\leq\; & \bar{\mathbf{Q}}_{ii}^{(\gamma)}(f^*) \cdot v_i^{(\gamma)}(f^*) \\ & = \sigma_i^{(\gamma)}(f^*)\; v_i^{(\gamma)}(f^*). \end{aligned} \tag{40}$$

Observe that $\sigma_1^{(\gamma)}(f^*) = \sigma^{(\gamma)}(f^*)$ and that each diagonal block $\bar{\mathbf{Q}}_{ii}^{(\gamma)}(f)$ in (38) may be reducible, and if $\bar{\mathbf{Q}}_{ii}^{(\gamma)}(f^*)$ is reducible then it can be decomposed similarly as in (19).

We make the following assumption:

Assumption GB. For a given value of the risk aversion coefficient γ a strict inequalities holds in the second part of (39), i.e.:

$$\sigma_1^{(\gamma)}(f^*) > \sigma_2^{(\gamma)}(f^*) > \ldots > \sigma_s^{(\gamma)}(f^*). \tag{41}$$

Remark Observe that the case $\sigma_i^{(\gamma)}(f) = \sigma_{i+1}^{(\gamma)}(f)$ can be easily excluded, since, if necessary, we may assume that after small perturbations of some values $q_{ij}(f_i)$ and $r(i,j)$ (i.e. the perturbation of $q_{ij}^{(\gamma)}(f_i)$), we arrive at $\sigma_i^{(\gamma)}(f) > \sigma_{i+1}^{(\gamma)}(f)$ and condition (41) will be fulfilled.

So we have arrived at the following

Result 9. Let Assumption GB hold. Then for the matrix $\bar{\mathbf{Q}}^{(\gamma)}(f^*)$ with $f^* \in \mathcal{F}$ decomposed in accordance with the basic partition of the state space $\hat{\mathcal{I}} = \mathcal{I}_1(f^*) \cup \mathcal{I}_2(f^*) \cup \ldots \cup \mathcal{I}_s(f^*)$ it holds: Maximum possible growth rate is the same for each $j \in \mathcal{I}_i(f^*)$ and is equal to $\sigma_i^{(\gamma)}(f^*)$. Moreover, this growth rate can be obtained if stationary policy $\pi^* \sim f^*$ is followed. Moreover, maximal average rewards $J_j^{\pi^*}(\gamma)$ are the same for each $j \in \mathcal{I}_i(f^*)$ and are equal to $(\gamma)^{-1}\sigma_i^{(\gamma)}(f^*)$.

6. Value and Policy Iteration Algorithms

Recalling (5) we get if policy $\pi = (f^n)$ is followed

$$\mathbf{U}^{\pi}(\gamma, n) = \bar{\mathbf{P}}^{(\gamma)}(f^0) \cdot \bar{\mathbf{P}}^{(\gamma)}(f^1) \cdot \ldots \cdot \bar{\mathbf{P}}^{(\gamma)}(f^{n-1}) \cdot \mathbf{e}.$$

In case that there exists decision vector $f^* \in \mathcal{F}$ such that

$$\bar{\mathbf{P}}^{(\gamma)}(f) \cdot \mathbf{v}(f^*) \leq \rho(f^*)\,\mathbf{v}(f^*) = \bar{\mathbf{P}}^{(\gamma)}(f^*) \cdot \mathbf{v}(f^*), \tag{42}$$

$$\rho(f) \leq \rho(f^*) \equiv \rho^* \quad \text{for all } f \in \mathcal{F}. \tag{43}$$

we can immediately conclude that for $\pi = (f^n)$

$$\prod_{k=0}^{n-1} \bar{\mathbf{P}}^{(\gamma)}(f^k) \cdot \mathbf{v}(f^*) \leq (\bar{\mathbf{P}}^{(\gamma)}(f^*))^n \cdot \mathbf{v}(f^*)$$

$$= (\rho(f^*))^n \cdot \mathbf{v}(f^*) \tag{44}$$

Hence, since $\mathbf{v}(f^*) > \mathbf{0}$, on selecting $\mathbf{v}(f^*) \geq \mathbf{e}$, say we set $\mathbf{v}_{\mathrm{upb}}(f^*) := \mathbf{v}(f^*) \geq \mathbf{e}$, (44) yields an lower bound on the growth of $\mathbf{U}^{\pi}(\gamma, n)$.

From (44) we can easily conclude that if (42), (43) hold maximum possible growth of every $U_i^{\pi^*}(\gamma, n)$ is given by ρ^*.

Then for the corresponding values of certainty equivalents we get for $i = 1, 2, \ldots, N$ since $\gamma \neq 0$

$$Z_i^{\pi^*}(\gamma, n) = \frac{1}{\gamma} \cdot \ln[U_i^{\pi^*}(\gamma, n)] = \frac{1}{\gamma} \cdot [n \ln(\rho^*) + w_i] \tag{45}$$

and for the mean value of certainty equivalents we have

$$J_i^{\pi^*}(\gamma) = \frac{1}{\gamma} \cdot \ln[\rho^*] \tag{46}$$

The above procedures also enables to generate upper and lower bounds of the minimum growth rate and the corresponding certainty equivalents.

To this end, let us generate a sequence of maximum possible expected utilities by the following dynamic programming recursion for $n = 0, 1, \ldots$:

$$\begin{aligned} \hat{\mathbf{U}}(n+1) &= \max_{f \in \mathcal{F}} \bar{\mathbf{P}}^{(\gamma)}(f) \cdot \hat{\mathbf{U}}(n) := \bar{\mathbf{P}}^{(\gamma)}(\bar{f}^n) \cdot \hat{\mathbf{U}}(n), \\ &\quad \text{with} \quad \hat{\mathbf{U}}(0) = \mathbf{e}. \end{aligned} \tag{47}$$

Then on employing elements of the sequence $\hat{\mathbf{U}}(n)$ we can easily generate upper and lower bounds on the maximal growth rate ρ^*, denoted $\rho_{\max}(n)$ and $\rho_{\min}(n)$ respectively, where

$$\rho_{\max}(n) := \max_{i \in \mathcal{I}} \frac{\hat{U}_i(n+1)}{\hat{U}_i(n)}, \qquad \rho_{\min}(n) := \max_{i \in \mathcal{I}} \frac{\hat{U}_i(n+1)}{\hat{U}_i(n)}$$

Result 10. If there exists $f^* \in \mathcal{F}$, and $\mathbf{v}(f^*) > \mathbf{0}$ such that for any $f \in \mathcal{F}$

$$\bar{\mathbf{P}}^{(\gamma)}(f) \cdot \mathbf{v}(f^*) \leq \bar{\mathbf{P}}^{(\gamma)}(f^*) \cdot \mathbf{v}(f^*)$$

the sequence $\{\rho_{\max}(n)\}$ resp. $\{\rho_{\min}(n)\}$ is nonincreasing, resp. nondecreasing, and if $\mathbf{P}^{(\gamma)}(f^*)$ is aperiodic then

$$\lim_{n\to\infty} \rho_{\max}(n) = \lim_{n\to\infty} \rho_{\min}(n) = \rho^*$$

where ρ^* is the miximal possible growth rate.

The same holds also for mean values of the corresponding certainty equivalents. In particular,
$J_{\max}(\gamma, n) := \frac{1}{\gamma} \ln[\rho_{\max}(n)]$, $\quad J_{\min}(\gamma, n) := \frac{1}{\gamma} \ln[\rho_{\min}(n)]$,
and the sequence $\{J_{\max}(\gamma, n)\}$, resp. $\{J_{\min}(\gamma, n)\}$ is nonincreasing, resp. nondecreasing, and if $\mathbf{P}^{(\gamma)}(f^*)$ is aperiodic then

$$\lim_{n\to\infty} J_{\max}(\gamma, n) = \lim_{n\to\infty} J_{\min}(\gamma, n) = J_i^{\pi^*}$$

In case that there exists no $\mathbf{v}(f^*) > \mathbf{0}$ such that (42), (43) hold we can proceed as follows:

Suppose (for simplicity) that $\{\bar{\mathbf{P}}^{(\gamma)}(f), f \in \mathcal{F}\}$ can be decomposed as:

$$\bar{\mathbf{P}}^{(\gamma)}(f) = \begin{bmatrix} \bar{\mathbf{P}}_{11}^{(\gamma)}(f) & \bar{\mathbf{P}}_{12}^{(\gamma)}(f) \\ \mathbf{0} & \bar{\mathbf{P}}_{22}^{(\gamma)}(f) \end{bmatrix}$$

and there exists $f^* \in \mathcal{F}$ such that

$$\rho_1(f^*) > \rho_2(f^*), \quad \mathbf{v}_1(f^*) > \mathbf{0}, \quad \mathbf{v}_2(f^*) > \mathbf{0}$$

and for any $\bar{\mathbf{P}}^{(\gamma)}(f)$ with $f \in \mathcal{F}$:

$$\begin{aligned} \bar{\mathbf{P}}_{11}^{(\gamma)}(f) \cdot \mathbf{v}_1(f^*) &\geq \rho_1(f^*)\mathbf{v}_1(f^*) \\ &= \bar{\mathbf{P}}_{11}^{(\gamma)}(f^*) \cdot \mathbf{v}_1(f^*) \end{aligned}$$

$$\begin{aligned} \bar{\mathbf{P}}_{22}^{(\gamma)}(f) \cdot \mathbf{v}_2(f^*) &\geq \rho_2(f^*)\mathbf{v}_2(f^*) \\ &= \bar{\mathbf{P}}_{22}^{(\gamma)}(f^*) \cdot \mathbf{v}_2(f^*) \end{aligned}$$

Let the rows of $\bar{\mathbf{P}}_{11}^{(\gamma)}(f)$, resp. $\bar{\mathbf{P}}_{11}^{(\gamma)}(f)$, be labelled by numerals from $\mathcal{I}_1$, resp. $\mathcal{I}_2$. (Obviously, $\mathcal{I} = \mathcal{I}_1 \cup \mathcal{I}_2$.)

Then on iterating (47) under the aperiodicity of $\bar{\mathbf{P}}^{(\gamma)}(f^*)$ we get:

$$\lim_{n\to\infty} \frac{\hat{U}_i(n+1)}{\hat{U}_i(n)} = \rho_1(f^*), \quad \text{for any} \quad i \in \mathcal{I}_1$$

$$\lim_{n\to\infty} \frac{\hat{U}_i(n+1)}{\hat{U}_i(n)} = \rho_2(f^*), \quad \text{for any} \quad i \in \mathcal{I}_2$$

and for

$$\rho_{\max}^{(1)}(n) := \max_{i\in\mathcal{I}_1} \frac{\hat{U}_i(n+1)}{\hat{U}_i(n)}, \quad \rho_{\min}^{(1)}(n) := \max_{i\in\mathcal{I}_1} \frac{\hat{U}_i(n+1)}{\hat{U}_i(n)}$$

the sequence $\{\rho_{\max}^{(1)}(n)\}$ is nonincreasing, the sequence $\{\rho_{\min}^{(1)}(n)\}$ nondecreasing, and under the aperiodicity of $\bar{\mathbf{P}}_{11}^{(\gamma)}(f^*)$

$$\lim_{n\to\infty} \rho_{\max}^{(1)}(n) = \lim_{n\to\infty} \rho_{\min}^{(1)}(n) = \rho_1(f^*)$$

where $\rho_1(f^*)$ is the maximum possible growth rate that can occur in states from $\mathcal{I}_1$.

The same holds also for mean values of the corresponding certainty equivalents.
Finally, we present a policy iteration algorithm for finding stationary policy $\pi^* \sim f^*$ fulfilling inequality (6), i.e.

$$\begin{aligned} \bar{\mathbf{P}}^{(\gamma)}(f) \cdot \mathbf{v}(f^*) &\leq \rho(f^*)\,\mathbf{v}(f^*) = \bar{\mathbf{P}}^{(\gamma)}(f^*) \cdot \mathbf{v}(f^*) \\ &\qquad \text{with} \quad \mathbf{v}(f^*) > \mathbf{0} \\ \rho(f) &\leq \rho(f^*) \qquad \text{for all} \quad f \in \mathcal{F}. \end{aligned}$$

The policy iteration algorithms generates a sequence of stationary policies such that the corresponding sequence of spectral radii $\rho(f^{(k)})$ is non-decreasing (i.e. $\rho(f^{(k+1)}) \geq \rho(f^{(k)})$, resp. increasing if $\bar{\mathbf{P}}^{(\gamma)}(f^{(k+1)})$ is irreducible, and the sequence $\bar{\mathbf{P}}^{(\gamma)}(f^{(k)})$ converges monotonously to the matrix $\bar{\mathbf{P}}^{(\gamma)}(f^*)$.

Policy Iteration Algorithm

- *Step 0.* Select matrix $\bar{\mathbf{P}}^{(\gamma)}(f^{(0)})$ with $f^{(0)} \in \mathcal{F}$ such that the row sums are maximal, i.e. it holds $\bar{\mathbf{P}}^{(\gamma)}(f^{(0)}) \cdot \mathbf{e} \geq \bar{\mathbf{P}}^{(\gamma)}(f) \cdot \mathbf{e}$ for any $f \in \mathcal{F}$.
- *Step 1.* For the matrix $\bar{\mathbf{P}}^{(\gamma)}(f^{(k)})$ with $f^{(k)} \in \mathcal{F}$, $k = 0, 1, \ldots$ calculate its spectral radius $\rho(f^{(k)})$ along with its right Perron eigenvector $\mathbf{v}(f^{(k)})$.
- *Step 2.* Construct (if possible) the matrix $\bar{\mathbf{P}}^{(\gamma)}(f^{(k+1)})$ with $f^{(k+1)} \in \mathcal{F}$, such that

$$\bar{\mathbf{P}}^{(\gamma)}(f^{(k+1)}) \cdot \mathbf{v}(f^{(k)}) > \rho(f^{(k)})\,\mathbf{v}(f^{(k)}) = \bar{\mathbf{P}}^{(\gamma)}(f^{(k)}) \cdot \mathbf{v}(f^{(k)})$$

  (i.e., a strict inequality holds at least for one $i \in \mathcal{I}$).
- *Step 3.* If such a matrix $\bar{\mathbf{P}}^{(\gamma)}(f^{(k+1)})$ exists, then set $\bar{\mathbf{P}}^{(\gamma)}(f^{(k+1)}) := \bar{\mathbf{P}}^{(\gamma)}(f^{(k)})$ and repeat Step 1, else set $\bar{\mathbf{P}}^{(\gamma)}(f^{*}) := \bar{\mathbf{P}}^{(\gamma)}(f^{(k)})$, $f^{*} := f^{(k)}$ and stop.

In the remainder of this section we rewrite our results in the fashion employing the γ-average cost optimality equation in additive fashion. To this end, observe that for $i = 1, 2, \ldots, N$

$$V_i(\gamma, n) := \gamma^{-1} \ln U_i(\gamma, n)$$

equations for the minimal average cost take on the following form

$$\mathrm{e}^{\gamma\, V_i(\gamma, n+1)} = \min_{a \in \mathcal{F}_i} \left\{ \sum_{j \in \mathcal{I}} p_{ij}(a)\, \mathrm{e}^{\gamma (c_{ij}(a) + V_j(\gamma, n))} \right\} \tag{48}$$

Since $U_i(\gamma, n+1)/U_i(\gamma, n) = \mathrm{e}^{\gamma (V_i(\gamma, n+1) - V_i(\gamma, n))}$ and if $\bar{\mathbf{P}}^{(\gamma)}(\hat{f})$ is aperiodic we immediately conclude that for $i = 1, 2, \ldots, N$

$$J_{\min}(\gamma, n) = \min_{i \in \mathcal{I}} \{ V_i(\gamma, n+1) - V_i(\gamma, n) \} \text{ is nondecreasing in } n$$
$$J_{\max}(\gamma, n) = \max_{i \in \mathcal{I}} \{ V_i(\gamma, n+1) - V_i(\gamma, n) \} \text{ is nonincreasing in } n$$
$$g(\hat{f}) = \frac{1}{\gamma} \ln \rho(\hat{f}) = \lim_{n \to \infty} \{ V_i(\gamma, n+1) - V_i(\gamma, n) \}$$

The above facts can be explicitly formulated as

Value Iteration Method.

- *Step 0.* Select $V_i(\gamma, 0) > 0$ for $i = 1, 2, \ldots, N$.
- *Step 1.* Employing $V_i(\gamma, n) > 0$ with $i = 1, 2, \ldots, N$, update them to $V_i(\gamma, n+1) > 0$ using the recursive formula

$$e^{\gamma V_i(\gamma, n+1)} = \min_{a \in \mathcal{F}_i} \left\{ \sum_{j \in \mathcal{I}} p_{ij}(a)\, e^{\gamma(c_{ij}(a) + V_j(\gamma, n))} \right\} \quad (49)$$

  and calculate the values

$$\begin{aligned} \hat{J}_{\min}(\gamma, n) &:= \min_{i \in \mathcal{I}} \{ V_i(\gamma, n+1) - V_i(\gamma, n) \} \\ \hat{J}_{\max}(\gamma, n) &:= \max_{i \in \mathcal{I}} \{ V_i(\gamma, n+1) - V_i(\gamma, n) \} \end{aligned}$$

  being the upper and the lower bound on $\hat{J}$ that converge monotonically to $\hat{J}$.
- *Step 2.* If the difference $\hat{J}_{\max}(\gamma, n) - \hat{J}_{\min}(\gamma, n)$ is less than a given $\delta > 0$ then stop. The current stationary policies $\pi^{(n)} \sim f^{(n)}$ guarantees that $J(f^{(n)}) \in [\hat{J}_{\min}(\gamma, n); \hat{J}_{\max}(\gamma, n)]$.

This procedure overlaps results obtained in Cavazos-Cadena, Montes de Oca (MOR (2003), JAP(2005)).

In a quite similar fashion we can also rewrite the policy iteration algorithm.

Policy Iteration Method.

- *Step 0.* Select matrix $\bar{\mathbf{P}}^{(\gamma)}(f^{(0)})$ with $f^{(0)} \in \mathcal{F}$ such that the row sums are minimal, i.e. it holds $\bar{\mathbf{P}}^{(\gamma)}(f^{(0)}) \cdot \mathbf{e} \leq \bar{\mathbf{P}}^{(\gamma)}(f) \cdot \mathbf{e}$.
- *Step 1.* (Policy evaluation.) For a given stationary $\pi^{(n)} \sim f^{(n)}$ calculate the values

$$V_i(\gamma, f^{(n)}) + g(f^{(n)}) = \frac{1}{\gamma} \ln \left[\sum_{j \in \mathcal{I}} p_{ij}(f_i^n)\, \mathrm{e}^{\gamma(c_{ij}(f_i^n) + V_j(f^{(n)}))} \right]$$

- *Step 2.* (Policy improvement.) Using the values $V_i(\gamma, f^{(n)})$ in each state $i \in \mathcal{I}$ select action $f_i^{(n+1)} \in \mathcal{F}_i$ minimizing

$$H_i(\gamma, n) := \min_{a \in \mathcal{F}_i} \sum_{j \in \mathcal{I}} p_{ij}(a)\, \mathrm{e}^{\gamma(c_{ij}(a) + V_j(\gamma, f^{(n)}))}$$

and calculate

$$H_{\min}(\gamma, n) := \min_{i \in \mathcal{I}} \left\{ \frac{1}{\gamma} \ln[H_i(\gamma, n) - V_i(\gamma, n)] \right\}$$

$$H_{\max}(\gamma, n) := \max_{i \in \mathcal{I}} \left\{ \frac{1}{\gamma} \ln[H_i(\gamma, n) - V_i(\gamma, n)] \right\}$$

being the lower and upper bound on the optimal γ-average cost policy, as well as on the current policy $\pi^{(n)} \sim f^{(n)}$.

- *Step 3.* If $f^{(n+1)} = f^{(n)}$ then policy $\pi \sim f^{(n)}$ is an optimal policy and stop, else go to Step 1.

Thank you for your attention!