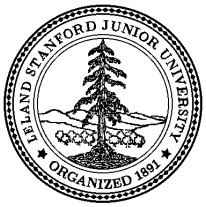
Lectures in Dynamic Programming and Stochastic Control

Arthur F. Veinott, Jr.



Spring 2008
MS&E 351 Dynamic Programming and Stochastic Control

Department of Management Science and Engineering
Stanford University
Stanford, California 94305

Contents

Homework Assignments

1

Discrete-Time-Parameter Finite Markov Population Decision Chains

1 FORMULATION

A *discrete-time-parameter finite Markov population decision chain* is a *system* that involves a finite population evolving over a sequence of periods labeled $1, 2, \ldots$ and over which one can exert some control. The system description depends on four data elements, viz., states, actions, rewards and transition rates. In each period N, each individual is in some *state* s in a set $\mathcal{S}$ of $S < \infty$ states. *The state summarizes all "relevant" information about the system history H as of period N.* Each individual in state s chooses an *action* a from a finite set $A_{sH} = A_s$ of possible actions, earns a *reward* $r(s, a, H) = r(s, a)$, and generates a (finite) expected[1] number $p(t \mid s, a, H)$ $= p(t \mid s, a) \geq 0$ of individuals in state $t \in \mathcal{S}$ in period $N+1$. Call $p(t \mid s, a)$ the *transition rate. The assumptions that A, r and p depend on H only through s are essential for a state s in a period to summarize all relevant information about the system history H as of the period.* Ob-

[1]For nearly all optimality concepts considered in the sequel, it suffices to consider only the expected numbers of individuals entering each state rather than the distribution of the actual random numbers of individuals entering each state. For that reason, those distributions are not considered explicitly here.

serve that the size of the population may vary over time. Also, there is no interaction among the individuals. Figure 1 illustrates this situation.

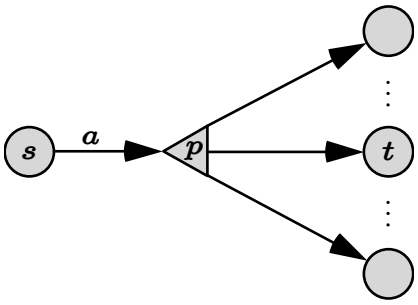


Figure 1

Call a system *stochastic* (resp., *substochastic*) if $\sum_t p(t \mid s,a) = 1$ (resp., ≤ 1) for each $a \in A_s$ and $s \in \mathcal{S}$. An important instance of a stochastic (resp., substochastic) system is that in which $p(t \mid s,a)$ is the probability that an individual in state s who takes action a in a period generates a single individual in state t.[2] Call a stochastic system *deterministic* if the $p(t \mid s,a)$ are all 0 or 1 because then for each s and a there will be a unique $t \in \mathcal{S}$ for which $p(t \mid s,a) = 1$ and $p(\tau \mid s,a) = 0$ for all $\tau \neq t$.

Examples of States and Actions in Various Applications. The table below gives examples of states and actions in several application areas.

Application	State	Action
Manage supply chain	Inventory levels of products	Choose product order times/quantities
Maintain road	Condition of road	Select resurfacing option
Invest in securities	Portfolio of securities	Buy/sell securities: times and amounts
Inspect lot	Number of defectives	Accept/reject lot, continue sampling
Route calls	Nodes in network	Send a call at one node to another
Control queueing network	Queue sizes at each station	Set service rates at each station
Overbook flight	Number of reservations	Book/decline reservation request
Market product	Goodwill	Advertise product
Manage reservoirs	Water levels at reservoirs	Release water from reservoirs
Insure asset	Risk category	Set policy premium
Patrol area	Car locations, service requests	Reposition cars
Guide rocket	Position and velocity	Choose retrorockets to fire
Care for a patient	Condition of patient	Conduct tests and treatments

In practice, it is often the case that the reward $r(s,a,T)$ that a substochastic system earns when an individual in a period takes action a in state s depends on s, a and an auxiliary random variable T whose conditional distribution given s, a and the system history at that time

[2]The last condition rules out a system in which an individual in a period generates more than one individual in the next period even though such a system may still be stochastic (resp., substochastic).

depends only on s and a. To reduce this problem to the one above, it suffices to let $r(s,a) = \mathrm{E}(r(s,a,T) \,|\, s,a)$ be the conditional expected reward that the system earns when it is in state s and takes action a therein. For example, if T is the next state that the system visits after taking action a in state s, then $r(s,a) = \sum_t r(s,a,t)p(t \,|\, s,a)$.

2 MAXIMUM N-PERIOD VALUE: RECURSION AND EXAMPLES

Why Maximize Expected Finite-Horizon Rewards? The most common goal in the above setting is to find a "policy", i.e., a rule that specifies the action to take in each state with N or less periods to go, that maximizes the expected N-period reward where there is a given terminal value of being in each state with no periods to go. Finite-horizon optimality concepts like this require users to specify the terminal value, though it is often difficult to do. While decision makers are generally not interested in a fixed number of periods, this approach is often used for several reasons.

- **Simplicity.** Though fixing a particular finite horizon is often rather arbitrary, the concept is simple. By contrast, optimality concepts for an infinite horizon—perhaps the main alternative—are more subtle and varied.

- **Realism.** Optimality over a finite horizon is often more realistic. Users frequently think that they can specify the terminal value well enough to facilitate making good decisions in earlier periods and are comfortable with finite horizons. This way, they do not have to make explicit projections beyond the finite horizon—which they might view as speculative anyway—though of course these projections must instead be reflected in the terminal-value function. Few users are prepared to think about the indefinite future of which their own lives occupy such a minuscule part.

- **Extension to Nonstationary Data.** Optimality concepts over a finite horizon generally adapt easily to nonstationary data without any increase in computational complexity. By contrast, the computations needed to assure infinite-horizon optimality rise rapidly with the extent of nonstationarity. This is an important advantage of finite-horizon concepts because nonstationarity is common in life.

Maximum N-Period Value. We begin by examining this problem informally. Let V_s^N be the maximum expected N-period reward, called the *N-period value*, that an individual (and his progeny) can earn starting from state s, assuming for the moment that this maximum is achieved. Now if an individual chooses an action $a \in A_s$ in the first period and the individual's progeny use an "optimal" policy in the remaining $N-1$ periods, then

$$V_s^N \geq \underbrace{r(s,a)}_{\substack{\text{reward in} \\ \text{first period}}} + \underbrace{\sum_{t \in \mathcal{S}} p(t \,|\, s,a) V_t^{N-1}}_{\substack{\text{expected reward in remaining } N-1 \text{ periods} \\ \text{when an optimal policy is used therein}}}.$$

Moreover, if the individual chooses an "optimal" action a in the first period, then equality occurs above. This idea, called the *principle of optimality,* implies the *dynamic-programming recursion*:

$$V_s^N = \max_{a \in A_s} [r(s,a) + \sum_{t \in \mathcal{S}} p(t \,|\, s,a) V_t^{N-1}] \tag{1}$$

for $s \in \mathcal{S}$ and $N = 1, 2, \ldots,$ where V_s^0 is the given *terminal value* in state s. This recursion permits one to calculate $\{V_s^1\}$, then $\{V_s^2\}$, then $\{V_s^3\}$, and so on. Once V_s^N is computed, one optimal action a_s^N in state s with N periods to go is simply any action in A_s that achieves the maximum on the right-hand side of (1). Thus, when there are N periods to go, each individual in state s can be assumed to take the same action without loss of optimality. *The recursion* (1) *is of fundamental importance in a broad class of applications.*

Minimum N-Period Cost. If $r(s, a)$ is instead a "cost" and the aim is to minimize expected N-period cost, then the "max" in (1) should be replaced by "min". Then V_s^N is the minimum expected N-period cost starting from state s. This alternate form of the dynamic-programming recursion appears often in the sequel with C_s^N replacing V_s^N.

Deterministic Case. In the deterministic case, there may be several actions that take an individual from state s to t. In that event, it is best to choose one with maximum reward and eliminate the others. Then one can identify actions in state s with states t visited next, so (1) simplifies to

$$V_s^N = \max_{t \in A_s} [r(s, t) + V_t^{N-1}] \tag{2}$$

for $s \in \mathcal{S}$ and $N = 1, 2, \ldots$. As Figure 2 illustrates, V_s^N can be thought of as the maximum N-period reward that an individual can earn in traversing an N-step chain that begins in state s and earns a terminal reward V_t^0 when it ends in state t.

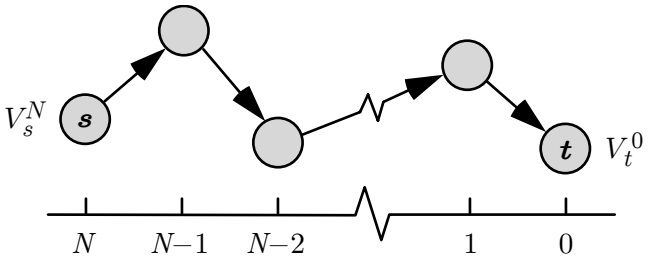


Figure 2

Examples

1 Minimum-Cost Chain. The minimum-cost-chain problem described above has a terminal cost in each state. One specialization of this problem is to find a minimum-cost chain from every state to a given terminal state τ in N steps or less. Thus if $c(s, t)$ is the cost of moving from state s to state t in one step and C_s^N is the minimum N-step-or-less cost of moving from state s to state τ, then $C_\tau^N = 0$ for $N \geq 1$, $C_s^1 = c(s, \tau)$ and

$$C_s^N = \min_{t \in A_s} [c(s, t) + C_t^{N-1}],\ N = 2, 3, \ldots \text{ and } s \in \mathcal{S} \setminus \{\tau\}. \tag{3}$$

If the associated (directed) graph $\mathcal{G} \equiv (\mathcal{S}, \mathcal{A})$ with node set $\mathcal{S}$ and arc set $\mathcal{A} \equiv \{(s,t) : t \in A_s\}$ has no *circuit* (i.e., directed cycle) around which the total cost incurred is negative, then

$$C_s^N = C_s^{S-1} \text{ for } s \in \mathcal{S} \setminus \{\tau\} \text{ and } N \geq S-1. \tag{4}$$

This is because a minimum-cost chain from s to τ need not visit a node twice. For if it did, the chain could be shortened and its cost reduced by eliminating the circuit created by visiting the indicated node twice as Figure 3 illustrates. Thus a minimum-cost chain from any node in $\mathcal{S} \setminus \{\tau\}$ to τ can be assumed to have $S-1$ arcs or less, justifying (4). If one takes care to choose the minimizer $t = t_s^N$ in (3) so that $t_s^N = t_s^{N-1}$ whenever $C_s^N = C_s^{N-1}$, then the $t_s^N = t_s$ will be independent of $N \geq S-1$ by (4). Also, the subgraph of $\mathcal{G}$ with arcs (s, t_s), $s \in \mathcal{S} \setminus \{\tau\}$, is a tree with the unique simple chain therein from each node $s \neq \tau$ to τ being a minimum-cost chain from s to τ.

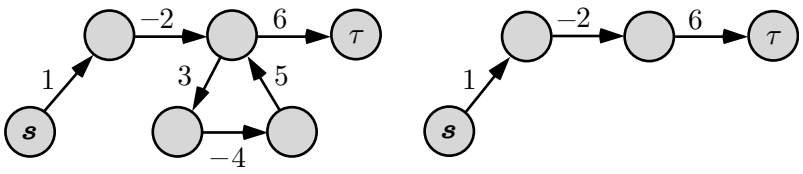


Figure 3

Computational Effort (or Complexity). How much computation is required to calculate $C_s \equiv C_s^{S-1}$ for all $s \in \mathcal{S} \setminus \{\tau\}$? Let A be the number of arcs in $\mathcal{A}$. Then for each N, evaluation of (3) requires A additions and nearly the same number of comparisons. Since this must be done for each $N = 2, \ldots, S-1$, the total number of additions and comparisons is about $(S-2)A$.

2 Supply Management. One of the areas in which dynamic programming has been used most widely is to find optimal supply-management policies. The reason for this is that nearly all firms carry inventories and the investment in them is sizable. For example, US manufacturing and trade inventories alone in 2000 were 1.205 trillion dollars, or 12% of the entire US gross national product of 9.963 trillion dollars that year![3]

As an illustration of the role of dynamic programming in such problems, suppose that the demands for a single product in successive periods are independent and identically-distributed random variables. At the beginning of a period, the supply manager observes the *initial stock* s, $0 \leq s \leq S$, and orders a nonnegative amount with immediate delivery bringing the *starting stock* to a, $s \leq a \leq S$. If the demand D in the period exceeds a, the excess demand $D-a$ is lost. There is an ordering cost $c(a-s)$ and a holding and penalty cost $h(a-D)$ in the period. Let C_s^N be the minimum expected N-period cost starting with the initial stock s. Then ($C_\cdot^0 \equiv 0$)

[3]2001 *Statistical Abstract of the United States,* Table 756 and 640.

$$C_s^N = \min_{s \le a \le S} [c(a-s) + \mathrm{E}h(a-D) + \mathrm{E}C_{(a-D)^+}^{N-1}]$$

for $0 \le s \le S$ and $N = 1, 2, \ldots$.

3 Exercising a Call Option. One of the most active places in which dynamic programming is used today is Wall Street. To illustrate, consider the problem of determining when to exercise an (American) *call option* to buy a stock ignoring commissions. The option gives the purchaser the right to buy the stock at the *strike price* $s^* \ge 0$ on any of the next N days.

Two questions arise. When should the option be exercised? What is its value? To answer these questions requires a stock-price model and a dynamic-programming recursion to find the value of the option as well as an optimal option-exercise policy.[4]

Consider the following stock-price model. Suppose that the stock price is s on *day* $N \ge 0$, i.e., N days before the option expires. If $N > 0$, assume that the stock price on the following day $N-1$ is sR_N where $R_1, R_2, \ldots$ are independent identically distributed nonnegative random variables with the same distribution as a nonnegative random variable R. Then, $r \equiv R - 1$ is the rate of return for a day, and $\mathrm{E}r$ is the expected rate of return that day.

Let V_s^N be the *value of the option* on day N when the market price of the stock is s. There are two alternatives that day. One is to exercise the option to buy the stock at the strike price and immediately resell it, which earns $s - s^*$. The other is not to exercise the option that day, in which case the maximum expected income in the remaining $N-1$ days is $\mathrm{E}V_{sR}^{N-1}$. Since one seeks the alternative with higher expected future income, the value of the option on expiration day is $V_s^0 = (s-s^*)^+$ and on day $N > 0$ is given recursively by

$$V_s^N = \max(s - s^*, \mathrm{E}V_{sR}^{N-1}), N = 1, 2, \ldots. \tag{5}$$

Thus it is optimal to exercise the option on expiration day if $s - s^* > 0$, and not do so otherwise. And it is optimal to exercise the option on day $N > 0$ if $s - s^* > \mathrm{E}V_{sR}^{N-1}$, and not do so otherwise.

Nonnegative Expected Rate of Return. Consider now the question when it is optimal to exercise the option. It turns out that as long as the expected rate of return is nonnegative, i.e., $\mathrm{E}r \ge 0$ or equivalently $\mathrm{E}R \ge 1$, the answer is to wait until the expiration day. To establish this fact, it suffices to show that

$$V_s^N = \mathrm{E}V_{sR}^{N-1} \tag{6}$$

[4]This formulation of the problem of when to exercise an (American) call option addresses the situation faced by an investor who wishes to profit from speculation on the market price of a stock. By contrast, the formulation of the problem in finance addresses the problem faced by an institution who wishes to price an option to avoid market risk and rely on commissions for profits. Though the assumptions differ, the computations are similar.

for each $N > 0$ and all $s \geq 0$. To that end, observe from (5) for $N > 1$ and by definition for $N = 1$ that $V_s^{N-1} \geq s - s^*$ for each $s \geq 0$. Thus because $\mathrm{E}R \geq 1$, it follows that $\mathrm{E}V_{sR}^{N-1} \geq \mathrm{E}(sR - s^*) \geq s - s^*$. Hence from (5) again, (6) holds. Hence, *if the expected rate of return is nonnegative, it is optimal not to exercise the option at any market price when $N > 0$ days remain until expiration.* Furthermore, $V_s^N = \mathrm{E}(sR^N - s^*)^+$ is the value of the option where $R^N \equiv \prod_1^N R_i$, so sR^N is the price at expiration.

Negative Expected Rate of Return. Suppose now that the expected rate of return is negative, i.e., $\mathrm{E}r < 0$. To analyze equation (5), it turns out to be useful to subtract $s - s^*$ from both sides of (5) and make the change of variables $U_s^N = V_s^N - (s - s^*)$. Then (5) reduces to the equivalent system

$$U_s^N = \max(0, s\mathrm{E}r + \mathrm{E}U_{sR}^{N-1}) \tag{5}'$$

for $N = 1, 2, \ldots$ where $U_s^0 = (s^* - s)^+$. Now we claim that U_s^N is decreasing and continuous in $s \geq 0$ for each N and $\lim_{s\to\infty} U_s^N = 0$. Certainly that is so for $N = 0$. Suppose it is so for $N-1$ and consider N. Then $\mathrm{E}U_{sR}^{N-1}$ is decreasing and continuous in s and approaches 0 as $s \to \infty$. Consequently, there is a smallest $s = s_N \geq 0$ such that $s\mathrm{E}r + \mathrm{E}U_{sR}^{N-1} \leq 0$. Thus, it follows that the maximum on the right side of (5)′ is $s\mathrm{E}r + \mathrm{E}U_{sR}^{N-1}$ if $s < s_N$ and 0 if $s \geq s_N$. Since (5) and (5)′ are equivalent, it follows that this rule is optimal with (5) as well, i.e., it is optimal to wait if $s < s_N$ and to exercise the option if $s \geq s_N$. In short, *if the expected rate of return is negative and if N days remain until expiration of the option, then there is a price limit s_N such that it is optimal to wait if the price is below s_N and to exercise the option if the price is s_N or higher.*

4 System Reliability. A system consists of a finite set Σ of components. The system is observed once a period and each component is found to be "working" or "failed." If s is the subset of components that is observed to be working in some period, then a subset $t \setminus s$ of the failed components may be replaced at a cost $r_{t\setminus s}$ where t ($\supseteq s$) is the set of working components after replacement. The expected operating cost incurred during the period is then p_t. Some components may fail during the period with the conditional distribution of the random set w of working components in the next period, given that t is the working set after replacement in the period and given the past history, depending on t, but not otherwise on the past history. Let C_s^N be the minimum expected N-period cost when s is the initial working set. Then ($C_\cdot^0 \equiv 0$),

$$C_s^N = \min_{s\subseteq t\subseteq\Sigma} [r_{t\setminus s} + p_t + \mathrm{E}(C_w^{N-1} \mid t)]$$

for $\emptyset \subseteq s \subseteq \Sigma$ and $N = 1, 2, \ldots$.

5 Maximum Expected N-Period Instantaneous Rate of Return: Portfolio Selection. Suppose that in a stochastic system, the return the system earns in periods $1, \ldots, N$ is the product $R_1 \cdots R_N$ of the nonnegative random returns $R_1, \ldots, R_N$ in those periods. For example, if the rate of return in period i is 15%, then the return in period i is $R_i = 1.15$. Now let $100\rho_N\%$ be the (random) instantaneous rate of return per period over the N periods assuming continuous compounding. Then $e^{\rho_N N} = R_1 \cdots R_N$, so

$$\rho_N = \frac{1}{N}\,[\ln R_1 + \cdots + \ln R_N]. \tag{7}$$

Now since $\ln R_i$ is the instantaneous rates of return in period i, the problem of maximizing the expected instantaneous rate of return over N periods reduces to maximizing the sum $\mathrm{E}\ln R_1 + \cdots + \mathrm{E}\ln R_N$ of the expected instantaneous rates of return in those N periods.

Now assume that the conditional distribution of R_i given that the system starts in state s in period i and takes action $a \in A_s$ in that period, and given the past history, is independent of i and of the past history. Then the corresponding conditional expected value $r(s,a)$ of $\ln R_i$ is independent of i.

Note that a necessary condition for $r(s,a)$ to be finite in this example is that the conditional probability that R_i is zero given that the system starts in state s in period i and takes action $a \in A_s$ in that period is zero. The reason is that a zero return corresponds to an instantaneous rate of return equal to $-\infty$ and so is to be avoided at all costs if the goal is to maximize the expected instantaneous rate of return.

Observe that maximizing the expected instantaneous rate of return is different from maximizing the instantaneous rate of expected return. The former is equivalent to maximizing the expected value of the sum of the logarithms of the returns whereas the later entails maximizing the expected value of the product of the returns. Thus, the former involves additive rewards while the latter involves multiplicative rewards. For example, if you have the opportunity to invest in a security whose return R in a period is 0 or 3, each with equal probability, then $\mathrm{E}\ln R = -\infty$ and $\mathrm{E}R = 1.5$. Thus, the expected instantaneous rate of return is $-\infty$ whereas the instantaneous rate of expected return is 50%.

Optimal Portfolio Selection. A simple example of the above development is portfolio selection. Suppose that the returns that various securities earn depend on the market state $s \in \mathcal{S}$, e.g., market index, earnings, interest rates, time, etc. The actions available in market state s is a set A_s of several portfolios of securities. The one-period return of a portfolio a in market state s is a random variable whose conditional distribution given s, a and the past history depends only on s and a. Denote by $r(s,a)$ the corresponding conditional expected value of the instantaneous rate of return. Finally the probability that the market state in a period is t given that the market state

in the prior period is s and that a portfolio a is chosen at that time, and given the history of the process, depends only on s and t. This reflects the fact that a single portfolio manager generally does not influence the market. Thus, $p(t \mid s, a) = p(t \mid s)$. Let V_s^N be the maximum expected value of the instantaneous rate of return over N periods when the market state is initially s. Then (1) holds with $p(t \mid s)$ replacing $p(t \mid s, a)$. Consequently, since an action in A_s maximizes the right-hand side of (1) if and only if it maximizes $r(s, a)$ because the term $\sum_t p(t \mid s)V_t^{N-1}$ is independent of a. This means that the optimal policy is *myopic*, i.e., maximizes the (expected) reward in each period alone without regard for the future. Thus, in the present case, *it suffices to separately maximize the expected instantaneous rate of return in each period.*

3 MAXIMUM EXPECTED N-PERIOD UTILITY WITH CONSTANT RISK POSTURE

[HM72], [Ar65], [Ro75a]

Expected Utility and Risk Aversion/Preference

Expected Utility. A decision maker's preferences among *gambles*, which we take to be random N-vectors, can often be expressed by a *utility function*, i.e., a real-valued function u defined on the range (assumed in $\Re^N$) of the set of gambles. When this is so, a decision maker who has a choice between two gambles will *prefer* one with higher expected utility, i.e., if X and Y are gambles, then the decision maker prefers X to Y if $\mathrm{E}u(X) \geq \mathrm{E}u(Y)$. Eminently plausible axioms implying that a decision maker's preferences can be represented by a utility function are discussed in books on the foundations of decision theory.

Risk Aversion/Preference. A decision maker whose preferences can be represented by a utility function u is called a *risk averter* (resp., *risk preferrer*) if $\mathrm{E}u(X) \leq u(\mathrm{E}X)$ (resp., $\mathrm{E}u(X) \geq u(\mathrm{E}X)$) for all gambles X with finite expectations, i.e., the decision maker prefers a certain (resp., an uncertain) gamble to an uncertain (resp., a certain) one having the same expected value. As examples, managers and investors are usually risk averters as are buyers of insurance. By contrast, those who gamble at casinos and at race tracks are usually risk preferrers. Of course, a risk averter may still prefer an uncertain gamble to a certain one if the former has higher expectation and u is increasing (which is usually the case). Indeed, in that event only such gambles will be of interest if they are real valued. For if Y is a constant random variable and X is an uncertain one that is preferred to Y, then $u(Y) \leq \mathrm{E}u(X) \leq u(\mathrm{E}X)$, whence $Y \leq \mathrm{E}X$. That is why investors do not exhibit much interest in risky securities whose expected returns are less than those available on safe ones.

A decision maker is a risk averter (resp., preferrer) if and only if u is concave (resp., convex). To see this, observe first that it suffices to establish the claim for a risk averter, since if u is the utility function for a risk preferrer, $-u$ is the utility function for a risk averter. To show the "only if" part, observe that for all N-vectors x, y and nonnegative numbers p, q with $p + q = 1$,

a risk averter would rather receive the certain gamble $px + qy$ than the gamble that yields x with probability p and y with probability q (and so has the same expected value), i.e., $u(px + qy) \geq pu(x) + qu(y)$, which is the definition of concavity. The "if part" is known as Jensen's inequality and may be proved as follows. Suppose X is a random N-vector with finite expectation $\mathrm{E}X$ and that u is concave. Then since u is concave, there is a $d \in \Re^N$ (the supergradient of u at $\mathrm{E}X$ or the gradient there if it exists) such that $u(x) \leq u(\mathrm{E}X) + d(x - \mathrm{E}X)$ for all $x \in \Re^N$. Thus $u(X) \leq u(\mathrm{E}X) + d(X - \mathrm{E}X)$, so on taking expected values, $\mathrm{E}u(X) \leq u(\mathrm{E}X)$.

Examples. Suppose $\lambda = (\lambda_i)$, $x = (x_i) \in \Re^N$. For $x \gg 0$, i.e., $x_i > 0$ for each i, let $x^\lambda \equiv x_1^{\lambda_1} \cdots x_N^{\lambda_N}$ and $\ln x \equiv (\ln x_i)$. Then the functions λx, $-e^{\lambda x}$ and, for $\lambda \geq 0$ and $x \gg 0$, $\ln x^\lambda = \sum_i \lambda_i \ln x_i$ exhibit risk aversion, while the first and the negatives of the last two exhibit risk preference. For $x \gg 0$, the functions $\pm x^\lambda$, and for $0 \not\leq \lambda \not\leq 0$, $\ln x^\lambda$ generally do not exhibit either risk aversion or risk preference.

General (even concave) utility functions are usually too complex to permit one to do the computations needed to choose between gambles. For that reason, it is of interest to consider more tractable utility functions with special structures that arise naturally in applications. One such class of utility functions that is tractable in dynamic programming is that with *constant risk posture*, i.e., for which one's posture towards a class of gambles is independent of one's level of wealth. We now discuss this concept for the classes of additive and multiplicative gambles.

Constant Additive Risk Posture

A decision maker has *constant additive risk posture* if his posture towards every additive gamble Y is independent of his wealth $x \in \Re^N$, i.e., $\mathrm{E}u(x + Y) - u(x)$ has constant sign in x whenever $\mathrm{E}u(x + Y)$ has finite expected value for every x. This hypothesis is plausible for large firms whose total assets are much larger than the investments they consider. In any case, the hypothesis is satisfied if and only if, apart from an affine transformation $au + b$ of u, u has the form

$$u(x) = e^{\lambda x} \quad \text{for all } x \tag{1}$$

or

$$u(x) = \lambda x \quad \text{for all } x \tag{2}$$

for some row vector $\lambda \in \Re^N$. For the "if" part of the above result, observe that the sign of

$$\mathrm{E}u(x + Y) - u(x) = \begin{cases} e^{\lambda x}[\mathrm{E}e^{\lambda Y} - 1], & \text{if (1) holds} \\ \lambda \mathrm{E}Y, & \text{if (2) holds} \end{cases}$$

is independent of x. Notice that $au + b$ is concave (resp., convex) if (1) holds and $a \leq 0$ (resp., $a \geq 0$), in which case the decision maker is a constant additive risk averter (resp., preferrer).

Constant Multiplicative Risk Posture

Alternately, a decision maker has *constant multiplicative risk posture* if his posture toward any positive multiplicative gamble $Y \gg 0$ is independent of his positive wealth $x \in \Re^N$, i.e., $\mathrm{E}u(x \circ Y) - u(x)$ has constant sign in $x \gg 0$ whenever $\mathrm{E}u(x \circ Y)$ has finite expected value for all $x \gg 0$ where $x \circ Y \equiv (x_i Y_i)$. In most situations this hypothesis seems more reasonable for a wide range of wealth levels than does constant additive risk posture. In any case, the hypothesis is satisfied if and only if, apart from an affine transformation $au + b$ of u, u has the form

$$u(x) = x^\lambda \qquad \text{for all } x \gg 0 \tag{3}$$

or

$$u(x) = \ln x^\lambda \quad \text{for all } x \gg 0 \tag{4}$$

and some $\lambda \in \Re^N$. This result follows from that for the additive case on making the change of variables $x' = \ln x$ and $Y' = \ln Y$, and defining u' by the rule $u'(x') = u(x)$. *Then $u(x \circ Y) = u'(x' + Y')$, so u exhibits constant multiplicative risk posture if and only if u' exhibits constant additive risk posture.* In particular,

$$u(x) = x^\lambda \qquad \text{if and only if} \quad u'(x') = e^{\lambda x'},$$

while

$$u(x) = \ln x^\lambda \quad \text{if and only if} \quad u'(x') = \lambda x'.$$

Additive and Multiplicative Utility Functions

A utility function $u(r)$ of rewards $r = (r_1, \ldots, r_N)$ earned in periods $1, \ldots, N$ that exhibits additive or multiplicative risk posture is either *additive*, i.e.,

$$u(r) = u_1(r_1) + \cdots + u_N(r_N) \tag{5a}$$

for some functions $u_1, \ldots, u_N$ on the real line, or *multiplicative*, i.e.,

$$u(r) = \pm\, u_1(r_1) \cdots u_N(r_N) \tag{5b}$$

for some nonnegative functions $u_1, \ldots, u_N$ on a subset of the real line. The situation may be summarized as follows.

- **Constant Additive Risk Posture.** Suppose $u(r)$ exhibits constant additive risk posture. Then up to a positive affine transformation, either $u(r) = \lambda r$ is additive with $u_i(v) = \lambda_i v$ or $u(r) = \pm\, e^{\lambda r}$ is multiplicative with $u_i(v) = e^{\lambda_i v}$. *Thus, the only strictly risk-adverting (resp., -preferring) constant-additive-risk-posture utility functions are multiplicative and exponential.*

- **Constant Multiplicative Risk Posture.** Suppose $u(r)$ exhibits constant multiplicative risk posture on the positive orthant. Then, up to a positive affine transformation, either $u(r) = \ln r^\lambda = \lambda \ln r$ is additive with $u_i(v) = \lambda_i \ln v$ or $u(r) = \pm\, r^\lambda$ is multiplicative with $u_i(v) = v^{\lambda_i}$ for $v > 0$. *Thus, the only strictly risk-averting (resp., -preferring) constant-multiplicative-risk-posture utility functions are additive and logarithmic.*

Maximum Expected N-Period Symmetric Multiplicative Utility

Now consider an S-state stochastic system that consists of a single individual with state-space $\mathcal{S}$, action sets A_s, transition probabilities $q(t \mid s, a)$ and, for this paragraph only, one-period rewards $r(s, a, t)$ that depend not only on s and a, but also on t. Suppose also that the utility $u(r)$ of the rewards $r = (r_1, \ldots, r_N)$ earned in periods $1, \ldots, N$ is additive or multiplicative, i.e., (5a) or (5b) holds. For simplicity, also assume that u is *symmetric*, i.e., $u_1 = \cdots = u_N = \widehat{u}$, say. When u is additive, the dynamic-programming recursion (1) in §1.2 applies directly by simply replacing the one-period rewards $r(s, a, t)$ by their utilities $\widehat{u}(r(s, a, t))$. Now consider the case when u is multiplicative. Then let V_s^N be the maximum expected N-period utility starting from state s. Hence, since $u(r_1, \ldots, r_N) = \widehat{u}(r_1)u(r_2, \ldots, r_N)$ for $N \geq 2$ and $u(r_1) = \pm\, \widehat{u}(r_1)$ where $\widehat{u} \geq 0$, one sees that

$$V_s^N = \max_{a \in A_s} \sum_{t \in \mathcal{S}} \widehat{u}(r(s, a, t))q(t \mid s, a)V_t^{N-1}$$

for $s \in \mathcal{S}$ and $N = 1, 2, \ldots$ where $V^0 \equiv \pm\, 1$ (1 is an S-vector of ones). Thus,

$$V_s^N = \max_{a \in A_s} \sum_{t \in \mathcal{S}} p(t \mid s, a)V_t^{N-1} \tag{6}$$

for $s \in \mathcal{S}$ and $N = 1, 2, \ldots$ where for each $a \in A_s$ and $s, t \in \mathcal{S}$,

$$p(t \mid s, a) \equiv \widehat{u}(r(s, a, t))q(t \mid s, a). \tag{7}$$

Observe that $\sum_{t \in \mathcal{S}} p(t \mid s, a)$ may exceed one even though $\sum_{t \in \mathcal{S}} q(t \mid s, a) = 1$. Thus (7) is an instance of branching and maximizing (resp., minimizing) the size of the expected total population in all states at the end of N periods from each initial state where $V^0 = 1$ (resp., $V^0 = -1$).

Maximum N-Period Instantaneous Rate of Expected Return.

The above development also applies to the problem in which the goal is to maximize the instantaneous rate of expected return. In that event, $\widehat{u}(r) = r$, so (7) simplifies to

$$p(t \mid s, a) \equiv r(s, a, t)q(t \mid s, a). \tag{7}'$$

4 MAXIMUM VALUE IN CIRCUITLESS SYSTEMS

In the preceding two subsections we studied the problem of maximizing the N-period value. In many circumstances one is interested in earning the maximum value over an infinite horizon. Problems of this type are not generally well posed because the sum of the expected rewards in periods $1, 2, \ldots$ may diverge.

The simplest situation in which this difficulty does not arise is that in which the system is "circuitless". To describe this concept, it is useful to introduce the *system graph* $\mathcal{G}$, i.e., the di-

rected graph whose nodes are the states and whose arcs are the ordered pairs (s,t) of states for which $p(t \mid s,a) > 0$ for some $a \in A_s$. Call the system *circuitless* if the system graph has no circuits, i.e., there is no sequence of states that begins and ends with the same state and for which each successive ordered pair (s,t) of states in the sequence is an arc of $\mathcal{G}$.

For example, the N-period problems of §1.2 and §1.3 are circuitless systems. To see this, include the period in the state so the state-space for the N-period problem consists of the pairs $(s,n) \in \mathcal{S} \times \mathcal{N}$ where $\mathcal{N} = \{1,\ldots,N\}$ is the set of periods. Then since no period can be revisited, the system is circuitless.

In circuitless systems, it is possible to relabel the states as $1,\ldots,S$ so that each state is accessible only to states with higher numbers. Consequently, the number of transitions before the population disappears is at most $|\mathcal{S}| - 1 = S - 1$ because no state can be revisited. Thus, the maximum value V_s starting from state s is finite because it is a sum of at most S finite terms. Then by an argument like that used to justify (1) of §1.2, one sees that V_s satisfies

$$V_s = \max_{a \in A_s} [r(s,a) + \sum_{s<t} p(t \mid s,a) V_t],\ s \in \mathcal{S}. \tag{1}$$

The V_s may then be calculated recursively in the order $V_S,\ldots,V_1$.

Examples

1 Knapsack Problem (Capital Budgeting). The problem of choosing items to fill a knapsack to maximize total value, or of allocating limited capital to projects to maximize revenue, are both instances of an important problem called the *knapsack* problem. To describe the problem, let C be a finite set of positive integers representing amounts of capital investment required for each of a group of projects. Let r_c be the projected revenue when c is invested and S be the (positive) capital available. The problem can be posed as the integer program of finding an integer vector $x = (x_c)$ that

$$\text{maximizes} \sum_{c \in C} r_c x_c$$

subject to

$$\sum_{c \in C} c x_c = S \text{ and } x \geq 0.$$

(This formulation allows replication of projects; if this is impossible, then assume $x \leq 1$.) Also assume $1 \in C$, so the knapsack problem is feasible for $S = 0,1,\ldots$. Let $\mathcal{S} = \{0,\ldots,S\}$ be the state-space. The actions available in state s are the investments $c \in C$ for which $c \leq s$. Since each investment reduces the amount of capital available for subsequent investments, the system is circuitless. Let V_s be the maximum revenue that can be earned when the capital available is s. Now $V_0 \equiv 0$.

A Dynamic-Programming Recursion. One dynamic-programming recursion for this problem is

$$V_s = \max_{\substack{c\in C\\ c\le s}} [r_c + V_{s-c}]$$

for $s = 1, \dots, S$. Of course, V_S is the desired maximum revenue, and $\Big(\frac{V_S}{S} - 1\Big)100\%$ is the internal rate of return on the invested capital S assuming that all revenue is received at a common time.

An Alternate Dynamic-Programming Recursion. Dynamic-programming recursions for solving a problem are not unique. For example, the V_s may also be determined from the alternate dynamic-programming (branching) recursion

$$V_s = \max_{\substack{u+v=s\\ u,v>0}} [V_u + V_v] \vee r_s,\ s \in C$$

and

$$V_s = \max_{\substack{u+v=s\\ u,v>0}} [V_u + V_v],\ s \notin C.$$

2 Portfolio Selection (Manne). In each period s, $1 \le s < S$, there is a set A_s of securities available for investment. One dollar invested in security $a \in A_s$ in period s generates $p(t \,|\, s, a) \ge 0$ dollars in period $t = s{+}1, \dots, S$. Of course $\sum_{t=s+1}^{S} p(t \,|\, s, a)$ normally exceeds one, so the system is branching. Then the state-space is $\mathcal{S} = \{1, \dots, S\}$. Since no period can be revisited, the system is circuitless. The goal is to find, for each s, the maximum income V_s that can be earned by period S from each dollar invested in period s. Then $V_S = 1$ and

$$V_s = \max_{a\in A_s} \sum_{t=s+1}^{S} p(t \,|\, s, a) V_t,\ s = 1, \dots, S{-}1.$$

In this case, the internal rate of return on capital invested in period $s < S$ is $((V_s)^{\frac{1}{S-s}} - 1)100\%$.

5 DECISIONS, POLICIES, OPTIMAL-RETURN OPERATOR

An informal definition of a "policy" appears on page 3. It is now time to make that concept precise. At the same time, we introduce matrix and operator notation to simplify the development. A *decision* is a function δ that assigns to each state $s \in \mathcal{S}$ an action $\delta^s \in A_s$. Thus $\Delta \equiv \times_{s\in\mathcal{S}} A_s$ is the *set of decisions*. A *policy* is a sequence $\pi = (\delta_1, \delta_2, \dots)$ of decisions. Using $\pi = (\delta_N)$ means that if an individual is in state s in period N, then the individual uses action δ_N^s in that period. Now $\Delta^\infty \equiv \Delta \times \Delta \times \cdots$ is the *set of policies*. Finally, call a policy $\delta^\infty = (\delta, \delta, \dots)$ *stationary* if it uses the same decision δ in each period.

For any $\delta \in \Delta$, let r_δ be the S-element column vector whose s^{th} element is $r(s, \delta^s)$, i.e., r_δ is the one-period *reward vector* when using the decision δ. Also let P_δ be the $S \times S$ transition matrix whose st^{th} element is $p(t \,|\, s, \delta^s)$, i.e., P_δ is the (*one-step*) *transition matrix* using δ. If $\pi = (\delta_1, \delta_2, \dots)$ is a policy, let $P_\pi^N \equiv P_{\delta_1} \cdots P_{\delta_N}$ be the N-step transition matrix using π ($P_\pi^0 \equiv I$). Thus $P_\delta^N \equiv P_{\delta^\infty}^N = (P_\delta)^N$. Observe that the st^{th} element of P_π^N is the expected number of indi-

viduals in state t in period $N+1$ that one individual in state s in period one and his progeny generate when they use π. Then ($V_\pi^0 \equiv 0$) the S-element column vector

$$V_\pi^N \equiv \sum_{i=1}^{N} P_\pi^{i-1} r_{\delta_i} \tag{1}$$

is the *expected N-period reward* or *N-period value* of π.

Define the *optimal-return operator* $\mathcal{R} : \Re^S \to \Re^S$ by

$$\mathcal{R}V \equiv \max_{\delta \in \Delta} [r_\delta + P_\delta V],\ V \in \Re^S. \tag{2}$$

Observe that $\mathcal{R}V$ is the maximum expected one-period reward when V is the terminal reward in that period. Of course $\mathcal{R}^0 V = V$, $\mathcal{R}^1 V = \mathcal{R}V$, $\mathcal{R}^2 V = \mathcal{R}(\mathcal{R}V)$, etc. It is important to recognize that the maximum in (2) is attained; indeed the s^{th} element $(\mathcal{R}V)_s$ of $\mathcal{R}V$ is

$$(\mathcal{R}V)_s = \max_{a \in A_s} [r(s,a) + \sum_{t \in \mathcal{S}} p(t \,|\, s,a) V_t]$$

for all $s \in \mathcal{S}$, i.e., the maximum in (2) is taken coordinatewise. Also note that $\mathcal{R}V$ is increasing in V, a fact that we will use often in the sequel.

6 MAXIMUM N-PERIOD VALUE: FORMALITIES

Our goal now is to justify the dynamic-programming recursion (1) of §1.2 in a more formal way. To that end, let $V_\pi^N(u) \equiv V_\pi^N + P_\pi^N u$ be the N-period value when using π and the *terminal value* vector at the end of period N is $u \in \Re^S$.

Proposition 1. Maximum N-Period Value. $\mathcal{R}^N u = \max_{\pi \in \Delta^\infty} V_\pi^N(u)$, $N = 0, 1, \ldots$ *and* $u \in \Re^S$.

Proof. The result is trivial for $N = 0$. Suppose it holds for $N - 1 \geq 0$ and consider N. Now

$$r_\delta + P_\delta V_\pi^{N-1}(u) = V_{\delta\pi}^N(u)$$

for all δ, so because $P_\delta \geq 0$,

$$\mathcal{R}^N u = \mathcal{R}(\mathcal{R}^{N-1} u) = \max_{\delta \in \Delta} [r_\delta + P_\delta \mathcal{R}^{N-1} u] = \max_{(\delta,\pi) \in \Delta^\infty} V_{\delta\pi}^N(u). \blacksquare \tag{1a}$$

Remark 1. The first equality in (1a) can be written in the alternate form

$$V^N = \max_{\delta \in \Delta} [r_\delta + P_\delta V^{N-1}],\ N = 1, 2, \ldots \tag{1b}$$

where $V^N \equiv \mathcal{R}^N u$ and $V^0 \equiv u$, or equivalently,

$$V_s^N = \max_{a \in A_s} [r(s,a) + \sum_{t \in \mathcal{S}} p(t \,|\, s,a) V_t^{N-1}] \tag{1c}$$

for $N = 1, 2, \ldots$ and $s \in \mathcal{S}$. Of course, (1c) is precisely (1) in §1.2.

Remark 2. The proof is constructive. For if we have found π maximizing $V_\cdot^{N-1}(u)$, then (δ, π) maximizes $V_\cdot^N(u)$ if and only if δ attains the maximum on the right-hand side of (1b).

Advantages of Maximum-N-Period-Value Policies

Policies having maximum N-period value are very useful in practice for several reasons.

- **Ease of Computation.** They are easy to compute for moderate size N by the recursion (1c).
- **Dependence on Horizon Length.** The recursion for computing them automatically finds the best first-period decision for each horizon length N, thus enabling one to study the impact of the horizon length on the best first-period decision.
- **Nonstationary Data.** The recursion extends immediately to the case of nonstationary data *without* any additional computational effort. However, in this event, it is no longer possible to study the impact of the horizon length on the best first-period decision without extra computation *except* in the deterministic case. We now explain briefly why this is so.

Rolling Horizons and the Backward and Forward Recursions

Consider first the deterministic case in which the reward $r^n(s,t)$ earned in period n when an individual moves from state s in period n to state t in period $n+1$ is nonstationary and one is given the states σ, τ in which one must begin and end in periods 1 and N respectively. Then the obvious generalization of the recursion (1c) is the *backward recursion*

$$B_s^n = \max_t [r^n(s,t) + B_t^{n+1}],\ s \in \mathcal{S} \text{ and } n = 1, \ldots, N \tag{2}$$

where B_s^n is the maximum reward that can be earned in periods $n, \ldots, N$ by an individual starting in state s in period n and $B_t^{N+1} = 0$ or $-\infty$ according as $t = \tau$ or $t \neq \tau$. Observe that although we have suppressed the fact in the notation, B_s^n depends on N since B_s^n is the maximum reward earned in periods $n, \ldots, N$. Thus, if one wishes to tabulate the maximum N-period value B_σ^1 for $N = 1, \ldots, M$, then one must compute the B_s^n from (2) for all $1 \leq n \leq N \leq M$ and $s \in \mathcal{S}$. If we fix the number of state-action pairs, this requires $O(M^2)$ additions and comparisons.[5]

However, it is possible to do this computation with at most $O(M)$ additions and comparisons by instead using the *forward recursion*

$$F_t^n = \max_s [r^n(s,t) + F_s^{n-1}],\ t \in \mathcal{S} \text{ and } n = 1, \ldots, N \tag{2}'$$

where F_t^n is the maximum reward that can be earned in periods $1, \ldots, n$ by an individual ending in state t in period n and $F_s^0 = 0$ or $-\infty$ according as $s = \sigma$ or $s \neq \sigma$. Now the maximum N-per-

[5]If f and g are real-valued functions, write $f(x) = O(g(x))$ if there is a constant K such that $\big|f(x)\big| \leq Kg(x)$ for all x.

iod value is simply F_τ^N which can be tabulated for all $1 \leq N \leq M$ with at most $O(M)$ additions and comparisons. Thus, if one is interested in studying the effect of changes in the horizon length on the maximum N-period value for a deterministic problem, it is much better to use the forward than the backward recursion.

Unfortunately, the forward recursion does not have a natural generalization to stochastic systems. For that reason one is left only with the nonstationary generalization of the backward recursion (2) in that case.

Limitations of Maximum-N-Period-Value Policies

Policies having maximum N-period value do have some significant limitations including the following.

- **Computational Burden.** The computational effort rises linearly with the horizon length N.

- **Nonstationarity of Optimal Policies.** Optimal N-period policies are generally nonstationary, even with stationary data, and so are difficult to implement.

- **Dependence on Horizon Length.** The dependence of the best first-period decision on the horizon length N is generally complex and, especially in the case of large N, requires the user to be rather arbitrary in choosing it.

These considerations lead us to look for an asymptotic theory as N gets large, or alternately, to consider instead the infinite-horizon problem. The latter approach is particularly fruitful because the symmetries present in the problem generally assure that there is an "optimal" policy that is stationary. Also, those policies are "nearly optimal" for all large enough N.

7 MAXIMUM VALUE IN TRANSIENT SYSTEMS [Sh53], [Ho60], [D'E63], [deG60], [Bl62], [Der62], [Den67], [Ve69a], [Ve74], [Ro75c], [Ro78], [RV92]

Why Study Infinite-Horizon Problems? What is the point of studying infinite-horizon problems in a stationary environment? After all, life is finite. No doubt for this reason men and women of practical affairs are usually concerned with the near term, often concerned with the intermediate term, and occasionally concerned with the long term. But is the long term infinite? Probably not. Also, is it reasonable to consider models in which the environment remains immutable and unchanging forever? Hard to believe.

In view of these observations, what is the justification for studying the stationary infinite-horizon problem? Here are some reasons for so doing.

- **Near Optimality for Long Finite-Horizons.** The stationary infinite-horizon problem provides a good approximation to the stationary long finite-horizon problem in the sense that optimal infinite-horizon policies are nearly optimal for long (and often surprisingly short) horizons.

• **Optimality of Stationary Policies.** The stationarity of the environment in the stationary infinite-horizon problem generally assures that there is an optimal policy that is stationary and independent of the horizon. This fact is important in practice because stationary policies are much easier to store and implement than nonstationary ones.

• **Computational Effort Independent of Horizon.** By contrast with the finite-horizon problem, the computational effort to find an optimal infinite-horizon policy is *independent* of the horizon.

• **Includes Many Nonstationary Problems.** The stationary infinite-horizon problem includes the n-period nonstationary problem as a special case. To see this, append to the original state s each period N that the system could be in that state, so the states of the augmented system are the pairs (s, N) for $s \in \mathcal{S}$ and $N = 1, \ldots, n$, and transitions are from each period $N < n$ to $N+1$ and from period n to exiting the system. If $n = m + p$ and transitions in period n are instead to period $m + 1$, the system repeats every p periods after period m. This reduces such "eventually p-periodic" problems to stationary ones.

A natural approach to the infinite-horizon problem is to seek a policy $\pi = (\delta_N)$ whose *value*

$$V_\pi \equiv \sum_{N=0}^{\infty} P_\pi^N r_{\delta_{N+1}} \tag{1}$$

is a maximum. However, this definition makes sense only if the right-hand side of (1) is defined and finite. In order for this to be the case, it suffices to assume that π is *transient*, i.e., $\sum_{N=0}^{\infty} P_\pi^N$ is finite. The st^{th} element of the last series is the sum of the expected sojourn times of all individuals in state t generated by one individual starting in state s in period one. Indeed, the *only* way that (1) can be defined and finite for every choice of $r_\cdot$ is that π be transient. And in that event, V_π is the S-vector of expected infinite-horizon rewards, i.e., the value earned by π starting in each state. Incidentally, it is often convenient to call a decision δ or its transition matrix P_δ *transient* if δ^∞ is transient. When is a transition matrix transient?

Characterization of Transient Matrices

The sequel often uses the (Tchebychev) *norm* $\|P\|$ of a matrix $P = (p_{ij})$ defined by $\|P\| = \max_i \sum_j |p_{ij}|$. This norm has the property that $\|PQ\| \le \|P\|\|Q\|$ whenever the matrix product PQ is defined. Though §1 of the Appendix discusses several properties of this and other norms, the term "norm" means the Tchebychev norm throughout the main text.

Lemma 2. Characterization of Transient Matrices. *If P is a square complex matrix, the following are equivalent.*

1° $P^N \to 0$.

2° *The Neumann series $\sum_{N=0}^{\infty} P^N$ converges absolutely.*

Moreover, each of the above implies

3° *$I - P$ is nonsingular and its inverse equals the Neumann series in 2°.*

Proof. 2° ⇒ 1°. Immediate.

1° ⇒ 2°. Since $P^N \to 0$, $\|P^M\| \le \frac{1}{2}$ for some $M \ge 1$. Let $L = \sum_{j=0}^{M-1} \|P^j\|$. Since $\|P^{j+kM}\| \le \|P^j\|\|P^M\|^k$, it follows that $\sum_{N=0}^{\infty} \|P^N\| = \sum_{j=0}^{M-1} \sum_{k=0}^{\infty} \|P^{j+kM}\| \le L \sum_{k=0}^{\infty} \|P^M\|^k \le 2L < \infty$.

$1^\circ \Rightarrow 3^\circ$. Evidently,

$$(\sum_{N=0}^{M-1} P^N)(I-P) = \sum_{N=0}^{M-1}(P^N - P^{N+1}) = I - P^M.$$

Since $P^M \to 0$, the right-hand side above is nonsingular for large M, so $I-P$ is also nonsingular. Thus, postmultiplying by $(I-P)^{-1}$ and letting $M \to \infty$ establishes 3°. ∎

An alternate (more advanced) proof of the equivalence of 1° and 2° of the Lemma is to apply the Spectral Mapping Theorem 4 and Corollary 2 of the Appendix.

Observe that Lemma 2 implies that γ^∞ is transient if and only if $P_\gamma^N \to 0$, i.e., the expected population size in period N generated by γ^∞ converges to zero as $N \to \infty$. And in that event,

$$0 \le \sum_{N=0}^{\infty} P_\gamma^N = (I - P_\gamma)^{-1}.$$

For example, if P_γ is *strictly substochastic*, i.e., $P_\gamma 1 \ll 1$, then γ is transient. More generally, if P_γ is substochastic, then γ is transient if and only if P_γ^S is strictly substochastic.

Transient System. As discussed above, V_π is defined and finite for every π and $r.$ if and only if *every* π is transient. In that event, call the system *transient*. As an example, note that circuitless systems are transient because then $P_\pi^S = 0$ for all π.

The goal now is to show that for transient systems, there exists a policy having maximum value, and one such policy is stationary. The constructive proof of these claims depends on the following fundamental Comparison Lemma. It asserts that the difference between the values of two transient policies $\pi = (\gamma_N)$ and θ is the sum over N of the product of two matrices that depend on N. The first is the matrix P_π^N of expected numbers of individuals in each state after N periods starting from each state when π is used in periods $1, \ldots, N$. The second is the difference $V_{\gamma_{N+1}\theta} - V_\theta$ between the values of the policies (γ_{N+1}, θ)—which uses γ_{N+1} and then θ—and θ.

Comparison Lemma

Lemma 3. Comparison Lemma. *If $\pi = (\gamma_N)$ and θ are transient policies, then*

$$V_\pi - V_\theta = \sum_{N=0}^{\infty} P_\pi^N G_{\gamma_{N+1}\theta}$$

where the comparison function is defined by

$$G_{\gamma\theta} \equiv r_\gamma + P_\gamma V_\theta - V_\theta \text{ for } \gamma \in \Delta.$$

If also $\pi = \gamma^\infty$, then

$$V_\gamma - V_\theta = \sum_{N=0}^{\infty} P_\gamma^N G_{\gamma\theta} = (I - P_\gamma)^{-1} G_{\gamma\theta}$$

where $V_\gamma \equiv V_{\gamma^\infty}$. Moreover, $V = V_\gamma$ if and only if $V = r_\gamma + P_\gamma V$.

Proof. Since π is transient, $V_\pi^N \to V_\pi$ is finite and $P_\pi^N \to 0$, so $V_\pi^N(V_\theta) = V_\pi^N + P_\pi^N V_\theta \to V_\pi$. Thus,

$$V_\pi - V_\theta = \lim_{N\to\infty}[V_\pi^{N+1}(V_\theta) - V_\pi^0(V_\theta)] = \sum_{N=0}^{\infty}[V_\pi^{N+1}(V_\theta) - V_\pi^N(V_\theta)]$$
$$= \sum_{N=0}^{\infty} P_\pi^N[V_{\gamma_{N+1}}^1(V_\theta) - V_\theta] = \sum_{N=0}^{\infty} P_\pi^N G_{\gamma_{N+1}\theta}.$$

If $\pi = \gamma^\infty$, the last term above becomes $[\sum_{N=0}^{\infty}(P_\gamma)^N]G_{\gamma\theta}$. The remaining assertions then follow from Lemma 2 and what was just shown. ∎

Policy-Improvement Method

It is now time to apply the Comparison Lemma to give a policy-improvement method for finding a maximum-value stationary policy for the case where each stationary policy is transient. Then since $(I - P_\gamma)^{-1} \geq I \geq 0$, it follows from the Comparison Lemma that $V_\gamma - V_\delta \geq G_{\gamma\delta} > 0$ (resp., $V_\gamma - V_\delta \leq G_{\gamma\delta} \leq 0$) if $G_{\gamma\delta} > 0$ (resp., ≤ 0). For this reason we say γ *improves* δ if $G_{\gamma\delta} > 0$. If no decision improves δ, we claim that $\max_{\gamma\in\Delta} G_{\gamma\delta} = 0$, so $V_\gamma \leq V_\delta$ for all γ. The claim is an easy consequence of the facts that $G_{\delta\delta} = 0$ and the s^{th} element $G_{\gamma\delta s} = r(s,\gamma^s) + \sum_{t\in\mathcal{S}} p(t \mid s,\gamma^s)V_{\delta t}$ of $G_{\gamma\delta}$ depends on γ^s, but *not* γ^t for $t \neq s$. For suppose that $G_{\gamma\delta s} > 0$ for some γ and s. Define decision η by $\eta^s = \gamma^s$ and $\eta^t = \delta^t$ for $t \neq s$. Then $G_{\eta\delta s} = G_{\gamma\delta s} > 0$ and $G_{\eta\delta t} = G_{\delta\delta t} = 0$ for $t \neq s$. Hence $G_{\eta\delta} > 0$ and η improves δ, which is a contradiction and establishes the claim. The above remarks permit us to prove the following result constructively.

Lemma 4. Existence of Maximum-Value Stationary Policy. *If every stationary policy is transient, there is a maximum-value stationary policy δ^∞. Moreover,*

$$0 = \max_{\gamma\in\Delta} G_{\gamma\delta}, \tag{2a}$$

or equivalently, $V \equiv V_\delta$ is a fixed point of $\mathcal{R}$, i.e., $V = \mathcal{R}V$, or what is the same thing,

$$V = \max_{\gamma\in\Delta}[r_\gamma + P_\gamma V]. \tag{2b}$$

Moreover, the maximum value over the stationary policies is the unique fixed point of $\mathcal{R}$.

Proof. Let δ_0 be arbitrary and choose decisions $\delta_1, \delta_2, \ldots$ inductively so δ_{N+1} improves δ_N, and terminates otherwise. Since the value of each policy in the sequence is higher than that of its predecessor, no decision can occur twice in the sequence. Thus because there are only finitely many decisions, the sequence must terminate with a $\delta = \delta_N$ having no improvement. Then from the discussion preceding the Lemma, (2a) or equivalently (2b) holds, so δ^∞ is a maximum-value stationary policy. Conversely, if V is a fixed point of $\mathcal{R}$, then $V = r_\eta + P_\eta V$ for some $\eta \in \Delta$, whence $V = V_\eta$. Thus $\max_{\gamma\in\Delta} G_{\gamma\eta} = 0$ and so $V = V_\eta$ is the maximum value over the stationary policies. ∎

The finite algorithm given in the above proof finds a maximum-value stationary policy. Call the algorithm the *policy-improvement method*. The process of finding an improvement of δ, if one exists, involves two steps, viz., finding the value of δ and then seeking an improvement of δ. The first step entails solving the system

$$(I - P_\delta)V = r_\delta$$

for $V = V_\delta$. The second is to find, for each state s, an action $\gamma^s \in A_s$ for which

$$r(s, \gamma^s) + \sum_{t \in \mathcal{S}} p(t \,|\, s, \gamma^s) V_{\delta t} \geq V_{\delta s}$$

with strict inequality holding for some s, i.e., a decision $\gamma = (\gamma^s)$ such that $r_\gamma + P_\gamma V_\delta > V_\delta$.

Stationary Maximum-Value Policies: Existence and Characterization

We now come to our main results for transient systems.

Theorem 5. Characterization of Transient Systems. *A system is transient if and only if every stationary policy is transient.*

Proof. It suffices to prove the "if" part. By Lemma 4, there is a maximum-value stationary policy δ^∞, say, for the case $r_\gamma \equiv 1$ for all γ. Let $\pi = (\gamma_1, \gamma_2, \dots)$ be any policy. Then since $1 \leq V_\delta \ll \infty$ is a fixed point of $\mathcal{R}$, $\mathcal{R}$ is increasing, and $\mathcal{R}^M 1$ is the maximum M-period value with terminal value 1, it follows that

$$V_\delta = \mathcal{R}V_\delta = \cdots = \mathcal{R}^M V_\delta \geq \mathcal{R}^M 1 \geq \sum_{N=0}^{M} P_\pi^N 1$$

for each $M = 0, 1, \dots$, so π is transient. ∎

Theorem 6. Existence and Characterization of Maximum-Value Policies and Maximum Value. *In a transient system, there is a stationary maximum-value policy. Moreover, a policy θ has maximum value if and only if* $\max_{\gamma \in \Delta} G_{\gamma\theta} = 0$. *Finally, the maximum value V^* is the unique fixed point of $\mathcal{R}$.*

Proof. For the first assertion, observe from Lemma 4 that there is a maximum-value stationary policy δ^∞, say, with $\max_{\gamma \in \Delta} G_{\gamma\delta} = 0$. Hence, by the Comparison Lemma, $V_\pi \leq V_\delta$ for all π, so δ^∞ is a maximum-value policy. For the second assertion, if θ has maximum value, $V_\theta = V_\delta$ so $\max_{\gamma \in \Delta} G_{\gamma\theta} = \max_{\gamma \in \Delta} G_{\gamma\delta} = 0$. Conversely, if $\max_{\gamma \in \Delta} G_{\gamma\theta} = 0$, then θ has maximum-value by the Comparison Lemma. The third assertion is immediate from the first assertion and Lemma 4. ∎

Key Ideas. The above developments illustrate three important ideas that will recur frequently in the sequel in studying optimality concepts for nontransient systems.

• **Policy Improvement and Comparison Lemma.** It is often possible to use suitably generalized versions of the policy-improvement method and Comparison Lemma together to establish constructively the existence of stationary "optimal" policies, and to characterize them for general systems—transient or not.

• **Convenient Terminal Values.** It is often useful to study a problem with the aid of a terminal value that makes calculations easy and then find the effect of changing that value to the desired one. For example, in the proof of Theorem 5, $V_\delta \geq 1$ is a fixed point of $\mathcal{R}$ whence $V_\delta = \mathcal{R}^M V_\delta \geq \mathcal{R}^M 1$.

• **System Properties.** To establish properties of the system that are independent of the rewards, e.g., transience in Theorem 5, it is frequently useful to set all rewards and/or terminal values equal to 1 (or 0 or -1), and then apply existence or characterization results about "optimal" policies and their values.

The fact that V^* is a fixed point of $\mathcal{R}$ is intuitive on writing that fact as

$$\underset{\substack{\text{max value}\\ \text{periods}\\ \text{1,2,...}}}{V^*} = \max_{\gamma\in\Delta} [\ \underset{\substack{\text{reward in}\\ \text{period 1}\\ \text{using } \gamma}}{r_\gamma} + \underset{\substack{\text{max value}\\ \text{in periods}\\ \text{2,3,... given}\\ \gamma \text{ used in 1}}}{P_\gamma V^*}\ ]. \tag{3}$$

The fact that V^* is the unique solution of the system of S nonlinear equations $V = \mathcal{R}V$ in S variables suggests the possibility of using classical iterative methods to solve the equations for V^*, e.g., Newton's method, successive approximations, Gauss-Seidel method, etc. We first discuss Newton's method.

Newton's Method: A Specialization of Policy-Improvement Method

Newton's method is an instance of the policy-improvement method. To see this, initiate Newton's method with $V = V_\delta$. The first step of the method is to find a linear approximation $\widetilde{\mathcal{R}}$ of $\mathcal{R}$ with $\widetilde{\mathcal{R}}V = \mathcal{R}V$. The natural linear approximation is $\widetilde{\mathcal{R}}U \equiv r_\gamma + P_\gamma U$ for $U \in \Re^S$ where γ is a maximizer of $r_\gamma + P_\gamma V$. This is equivalent to choosing γ to maximize $G_{\gamma\delta}$, i.e., choosing the "best" improvement of δ. The linear approximation $\widetilde{\mathcal{R}}$ has the properties that $\widetilde{\mathcal{R}}V = \mathcal{R}V$ and $\widetilde{\mathcal{R}}\,U \leq \mathcal{R}U$ for all U. The second step is to solve the linear equations $U = \widetilde{\mathcal{R}}U$, i.e., $U = r_\gamma + P_\gamma U$, for $U = V_\gamma$.

Now replace δ by γ and repeat the above two steps iteratively. Thus Newton's method for solving $V = \mathcal{R}V$ is the *specialization of the policy improvement method* that, given δ, chooses a γ that not only improves δ, but also maximizes $G_{\gamma\delta}$. Moreover, the method converges in finitely many steps.

Successive Approximations: Maximum N-Period Value Converges to Maximum Value

Another classical method for finding the fixed point of $\mathcal{R}$ is *successive approximations*. This method entails choosing the sequence $V^0, V^1, \ldots$ iteratively by the rule $V^N = \mathcal{R}V^{N-1}$, $N = 1, 2, \ldots$. The method has special interest in dynamic programming because $V^N = \mathcal{R}^N V$ is the maxi-

mum N-period value with terminal reward $V = V^0$. Since in transient systems, $P_\delta^N \to 0$ for each δ, it is plausible that $V^N \to V^*$ in such systems. The sequel establishes this fact and provides the rate of convergence as well.

In order to see what the rate of convergence might be, consider the special case where Δ consists of only a single decision δ. Then

$$\|V^* - \mathcal{R}^N V\| = \|\mathcal{R}^N V^* - \mathcal{R}^N V\| = \|P_\delta^N (V^* - V)\| \le \|P_\delta^N\| \|V^* - V\|.$$

Thus the rate at which $\mathcal{R}^N V$ converges to V^* is bounded above by the rate at which P_δ^N converges to 0. And the rate of convergence of the latter is determined by the spectral radius of P_δ. Before establishing this fact, it seems useful to give an example.

Geometric Interpretation: a Single-State Reliability Example

It is informative to illustrate the above results geometrically on a concrete example for the case of a transient system with a single state and three decisions σ, μ, γ. The example is interesting for another reason too, viz., the time between transitions is random. Such systems are *semi-Markov decision chains*, though this does not really introduce anything new for infinite-horizon transient systems as the discussion below reveals.

Suppose that a *contractor* provides a service, e.g., lighting, automotive transportation, or desktop computing, for a *client*. The client can cancel the service at the beginning of any period without prior notice. The contractor estimates that the client will cancel independently in each period with probability $1 - \lambda$, $0 < \lambda < 1$. The contractor has three products with which to provide the service, e.g., three types of lights, cars or computers. The service lives of the copies of product δ that the contractor places in service are independent random variables whose distributions coincide with that of another nonnegative integer-valued random variable L_δ. Let $P_\delta \equiv \mathrm{E}\lambda^{L_\delta}$ be the probability that the client does not cancel the service during the life of product δ. Also let r_δ be the expected revenue that product δ earns during its lifetime. Now $V^* = \max_\delta V_\delta$.

The table below provides the problem data and the value of each policy. The system is

δ	r_δ	P_δ	$(I - P_\delta)^{-1}$	V_δ
σ	8	$\frac{1}{5}$	$\frac{5}{4}$	10
μ	12	$\frac{1}{4}$	$\frac{4}{3}$	16
γ	6	$\frac{3}{4}$	4	24

is transient since $\max(P_\sigma, P_\mu, P_\gamma) < 1$. Also, the maximum expected revenue that the contractor can earn, viz., $V = V^* = 24$, is the unique solution of the nonlinear equation

$$V = \max(\overbrace{8 + \tfrac{1}{5}V}^{\sigma}, \overbrace{12 + \tfrac{1}{4}V}^{\mu}, \overbrace{6 + \tfrac{3}{4}V}^{\gamma}).$$

Note that γ is optimal even though its expected lifetime revenue is smallest.

It is useful to examine how policy improvement method would produce this result with Newton improvements, i.e., the best improvement at each step. To that end, note that the optimal-return operator is given by $\mathcal{R}V = \max(8 + \frac{1}{5}V, 12 + \frac{1}{4}V, 6 + \frac{3}{4}V)$.

• **Policy σ.** Suppose one begins with σ. The first step is to solve $V = 8 + \frac{1}{5}V$ for the value of σ, i.e., for $V = V_\sigma = 10$. Next $\mathcal{R}(10) = \max(8 + \frac{1}{5}10, 12 + \frac{1}{4}10, 6 + \frac{3}{4}10) = \max(10, 14\frac{1}{2}, 13\frac{1}{2}) = 14\frac{1}{2}$, so $G_{\mu\sigma} = 14\frac{1}{2} - 10 = 4\frac{1}{2}$. Therefore μ improves σ.

• **Policy μ.** Now solve $V = 12 + \frac{1}{4}V$ for $V = V_\mu = 16$. Then $\mathcal{R}(16) = \max(8 + \frac{1}{5}16, 12 + \frac{1}{4}16, 6 + \frac{3}{4}16) = \max(11\frac{1}{5}, 16, 18) = 18$. Thus, $G_{\gamma\mu} = 18 - 16 = 2$. Therefore γ improves μ.

• **Policy γ.** Now solve $V = 6 + \frac{3}{4}V$ for $V = V_\gamma = 24$. Then $\mathcal{R}(24) = \max(8 + \frac{1}{5}24, 12 + \frac{1}{4}24, 6 + \frac{3}{4}24) = 24$. Thus, $G_{\delta\gamma} \leq 0$ for all δ, establishing that γ is optimal.

Figure 4a displays the graph of the optimal-return operator for the above example as a wide bold line. Note that $V - \mathcal{R}V$ is continuous, strictly increasing and changes sign once from $-$ to $+$, and so has a unique zero V^*, viz., the unique fixed point of $\mathcal{R}$. Also $\mathcal{R}V_\delta \geq V_\delta$ for all decisions δ, and equality holds if and only if $V_\delta \equiv V^*$.

Figures 4b and 4c respectively illustrate Newton's method and successive approximations for the above example. In each case, if U and V are values calculated at successive iterations, wide bold line segments are drawn from (U, U) to $(U, \mathcal{R}U)$ to (V, V). Together these line segments form a path that traces the trajectory of successive iterates from the initial value to V^*. In particular, starting from decision σ, Newton's method first chooses μ and then γ. By contrast, starting from the value V^0, successive approximations follows an infinite staircase trajectory.

Observe from Figure 4b,c that starting from an initial value V^0 for which $V^0 \leq \mathcal{R}V^0$, the value that Newton's method calculates at each iteration lies above the one that successive approximations computes, and this is so in general. Thus, Newton's method converges at least as fast as successive approximations. Of course this does not necessarily mean that Newton's method is more efficient than successive approximations because the former requires more work at each iteration than the latter.

System Degree, Spectral Radius and Polynomial Boundedness

The *degree* of a sequence $P_0, P_1, \ldots$ of matrices is 0 if $\|P_N\| = O(\alpha^N)$ for some $0 < \alpha < 1$ (i.e., $\|P_N\| \leq K\alpha^N$ for some K), the smallest positive integer k for which $\|P_N\| = O(N^{k-1})$ if such an integer exists and the degree is not 0, and $+\infty$ otherwise. The *degree* d_P of a square matrix P is the degree of the sequence $P^0, P^1, P^2, \ldots$ of nonnegative integer powers of P.

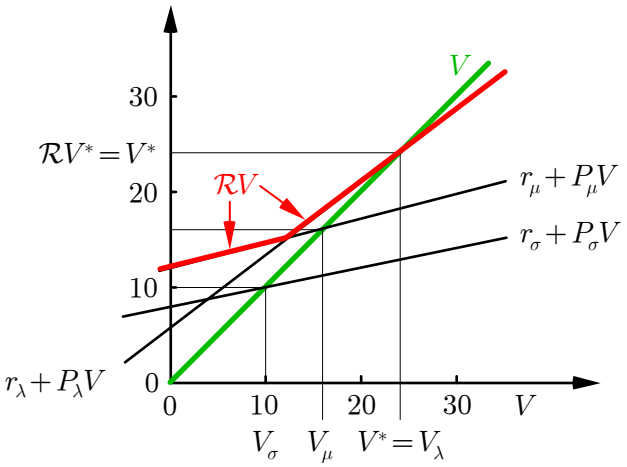


Figure 4a. Maximum Value is Unique Fixed Point of Optimal-Return Operator

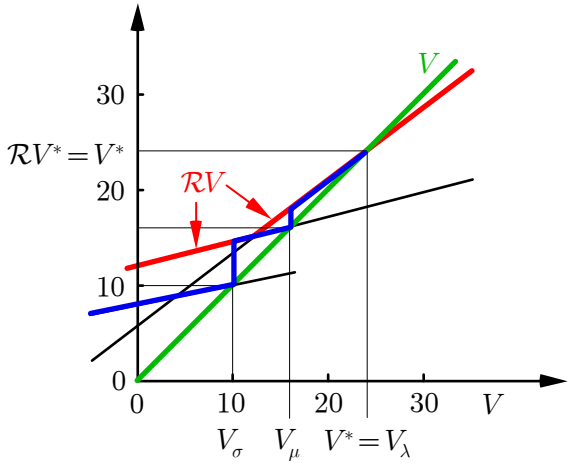


Figure 4b. Newton's Method

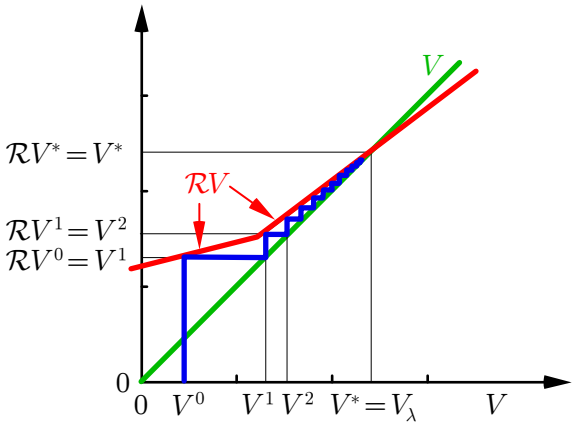


Figure 4c. Successive Approximations

Examples of Degrees of Sequences

Degree of $\{P_N\}$	0	1	2	3	∞
P_N	$\left(\frac{1}{2}\right)^N$	$(-1)^N$	N	$-N^2$	2^N

The *spectral radius* σ_P of a square complex matrix P is the radius of the smallest circle in the complex plane that is centered at the origin and contains the eigenvalues of P as Figure 5 illustrates. If $\sigma_P > 0$, then the *normalized matrix* $\overline{P} \equiv \sigma_P^{-1} P$ has spectral radius one.

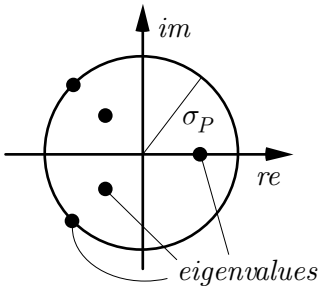


Figure 5. Spectral Circle

Lemma 7. *If P is an $S \times S$ complex matrix, the following are equivalent*:

1° $1 \le d_P < \infty$; $1 \le d_P \le S$; *and* $\sigma_P = 1$. (*polynomially bounded*)

Moreover, the following are equivalent:

2° $d_P = 0$; $P^N \to 0$; *and* $\sigma_P < 1$. (*transient*)

Finally, the following are equivalent:

3° $P^N = 0$ *for some* $N \ge 0$; $\sigma_P = 0$; *and* $P = TUT^{-1}$ *for some upper-triangular matrix* U *with zero diagonal elements and some nonsingular matrix* T. *If also,* P *is real and nonnegative, one such* T *is a permutation matrix, so the system graph of* P *is circuitless.* (*circuitless*)

Proof. Use the Jordan form of P. ∎

Corollary 8. *If P is an $S \times S$ complex matrix with $\sigma \equiv \sigma_P > 0$ and $\sigma^{-1}P$ has degree $\overline{d}$, then $1 \le \overline{d} \le S$ and $P^N = O(N^{\overline{d}-1}\sigma^N)$.*

Proof. Since $\sigma^{-1}P$ has spectral radius one, the result follows from 1° of Lemma 7. ∎

The *degree* d_π of a policy π is the degree of the sequence (P_π^N) of endogenous expected population sizes that π generates. The degree of a policy is of fundamental importance in the development of the theory and is a natural measure of the endogenous rate of growth of the sequence of expected population sizes when using the policy.

In particular, the degree of π is zero if and only if the sequence of endogenous expected population sizes when using π converges to zero geometrically, so π is transient. If the degree of π is nonzero and finite, it is one plus the order of the lowest-order polynomial majorizing the sequence of endogenous expected population sizes when using π. For example, if the system is stochastic, the degree of each policy is one because the sequence of endogenous expected population sizes is bounded (by one). If the degree of π is infinite, the sequence of endogenous expected population sizes when using π is not majorized by any polynomial.

Put $d_\delta \equiv d_{\delta^\infty}$ and call $d \equiv \max_{\delta\in\Delta} d_\delta$ the *system degree*. Let $\sigma_\delta \equiv \sigma_{P_\delta}$ and call $\sigma \equiv \max_{\delta\in\Delta}\sigma_\delta$ the *system spectral radius*. If $\sigma > 0$, call the system with transition matrices $\overline{P}_\delta \equiv \sigma^{-1}P_\delta$ for all $\delta \in \Delta$ the *normalized system*.

The next result is *very important*. It implies that $\max_{\pi\in\Delta^\infty} d_\pi = d$, i.e., the *system degree majorizes the degree of every policy*.

Theorem 9. System Degree. *The sequence* $\max_{\pi\in\Delta^\infty}\|P_\pi^N\|$, $N = 0, 1, \ldots,$ *has degree* d.

Proof. We give the proof only for the case $d = 0$, (Later we give the proof for the case $d = 1$). Put $\widehat{\mathcal{R}}V \equiv \max_{\delta\in\Delta}[1 + P_\delta V]$ and $\mathcal{R}V \equiv \max_{\delta\in\Delta} P_\delta V$. Since each stationary policy is transient, $\widehat{\mathcal{R}}$ has a unique fixed point V by Lemma 4, and $V = 1 + P_\delta V \geq 1$. Also $0 \leq \mathcal{R}V \leq \widehat{\mathcal{R}}V = V$. Thus $\mathcal{R}^N V$ is decreasing in N and is bounded below by 0, so $\mathcal{R}^N V \downarrow U$, say. Since $\mathcal{R}$ is continuous, $U = \lim_{N\to\infty}\mathcal{R}(\mathcal{R}^N V) = \mathcal{R}U$, so U is a fixed point of $\mathcal{R}$, whence $U = 0$ because 0 is the unique fixed point of $\mathcal{R}$ by Lemma 4. Thus $\mathcal{R}^N V \downarrow 0$. Hence $\mathcal{R}^M V \leq \frac{1}{2}V$ for some $M \geq 1$. Then $\mathcal{R}^{kM}V \leq (\frac{1}{2})^k V$ for $k = 1, 2, \ldots,$ from which fact it follows that $\mathcal{R}^N V = O(\alpha^N)$ with $\alpha \equiv (\frac{1}{2})^{1/M}$. Since $\mathcal{R}^N V = \max_{\pi\in\Delta^\infty} P_\pi^N V$, $\max_{\pi\in\Delta^\infty}\|P_\pi^N\| = O(\alpha^N)$. ∎

Transient Systems. For the case $d = 0$, Theorem 9 strengthens Theorem 5 by implying not only that every policy is transient, but also that every policy has degree zero.

Geometric Convergence of Successive Approximations

It is now possible to establish the convergence rate of $\mathcal{R}^N V$ to V^* in transient systems.

Theorem 10. Geometric Convergence of Successive Approximations. *In a system with positive maximum spectral radius* σ *and degree* $\overline{d}$ *of the normalized system,*

$$\|\mathcal{R}^N U - \mathcal{R}^N V\| = O(N^{\overline{d}-1}\sigma^N)\|U - V\| \text{ uniformly in } U, V \in \Re^S.$$

If also the system is transient, i.e., $\sigma < 1$, *then* $\mathcal{R}^N V \to V^*$.

Proof. Suppose $U, V \in \Re^S$ and let $\overline{P}_\pi^N$ be the N-step transition matrix for π in the normalized system. Choose π so $\mathcal{R}^N U = V_\pi^N + P_\pi^N U$. Then $\mathcal{R}^N V \geq V_\pi^N + P_\pi^N V$, so by Theorem 9 and Lemma 7,

$$\mathcal{R}^N U - \mathcal{R}^N V \leq P_\pi^N(U - V) = \sigma^N \overline{P}_\pi^N(U - V) = O(N^{\overline{d}-1}\sigma^N)\|U - V\|$$

uniformly in U, V. Interchanging the roles of U, V establishes the first claim. The second claim then follows from the first by choosing $U = V^*$ and using the facts that V^* is finite (because the system is transient) and $\mathcal{R}^N V^* = V^*$. ∎

Theorem 10 depends on the proof of Theorem 9 for arbitrary d. There is a weaker result that depends on Theorem 9 only for $d = 0$, the case for which that theorem was proved above. The weaker result is that $\left\|V^* - \mathcal{R}^N V\right\| = O(\overline{\sigma}^N)\|V^* - V\|$ uniformly in V for $\overline{\sigma} > \sigma$. The proof is identical to that of Theorem 10 except that one normalizes the P_δ with respect to $\overline{\sigma}$ rather than σ.

In a transient system, Theorem 10 provides a sharp estimate of the rate of convergence of $\mathcal{R}^N V$ to V^*. However, the estimate depends on σ which is, in general, difficult to compute. For that reason it is useful to have an upper bound on σ that is easier to compute. The next result gives such a bound.

Lemma 11. *If P is a square complex matrix, then $\sigma_P \leq \|P\|$.*

Proof. If λ is an eigenvalue of P, then there is a $v \neq 0$ such that $\lambda v = Pv$. Hence $|\lambda|\|v\| = \|Pv\| \leq \|P\|\|v\|$. On dividing by $\|v\|$, the result follows. ∎

Corollary 12. *If $\mu \equiv \max_{\delta\in\Delta}\|P_\delta\| < 1$, the system is transient and $\|V^* - \mathcal{R}^N V\| \leq \mu^N\|V^* - V\|$.*

Proof. Iterate the inequality $\|V^* - \mathcal{R}V\| \leq \mu\|V^* - V\|$. ∎

Remark. Notice that Corollary 12 is weaker than Theorem 10 when $\sigma \neq \mu$ because then, by Lemma 11, $\sigma < \mu$. If $\sigma = \mu$, then the Corollary gives the constant in Theorem 10. However, then the rate of convergence is the same in both cases because $\overline{d} = 1$ since $\|\overline{P}_\pi^N\| \leq 1$ for all N and $\overline{d}$ is the degree of $\overline{P}_\delta$ for some δ.

Contraction Mappings. The usual proof of Corollary 12 employs the contraction-mapping theorem. Indeed, when the hypothesis of Corollary 12 is valid, an easy modification of the proof of Theorem 10 shows that the mapping $\mathcal{R}$ is a *contraction with modulus* $0 \leq \mu < 1$, i.e., one has $\|\mathcal{R}U - \mathcal{R}V\| \leq \mu\|U - V\|$ for all $U, V \in \Re^S$. The contraction-mapping theorem in the Appendix then assures that $\{\mathcal{R}^N V\}$ is a Cauchy sequence and converges geometrically to the unique fixed point V^* of $\mathcal{R}$ at the rate μ that Corollary 12 provides. The development here employs a different approach because, as the above remark discusses, it gives the sharper result of Theorem 10.

Actually, when the system is transient, but the hypothesis of Corollary 12 does not hold, it is possible to modify the contraction-mapping method to establish geometric convergence of $\mathcal{R}^N V$ to V^* at any rate $\bar{\sigma}$ with $\sigma < \bar{\sigma} < 1$ in two different ways. One is to apply Theorem 10 to show that even though $\mathcal{R}$ is not a contraction with modulus $\bar{\sigma}$ in the usual (Tchebychev) norm, $\mathcal{R}^k$ is a contraction with modulus $\bar{\sigma}^k$ for some $k > 0$ in that norm. The other is to show that $\mathcal{R}$ is in fact a contraction with modulus $\bar{\sigma}$ in a different "weighted" (Tchebychev) norm.

Linear-Programming Method

So far in this section we have explored the possibility of finding maximum-value policies in transient systems by finding the unique fixed point of $\mathcal{R}$. It is possible to find this fixed point by linear programming if we consider instead the larger set of vectors V satisfying $V \geq \mathcal{R}V$, or equivalently the linear inequalities $V \geq r_\delta + P_\delta V$ for all $\delta \in \Delta$. Call such vectors V *excessive*. As we shall see, there is a least excessive vector, and it is the fixed point. The simplest way to see that there is a least excessive vector is to observe that the set of excessive vectors is nonempty, closed, bounded below (by each V_δ) and a *meet sublattice*, i.e., whenever U and V are excessive, so is the pointwise minimum $U \wedge V$. Moreover, the least excessive vector may be found by minimizing any positive linear function of V over the set of excessive vectors V. It might seem odd that this linear program entails minimization while our goal is to maximize value. The explanation for this apparent paradox is simply that this linear program is, as the sequel shows, the *dual* of the maximum-value problem.

The linear-programming method is interesting in its own right and because the theory and algorithms of linear programming are well developed, codes for solving such problems are widely available, and the method is readily adapted to encompass additional stochastic constraints. A typical stochastic constraint might involve a bound on a linear combination of the expected number of times individuals in various states take various actions, e.g., an upper bound on the expected number of times that individuals enter an undesirable set of states. Another stochastic constraint might assure that if one uses an action in one of a subset of states, then one uses the action in all states in the subset. The last constraint can be expressed in terms of linear inequalities in integer and continuous variables.

In order to formulate the linear programs, suppose that there is an S-element row vector $w \geq 0$ of initial population sizes of individuals in each state in period one. Then consider the following pair of dual linear programs for a transient system.

Primal Linear Program. Find S-element row vectors x_δ, $\delta \in \Delta$, that

$$\text{maximize} \qquad \sum_{\delta \in \Delta} x_\delta r_\delta$$

subject to

$$\sum_{\delta \in \Delta} x_\delta (I - P_\delta) = w$$

$$x_\delta \geq 0,\ \delta \in \Delta$$

Dual Linear Program. Find an S-element column vector V that

$$\text{minimizes} \quad wV$$

subject to

$$V \geq r_\delta + P_\delta V,\ \delta \in \Delta.$$

In order to understand the primal program, let $x = (x_\gamma)$ be such that $x_\delta(I - P_\delta) = w$ for one $\delta \in \Delta$ and $x_\gamma = 0$ for all $\gamma \neq \delta$. Then by Lemma 2, $I - P_\delta$ is nonsingular and $x_\delta = w(I - P_\delta)^{-1}$ $= w\sum_{N=0}^{\infty} P_\delta^N \geq 0$, so x is a basic feasible solution of the primal. Call $I - P_\delta$ the *basis associated with* δ. The basic feasible solution x has the following interpretation. Since wP_δ^N is the vector of expected population sizes in each state in period $N+1$ when all individuals use δ^∞, x_δ is the sum of the expected sojourn times of all individuals in each state when they all use δ^∞. Hence since r_δ is the vector of one-period rewards earned by one individual in each state who uses δ^∞, $x_\delta r_\delta = w(I - P_\delta)^{-1} r_\delta = wV_\delta$ is the expected total reward earned by all individuals in all periods when they use δ^∞. The aim is to find a δ^∞ that maximizes the expected total reward, which motivates the primal objective function. The next result establishes that the primal and dual always have optimal solutions in transient systems.

Theorem 13. Linear Programming. *In a transient system, the maximum value V^* is the least feasible solution of the dual, is optimal for the dual for all $w \geq 0$, and is the unique optimal solution of the dual for $w \gg 0$. Moreover, there is a δ such that $V^* = r_\delta + P_\delta V^*$; also each such δ^∞ has maximum value and $I - P_\delta$ is an optimal basis for the primal for all $w \geq 0$.*

Proof. The maximum value V^* is a fixed point of $\mathcal{R}$, and so is feasible for the dual. Choose $\delta \in \Delta$ so $V^* = r_\delta + P_\delta V^*$, whence $V^* = V_\delta$. Suppose $w \geq 0$. Choose $x = (x_\gamma)$ so $x_\delta(I - P_\delta) = w$ and $x_\gamma = 0$ otherwise. As discussed above, x is a basic feasible solution of the primal. Hence, since x and V^* satisfy complementary slackness, they are optimal for the primal and dual respectively. Moreover, since $I - P_\delta$ has a nonnegative inverse, it is an optimal basis for all $w \geq 0$. Thus, since V^* is optimal for the dual for all $w \geq 0$, V^* is the least feasible solution of the dual. Also, V^* is the unique optimal solution of the dual for $w \gg 0$ since then $wV > wV^*$ for $V > V^*$. ∎

It is important to recognize that although the matrix formulations of the primal and dual programs given above are convenient for theoretical analyses like those above, both are highly redundant because there are $\prod_{t \neq s} |A_t|$ distinct decisions that take action $a \in A_s$ in state s. As a consequence the primal will have a distinct variable and the dual a distinct inequality for each of those distinct decisions that take action a in state s. On the other hand, since the dual ine-

qualities associated with taking action a in state s are identical, all but one can be omitted because they are redundant. Hence, the primal variables corresponding to those redundant dual inequalities are also redundant and so can also be eliminated.

After stripping out the redundant variables of the primal and corresponding redundant inequalities of the dual, let x_{sa} be the (unique) primal variable corresponding to taking action a in state s. Then the primal and dual programs take the following irredundant forms. It is these irredundant forms that should be used for computation.

Irredundant Primal. Choose $x = (x_{sa}) \geq 0$ that

$$\text{maximizes} \sum_{a \in A_s, s \in \mathcal{S}} r(s,a) x_{sa}$$

subject to

$$\sum_{a \in A_s} x_{sa} - \sum_{a \in A_t, t \in \mathcal{S}} p(s \,|\, t, a) x_{ta} = w_s,\ s \in \mathcal{S}.$$

Irredundant Dual. Choose $V = (V_s)$ that

$$\text{minimizes} \quad \sum_{s \in \mathcal{S}} w_s V_s$$

subject to

$$V_s \geq r(s,a) + \sum_{t \in \mathcal{S}} p(t \,|\, s, a) V_t,\ a \in A_s,\ s \in \mathcal{S}.$$

We next study the set of all feasible solutions of the irredundant primal. To do this, it is convenient to introduce two definitions.

State-Action Frequency. The *state-action frequency* of a policy is the vector whose sa^{th} component is the expected number of times that individuals are in state s and take action a when the policy is used. If the state-action frequency $x = (x_{sa})$ of a policy π is finite, e.g., as when π is transient, x is feasible for the irredundant primal. To see this, observe that the expected number of times that individuals leave state s when π is used is $\sum_{a \in A_s} x_{sa}$. On the other hand the expected number of times that individuals enter state s when π is used is $w_s + \sum_{a \in A_t, t \in \mathcal{S}} p(s \,|\, t, a) x_{ta}$. The first term accounts for the initial population in state s in period one and the second for the expected number of times (after period one) that individuals enter state s from the preceding period when π is used. Since $x \geq 0$ and the expected number of times that individuals enter and leave state s must coincide when π is used, the state-action frequency x of π is feasible for the irredundant primal.

The above development shows that the state-action frequency of every transient policy is feasible for the irredundant primal. The question arises whether the converse is also true. In short, is every feasible solution of the irredundant primal the state-action frequency of some policy? The answer is 'no'—at least for the "nonrandomized" policies considered so far—as the following example shows.

Example. Nonconvexity of Set of State-Action Frequencies of Nonrandomized Policies. Suppose $\mathcal{S} = \{1, 2\}$ and $w = (1 \quad 0)$. There are only two actions in state one. The first generates no individuals in either state in the next period and the second sends the individual in state one to state two in the next period. There is only one action in state two and it generates no individuals in either state in the next period. Then there are effectively only two nonrandomized policies—both transient—and the set of state-action frequencies is $\{(1 \quad 0 \quad 0), (0 \quad 1 \quad 1)\}$, which is not convex. Thus there are many feasible solutions of the irredundant primal that are not state-action frequencies of nonrandomized policies.

Initially Randomized Polices. On the other hand, Theorem 15 below asserts that every feasible solution of the irredundant primal is a state-action frequency if the system is transient and the class of admissible policies is enlarged to permit initial randomization. Call a policy *initially randomized* if at the beginning of period one, the policy selects each stationary policy δ^∞ with some given probability $\mu_\delta \geq 0$, $\sum_{\delta \in \Delta} \mu_\delta = 1$. Call such a policy *nonrandomized* if all but one of the μ_δ's is zero. In particular, if only μ_δ is nonzero, and so equal to one, the corresponding nonrandomized policy is simply the stationary one δ^∞.

Theorem 15 also asserts that the extreme solutions of the irredundant primal are state-action frequencies of nonrandomized stationary policies even when the system is not transient. To establish this result, we first require the following characterization of the existence of a solution and of a least solution of the following system of linear inequalities where w and P are nonnegative.

$$x = w + xP,\ x \geq 0. \tag{4}$$

Lemma 14. *Suppose that the row vector w and square matrix P are nonnegative and have a like number of columns, and let $y \equiv \sum_{N=0}^\infty wP^N$. Then the following are equivalent.*

1° *y is finite.*

2° *The inequalities* (4) *have a solution, e.g., y.*

3° *y is finite and is the least solution of the inequalities* (4).

Proof. 1° ⇒ 2°. Observe that $y = w + \sum_{N=1}^\infty wP^N = w + yP \geq 0$, i.e., y satisfies (4).

2° ⇒ 3°. If x satisfies (4), then $x = w + xP = \cdots = \sum_{i=0}^{N-1} wP^i + xP^N \geq \sum_{i=0}^{N-1} wP^i$ for $N = 1, 2, \ldots$, so $x \geq y$. Thus since 1° ⇒ 2°, y is the least solution of (4).

3° ⇒ 1°. Immediate.

Theorem 15. Feasible Solutions of Irredundant Primal are State-Action Frequencies of Initially Randomized Policies. *If $w \geq 0$, then x is an extreme solution of the irredundant primal if and only if x is the finite state-action frequency of a nonrandomized stationary policy. If $w \gg 0$, the feasible row bases of the irredundant primal are the matrices $I - P_\delta$ for which δ^∞ is transient. If $w \geq 0$ and the system is transient, then the set of feasible solutions of the irredundant primal coincides with the set of state-action frequencies of the initially randomized policies.*

Proof. Suppose $w \geq 0$. We begin by proving the sufficiency in the first assertion. Suppose x is the finite state-action frequency of δ, so $x_\delta = \sum_{N=0}^\infty w P_\delta^N$ is finite. Also suppose $x = \frac{1}{2}(u + v)$ for some feasible solutions u and v of the irredundant primal. Then the nonzero elements of x, u and v are contained among the state-action pairs defining δ, so u_δ and v_δ satisfy (4) with $P = P_\delta$. Since x_δ is the least solution of (4) by Lemma 14, $x_\delta \leq u_\delta, v_\delta$, whence because $x_\delta = \frac{1}{2}u_\delta + \frac{1}{2}v_\delta$, $x_\delta = u_\delta = v_\delta$. Hence, $x = u = v$, so x is an extreme solution of the irredundant primal as asserted.

To establish the necessity in the first assertion, let L be the $\sum_{s\in\mathcal{S}} |A_s|$ row by S column constraint matrix of the irredundant primal. Let x be an extreme solution of $xL = w$, $x \geq 0$ and x_+ be the row subvector of positive elements of x. Then the rows B of L corresponding to x_+ are linearly independent. Let B_+ (resp., B_0) be the columns of B that contain at least one positive (resp., no positive) element, and let w_+ (resp., w_0) be the corresponding row subvector of w. Since $x_+ B_0 = w_0$, $x_+ \gg 0$, $B_0 \leq 0$ and $w_0 \geq 0$, it follows that $B_0 = 0$ and $w_0 = 0$. Thus, since each row of L, and hence of B, has at most one positive element, each column of B_+ has exactly one positive element. For if not, B_+ has more rows than columns, contradicting the fact that the rows of B, and hence B_+, are linearly independent. Thus B_+ is nonsingular. Let δ be a decision that uses the state-action pairs corresponding to the rows of B and arbitrary actions in other states. On possibly permuting the rows of L, we can write $B_+ = I - P_+$ for the restriction $P_+ \geq 0$ of P_δ to the set $\mathcal{S}_+$ of states corresponding to w_+. Since B_+ is nonsingular, x_+ is the unique solution of $x_+ = w_+ + x_+ P_+ \geq 0$, so $x_+ = \sum_{N=0}^\infty w_+ P_+^N$ by Lemma 14. Hence since $\mathcal{S}_+$ is closed under δ, x is the state-action frequency of δ^∞, which proves the first assertion. The second assertion follows from what was just proved on noting that $\mathcal{S}_+ = \mathcal{S}$ when $w \gg 0$.

To establish the third assertion, suppose $w \geq 0$ and the system is transient. Then the set of feasible solutions of the irredundant primal is bounded because it has an optimal solution when $r(s, a) = 1$ for all $a \in A_s$ and $s \in \mathcal{S}$ by Theorem 13, and so is a polytope, i.e., is polyhedral and bounded. Also, it is known that each element of a polytope is a convex combination of its extreme points. The third assertion then follows from the first if we take the weights on the extreme points in the convex combination to be the initial probabilities of choosing the various nonrandomized stationary policies. ∎

Finding An Optimal Policy. Observe from Theorems 13 and 15 that if the system is transient and $x = (x_{sa})$ is a basic optimal solution of the irredundant primal program, then one maximum-value decision in state s is to take the unique action $a \in A_s$ for which $x_{sa} > 0$ if such an action exists and to take any action if $x_{sa} = 0$ for all $a \in A_s$.

Simplex Method: A Specialization of Policy Improvement Method. Application of the simplex method to solve the irredundant primal is equivalent to the specialization of the policy-improvement method in which the action for only a single state is changed at each iteration. To see this, suppose that $\delta \in \Delta$. The basis corresponding to δ is $I - P_\delta$ and the price vector is given by $V_\delta = (I - P_\delta)^{-1} r_\delta$, so that the reduced profits corresponding to $\gamma \in \Delta$ are $r_\gamma - (I - P_\gamma)V_\delta = G_{\gamma\delta}$. The simplex method selects a γ and an s for which $G_{\gamma\delta s} > 0$ and γ differs from δ only in state s. (This selection rule agrees with the ordinary rule by which the simplex method chooses a variable to leave the basis when $w \gg 0$ because then the only feasible bases are associated with decisions. The selection rule breaks ties in the choice of the exiting variable when $w \geq 0$ and $w \not\gg 0$.) By contrast, the policy-improvement method amounts to carrying out block pivots in the irredundant primal program, which although not generally justified in linear programming, is valid here because of the special "totally Leontief" structure of the transpose of the irredundant primal constraint matrix L defined in the above proof.

We illustrate the simplex method in Figure 6 for a two-state problem in the space of dual inequalities. In that example there are two states $1, 2$, three actions b, c, d in state 1, and two actions b, c in state 2. There are five state-action pairs. For each state-action pair sa, we plot the line that is the set of solutions to the linear equations

$$V_s = r(s, a) + \sum_{t \in \mathcal{S}} p(t \mid s, a) V_t.$$

The value of a decision is the intersection of the two lines corresponding to the two state-action pairs defining the decision. A decision is a pair of state-action pairs, one for each state. For example, the maximum-value decision $(1d, 2b)$ takes action d in state one and action b in state two, and the maximum value is the least excessive point. One possible sequence of decisions selected by applying the simplex method to the irredundant primal is: $(1b, 2c) \to (1c, 2c) \to (1c, 2b) \to (1d, 2b)$. The heavy bold line in Figure 6 traces the progress of the simplex method in order through the values of these decisions in the space of dual variables. Notice that the action for exactly one state is changed at each iteration.

Running Times

So far we have discussed three related methods for finding maximum-value policies in transient systems, viz., successive approximations, policy improvement (including Newton's method)

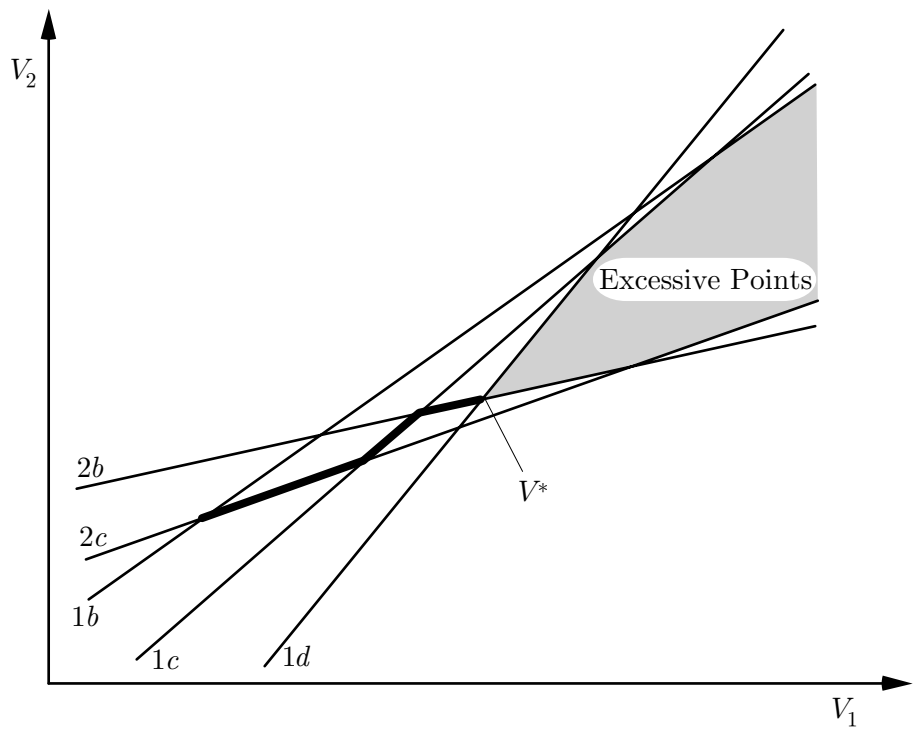


Figure 6. Progress of Primal Simplex Method in Dual Space for Two-State Problem

and linear-programming methods (including the simplex method and interior-point algorithms). Each of these methods can be used in practice to solve problems with thousands—and sometimes even millions—of states if implemented properly on a computer.

Uniformly Strictly Substochastic Systems: *Strongly Polynomial-Time Exact and Linear-Time ϵ-Approximation Algorithms.* Suppose the system is *uniformly strictly substochastic*, i.e., each of the row sums of the transition matrices is uniformly bounded by a fixed constant $0 \leq \beta < 1$. With rational data, Ye [Ye03] has shown that a maximum-value policy can be found in strongly polynomial time with interior-point methods. Even without rational data, the running time can be reduced to linear time using successive approximations if instead one is satisfied to find an ϵ-optimal policy. We outline briefly why this is so, leaving a more complete account to a homework problem.

Call a policy *ϵ-optimal* if its value is within $100\epsilon\% > 0$ of the maximum value. Call an algorithm that finds such a policy an *ϵ-approximation algorithm.* It turns out that for fixed $\epsilon > 0$, successive approximations is a strongly linear-time ϵ-approximation algorithm. That is so because the method can be implemented to find a policy that is ϵ-optimal with $O(n)$ arithmetic operations $(+, -, \times, \div, \vee, \wedge)$ where n is the number of nonzero one-period rewards and transition rates over all states and actions. To that end, it is necessary to compute $V^N = \mathcal{R}V^{N-1}$ for $N = 1, \ldots, M$ with $V^0 \equiv 0$ and $M \equiv \left\lceil \frac{\ln(1/\epsilon)}{\ln(1/\beta)} \right\rceil$ to assure that any optimal decision in the first period of an M-period problem has value within $100\epsilon\%$ of the maximum with an infinite horizon. Since each of the M needed evaluations of $\mathcal{R}V^{N-1}$ requires $O(n)$ arithmetic operations and M depends only on ϵ and β, successive approximations finds an ϵ-optimal policy in strongly linear time.

Since one must examine all the data of a transient system at least once to be assured of finding a maximum-value policy, the running time of every algorithm for finding a maximum-value policy is at least $O(n)$. Hence, under the above hypothesis and for fixed ϵ, successive approximations is as efficient as possible up to a constant factor.

Gauss-Seidel Method. One way to improve successive approximations is to update the value V after computing $(\mathcal{R}V)_s$ for a state s, rather than waiting to compute $(\mathcal{R}V)_s$ for all states s before updating V. This is the Gauss-Seidel method, and it converges faster than successive approximations.

Newton's Method. A different way to reduce the number of iterations of successive approximations is to use Newton's method. However, this method requires the added work of computing the value of the policy found at each iteration. The number of arithmetical operations to do this by Gaussian elimination is $O(S^3)$, which is higher than linear time if the number A of state-action pairs is much larger than S. If instead, one uses successive approximations to estimate these values, then the overall number of arithmetic operations to execute Newton's method falls to $O(n)$ for fixed ϵ as with successive approximations.

Balancing Pricing and Updating Values. In all three algorithms, one should take care to strike an appropriate balance between the work of pricing (computing $\mathcal{R}V$) and updating the estimate of the value V. In particular, this suggests that successive approximations is likely to be more or less efficient than Newton's method according as the number of actions in each state is respectively relatively small or large. This is simply because $\mathcal{R}V$ is relatively cheap to compute in the former case and relatively expensive to compute in the latter when compared to the work to update V. Also, successive approximations is likely to be more or less efficient than Newton's method using Gaussian elimination according as the maximum spectral radius of the transition matrices is respectively relatively small or large. This is because the work to compute the new value by successive approximations relative to Gaussian elimination is relatively large or small according as the maximum spectral radius is respectively relatively large or small.

Simplex Method. If the simplex method is used to solve the irredundant primal, it is probably not a good idea to compute $\mathcal{R}V$ at each iteration to find the best single state-action pair to improve. (For with the same work, one could use Newton's method to improve the actions in all states at once!) The reason is simply that too much work is expended in pricing. A better approach is either to price out the actions for a single state s at each iteration (which requires merely computing $(\mathcal{R}V)_s$) and perhaps cycle through the states to be priced at each iteration periodically. In this way one strikes a better balance between the work of pricing and finding a good estimate of the value.

Transient Systems: *Polynomial-Time Algorithms.* It is possible to check whether or not a substochastic system is transient in $O(n)$ time with a combinatorial algorithm that depends only on

the number n of state-actions pairs and positive transition rates. It is possible to check whether or not an arbitrary system is transient by solving the irredundant primal linear program with objective function $r(s,a) = 1$ for all $a \in A_s$ and $s \in \mathcal{S}$. If the data are rational, this linear program and its dual can be solved with $O((AS^2 + A^{1.5}S)L)$ arithmetic operations where L is the number of bits required to represent the reward and transition rates. Indeed, this can be done with an interior-point algorithm of Vaidya [Va90] and a refinement of Ye, Todd and Mizuno [YTM94]. If the system is transient, a maximum-value policy can then be found by solving the irredundant primal and dual linear programs with the same interior-point method and running time. An open question is whether these methods can be refined to run in linear time.

Stochastic Constraints. Practitioners often want to impose *stochastic constraints*, i.e., constraints on the state-action frequencies, to achieve service goals or avoid undesirable states. Here are some examples. The expected number of stockouts does not exceed σ. The expected number of backorders is at most β. The expected number of equipment failures is at most ϕ. Constraints of this type can often be expressed by appending linear inequalities to the irredundant primal. Alternately, a stochastic constraint might be necessary to assure that if the system uses an action in one of a subset of states, then the system uses the action in all states in the subset. A constraint of this type can be expressed in terms of linear inequalities in integer and continuous variables.

If stochastic constraints are imposed on the irredundant primal, then the optimal solution of the linear program will not generally be an extreme point of the irredundant primal. For this reason, it is necessary to use randomized policies. As we have seen, if the system is transient, then one way of doing this is to use an initially randomized policy. However, such a policy has the unattractive feature that the policies used after the initial randomization can have radically different performance. This feature seems artificial and is likely to be unacceptable to users.

Stationary Randomized Policies. In practice it is often more desirable to use a *stationary randomized policy* δ^∞ with $\delta = (\delta^s)$ where, for each $s \in \mathcal{S}$, $\delta^s = (\delta^{s\alpha})$ is a probability vector on A_s, i.e., a nonnegative vector on A_s whose sum is one. The interpretation is that δ^∞ chooses action $a \in A_s$ in state $s \in \mathcal{S}$ with probability δ^{sa} each time an individual enters that state. The stationary (nonrandomized) policies are equivalent to the stationary randomized policies where the probability vectors are all unit vectors.

Theorem 16. Feasible Solutions of the Irredundant Primal are State-Action Frequencies of Stationary Randomized Policies. *Suppose $w \geq 0$ and the system is transient. Then the set of feasible solutions of the irredundant primal coincides with the set of state-action frequencies of the stationary randomized policies. In particular, if x is a feasible solution of the irredundant primal, then x is the state-action frequency of the stationary randomized policy $\delta^\infty = (\delta^s)$ where the probability vector $\delta^s = (\delta^{sa})$ on A_s is given by $\delta^{sa} = x_{sa}/\sum_{\alpha \in A_s} x_{s\alpha}$ for each action $a \in A_s$ for states s with $\sum_{\alpha \in A_s} x_{s\alpha} > 0$ and δ^s is arbitrary for states s with $\sum_{\alpha \in A_s} x_{s\alpha} = 0$.*

Before turning to the proof, it is worthwhile noting why the last claim of the Theorem is intuitive. The reason is that the probability of choosing action a in state s can reasonably be expected to be the ratio of the expected number x_{sa} of individual visits to state s that choose action a therein to the expected number $\sum_{\alpha\in A_s} x_{s\alpha}$ of individual visits to state s.

Proof. We show first that each stationary randomized policy is transient. To that end, for each state $s \in \mathcal{S}$, let $\mathbb{P}_s$ be the set of probability vectors $\gamma^s = (\gamma^{sa})$ on A_s. Consider the stationary randomized policy γ^∞ where $\gamma = (\gamma^s)$ and $\gamma^s = (\gamma^{sa}) \in \mathbb{P}_s$ for each $s \in \mathcal{S}$. Let P_γ be the transition matrix of γ, i.e., the st^{th} element of P_γ is $\sum_a p(t \,|\, s, a)\gamma^{sa}$. To see that P_γ is transient, observe from Lemma 4 that because the system is transient, there is a unique vector $V \geq 1$ such that

$$V_s = \max_{a\in A_s} (1 + \sum_t p(t \,|\, s, a)V_t),\ s \in \mathcal{S}.$$

This implies that

$$V_s = \max_{\eta^s\in\mathbb{P}_s} (1 + \sum_{t,a} p(t \,|\, s, a)\eta^{sa}V_t),\ s \in \mathcal{S},$$

because, for each $s \in \mathcal{S}$, the maximum on the right-hand side of the above equation is attained by setting $\eta^{sa} = 1$ for some $a \in A_s$ and $\eta^{s\alpha} = 0$ otherwise. Now on setting $\eta^s = \gamma^s$ for each s, it follows from the above equation that $V \geq 1 + P_\gamma V$. Iterating this inequality shows that

$$V \geq \sum_{i=0}^{N-1} P_\gamma^i 1 + P_\gamma^N V \geq \sum_{i=0}^{N} P_\gamma^i 1,$$

so $\sum_{i=0}^N P_\gamma^i 1$ is increasing in N and bounded above by V, whence P_γ is transient as claimed.

Next oberve that since each stationary randomized policy is transient, its state-action frequency is finite and so feasible for the irredundant primal by the argument in the paragraph titled "state-action frequency" on page 31. Conversely, it remains to show that if x is a feasible solution of the irredundant primal, then x is the state-action frequency of the stationary randomized policy δ^∞ defined in the Theorem. To that end, let X_s be the expected number of exits from state s when using δ^∞. Since δ^∞ is transient, the X_s are finite. It remains to show that $X_s = \sum_{\alpha\in A_s} x_{s\alpha}$ for each state s. To see this, observe first that because the expected number of entries into state s equals the expected number of exits therefrom, it follows that

$$X_s - \sum_{t,a} p(s \,|\, t, a)\delta^{ta}X_t = w_s,\ s \in \mathcal{S}. \tag{5}$$

It is enough to show that this equation has a unique solution since x is feasible for the irredundant primal and so, as is readily verified by substitution, $X_s = \sum_{\alpha\in A_s} x_{s\alpha}$ is one such solution. Let P_δ be the matrix whose ts^{th} element is $\sum_a p(s \,|\, t, a)\delta^{ta}$. Rewriting (5) in matrix form yields

$$X - XP_\delta = w,$$

which as a unique solution because P_δ is transient. ∎

Example. Supply Management with Service Constraints. Consider a dynamic supply problem with an obsolescence probability, backorders up to B, say, starting inventories up to I, say, and a zero lead time for delivery of orders. Then the states are $-B, \ldots, I$. Now the constraint that the expected number of stockouts is at most σ can be expressed by the inequality

$$\sum_{s=-B}^{-1} \sum_{a} x_{sa} \leq \sigma.$$

Another possible constraint is to require that the expected fraction of the periods in which stock is short be at most f. This can be expressed as the ratio constraint

$$\frac{\sum_{s=-B}^{-1} \sum_{a} x_{sa}}{\sum_{s=-B}^{I} \sum_{a} x_{sa}} \leq f,$$

or equivalently, the linear inequality

$$\sum_{s=-B}^{-1} \sum_{a} x_{sa} \leq f \sum_{s=-B}^{I} \sum_{a} x_{sa},$$

On the other hand, the constraint that the expected number of backorders be at most β would take the form

$$-\sum_{s=-B}^{-1} s(\sum_{a} x_{sa}) \leq \beta.$$

Augmented Irredundant Primal. In order to find optimal policies with stochastic constraints, it is necessary to append the stochastic constraints to the irredundant primal program and solve the resulting program. Call this the *augmented irredundant primal*.

Number of Randomized Actions. If there are, say, k linear stochastic constraints, then each basic feasible solution of the augmented irredundant primal will have at most $S + k$ positive variables because there can be at most one basic variable for each constraint. Thus, there can be up to $S + k$ state-action pairs that are positive in a basic feasible solution of the augmented irredundant primal. In that event, the number of states in which it may be optimal to randomize can be as large as k. If k is small in comparison with S, relatively little randomization is necessary. But if k is large in comparison with S, randomization becomes important. In the supply problem above, there are only three stochastic constraints. Thus, it will be optimal to randomize in at most three initial inventory levels.

Maximum Present Value

The maximum-value criterion may be a reasonable one if the population is indifferent between receiving income in one period or another. However this is often not the case because income has a time value. One reflection of this is that it is usually possible to borrow and lend money at appropriate interest rates. Assume now that these two interest rates both equal $100\rho\%$ per period, say. In that event, the amount of money that must be invested in one period to generate one unit of money in the following period is $\beta \equiv \frac{1}{1+\rho}$. Call β the *discount factor*, reflecting the discount that one enjoys for buying one unit of money one period before receiving it. Thus, the amount of money that would be needed in period zero to generate the income stream $r^1, \ldots, r^N$ in periods $1, \ldots, N$ at the interest rate $100\rho\%$ is

$$V^{\rho N} \equiv \sum_{i=1}^{N} \beta^i r^i.$$

Call this sum the *present value* of the stream (r^i) discounted to period zero.

The above development tacitly assumes that interest is not taxed. If, as is usually the case, interest is taxed, it is more natural to let $100\rho\%$ be the *after-tax interest rate*, e.g., 60% of the original interest rate for populations in the 40% tax bracket. Then $V^{\rho N}$ is the after-tax present value of the income stream (r^i).

In the above discussion we have also ignored another important factor, viz., inflation. In inflationary times, income earned in one period has greater purchasing power than income earned later. This suggests that income streams and interest rates should be adjusted to account for inflation. To be concrete about this, suppose that $100\rho'\%$ is the ordinary after-tax interest rate and $100f\%$ is the rate of inflation. Then the *inflation-adjusted* after-tax interest rate $100\rho\%$ is given by $1 + \rho = \frac{1+\rho'}{1+f}$. Thus, even if ρ' is positive, ρ will be negative if the rate of inflation exceeds the after-tax rate of interest, a not uncommon situation. In any event $V^{\rho N}$ is the *inflation-adjusted after-tax present value* of the income stream (r^i). Since the inflation-adjusted after-tax present value is equivalent to an ordinary present value with an appropriate choice of the interest rate, we can and do simply speak about interest rates and present values in the sequel.

The above discussion leads us to consider both positive and negative interest rates. We will make the fairly mild assumption that $\rho > -1$, however. For if $\rho \leq -1$, then either $\rho = -1$, in which case money invested in one period is lost in the following period, or $\rho < -1$, in which case a unit investment in one period becomes a liability of $1 + \rho < 0$ in the following period. Both cases rarely occur, and almost never repeatedly.

The above considerations suggest that policies $\pi = (\delta_i)$ be compared on the basis of their *N-period present values* (discounted to period zero)

$$V_\pi^{\rho N} \equiv \sum_{i=1}^{N} \beta^i P_\pi^{i-1} r_{\delta_i}.$$

On setting $\widehat{P}_\delta \equiv \beta P_\delta$, $\widehat{r}_\delta \equiv \beta r_\delta$ and $\widehat{P}_\pi^i \equiv \widehat{P}_{\delta_1} \cdots \widehat{P}_{\delta_i}$, the above can be rewritten as

$$V_\pi^{\rho N} \equiv \sum_{i=1}^{N} \widehat{P}_\pi^{i-1} \widehat{r}_{\delta_i}.$$

This reduces the problem of finding a policy with maximum N-period present value to an equivalent problem of finding a policy with maximum N-period value.

Observe that $\sigma_{\widehat{P}_\delta} < 1$ if and only if $\beta \sigma_{P_\delta} < 1$. Thus the equivalent problem with transition matrices $\widehat{P}_\delta$ is transient if and only if $\beta\sigma < 1$, or equivalently, the *present value* (discounted to period zero)

$$V_\pi^\rho \equiv \sum_{i=1}^{\infty} \beta^i P_\pi^{i-1} r_{\delta_i}$$

of each policy $\pi = (\delta_i)$ is absolutely convergent for every choice of $r_\cdot$. When this is so, the *maximum-present-value problem is equivalent to a maximum-value problem* and so the results for the latter problem apply at once. In particular, there is a stationary maximum-present-value policy.

Multi-Armed Bandit Problem: Optimality of Largest-Index Rule [GJ74], [KV87], [Gi89]

A fundamental sequential statistical decision problem that attracted the interest of statisticians at least as early as the 1940s is the multi-armed bandit problem. The problem entails finding a maximum-present-value policy in a particular multidimensional Markov decision chain. The difficulty of the problem was thought to be so great that Allied scientists during World War II joked that dropping leaflets describing the problem over Axis countries might be a good way to distract their scientists from the war effort. The problem remained unsolved until 1972 when Gittins and Jones found a remarkably simple solution. There are many applications of this problem including sequential allocation of several available treatments to patients and sequential allocation of effort to several parallel projects. For definiteness in the sequel, consider the problem to be one of project scheduling.

There are N independent projects, labeled $1, \ldots, N$. In each period, only one project may be active. The states of the $N-1$ inactive projects in a period are frozen. The state s_t^n of project n at the beginning of the t^{th} period it is active is a finite-state Markov chain with stationary transition probabilities. Since the state of project n does not change while it is inactive, s_1^n is also the state of that project at the beginning of period 1. Put $s^n = (s_1^n, s_2^n, \ldots)$ and assume that s^j and s^n are independent for all $j \neq n$. If project n is active in a period when it is in state s, the project earns a reward r_s^n. Inactive projects earn no rewards. There is a discount factor $0 < \beta < 1$. A policy is a nonanticipative rule for deciding which project to activate in each period, i.e., the choice of the project to activate in a period depends only on the projects active in prior periods and the states of the projects observed in that and prior periods. This is a finite-state-and-action Markov decision chain. The goal is to find a maximum-present-value policy.

Random and Stopping Times. To do this requires a few definitions. A *random time* is a $+\infty$ or nonnegative integer-valued random variable. A *stopping time for project n* is a random time σ that is *nonanticipative* for project n, i.e., $\Pr(\sigma \le t \,|\, \boldsymbol{s})$ is independent of $s^n_{t+1}, s^n_{t+2}, \ldots$ for each nonnegative integer t and $\boldsymbol{s} = (s^1, \ldots, s^N)$. A policy may be described inductively by *selecting* for each $t = 1, 2, \ldots$ a period τ_t and a project that will be active during periods $\tau_{t-1}{+}1, \ldots, \tau_t$ where $\tau_0 \equiv 0$ and $\tau_{t-1} < \tau_t$ whenever $\tau_{t-1} < \infty$. Since policies are nonanticipative and since s^j and s^n are independent for $j \ne n$, $\tau_t + 1 - \tau_{t-1}$ is the *stopping time for the project selected* in period $\tau_{t-1}{+}1$ relative to the state of the project in that period.

Index of a Project in a State. Let $R^n_t \equiv \sum_1^t \beta^j r^n_{s^n_j}$ be the present value of the rewards that project n earns when it is active in the first t periods. Put $B_t \equiv 1 - \beta^t$. The sequel shows that B_t is proportional to $\sum_1^t \beta^j$. The *index of project n in state s* is $m^n_s \equiv \max_{\tau \ge 1} \mathrm{E}R^n_\tau / \mathrm{E}B_\tau$ where the maximum is over stopping times $\tau{+}1$ for project n in state s. Note that the indices for a particular project in its various states depend only on the rewards and transition law of the Markov chain for that project—not the others. Consequently, as the development below shows, finding the indices for the project entails finding maximum-present-value policies for a family of Markov decision chains for that project alone, one for each state.

Largest-Index Rule. The *largest-index rule* is a policy that, in each period, selects a project with maximum index. This simple policy turns out to have maximum present value. This result reduces the N-dimensional Markov decision problem to a collection of one-dimensional ones.

Present Value a Project Earns. The key fact underlying the development of this result is that for any sequence $1 \ge \alpha_1 \ge \alpha_2 \ge \cdots \ge 0$ of (Borel) functions of $\boldsymbol{s}$, there is a random time σ for which $\Pr(\sigma \ge t \,|\, \boldsymbol{s}) = \alpha_t$, for $t = 1, 2, \ldots$, so

$$\mathrm{E}R^n_\sigma = \mathrm{E}\sum_{t=1}^{\infty} \alpha_t \beta^t r^n_{s^n_t}. \tag{6}$$

For example, if k_t is the t^{th} period that project n is active when using a given policy and if σ is the random time for which $\Pr(\sigma \ge t \,|\, \boldsymbol{s}) = \alpha_t = \beta^{k_t - t}$, then (6) implies that $\mathrm{E}R^n_\sigma = \mathrm{E}\sum_{t=1}^{\infty} \beta^{k_t} r^n_{s^n_t}$ is the present value of the rewards that the policy earns from project n. Thus, the random time σ adjusts the discounting of rewards that project n earns to reflect its active periods $k_1, k_2, \ldots$.

Theorem 17. Optimality of Largest-Index Rule. *In the multi-armed-bandit problem, the largest-index rule has maximum present value.*

Proof. Let π^* be a policy that uses the largest-index rule and let $\tau_{t-1}{+}1$ be the t^{th} period in which π^* selects a project. By relabeling, assume that π^* selects project one, say, in the first period. Put $\tau \equiv \tau_1$ and $m^1 = \mathrm{E}R^1_\tau / \mathrm{E}B_\tau$. By hypothesis, $m^1 \ge m^n_{s^n_1}$, so for each stopping time $\sigma{+}1$ for project n,

$$\mathrm{E}R^n_\sigma \le m^1 \mathrm{E}B_\sigma. \tag{7}$$

Let V_π herein be the present value of any policy π starting from the given initial state. Let π' be an arbitrary policy, $\epsilon > 0$ be a number and π_0 be a policy that activates each project infinitely often and for which $V_{\pi_0} \ge V_{\pi'} - \epsilon$. Because rewards in distant periods are heavily discounted, π_0 can be formed, for example, by modifying π' to cyclically activate each project for one period every N periods starting in a sufficiently distant period. Let π_1 be the policy that permutes the order in which π_0 activates projects by shifting the first τ times that π_0 activates project one to the first τ periods.[6] Let T be the period in which π_0 activates project one for the τ^{th} time if τ is finite and let $T = \infty$ otherwise. Figure 7 illustrates these definitions for the case $N = 3$.

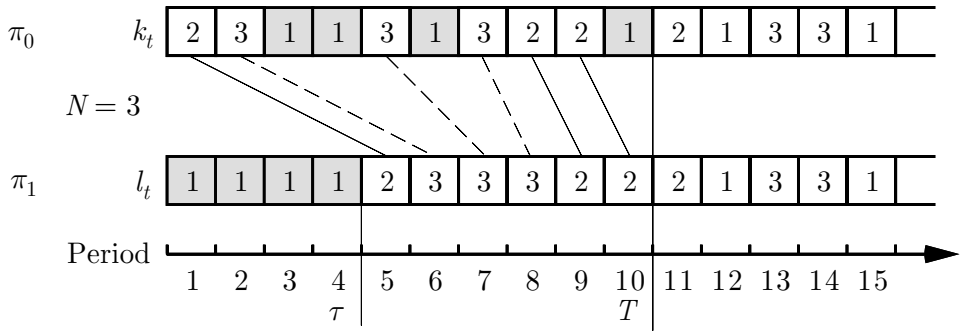


Figure 7. Active Projects in each Period

For each fixed n and $t \ge 1$, let k_t (resp., l_t) be the period in which π_0 (resp., π_1) activates project n for the t^{th} time. For $1 \le k_t \le T$, set $\alpha_t = \beta^{k_t - t}$ if $n = 1$ and $\alpha_t = \beta^{k_t - t} - \beta^{l_t - t}$ if $n > 1$. For $k_t > T$, $k_t = l_t$ and set $\alpha_t = 0$. Observe that $0 \le \alpha_t \le 1$ is decreasing in $t \ge 1$ because $k_t - t > 0$ is increasing and, if $n > 1$, $k_t - l_t \le 0$ is increasing in $t \ge 1$. Thus, from (6), there is a random time $\sigma^n \le T$ with $\Pr(\sigma^n \ge t \,|\, \boldsymbol{s}) = \alpha_t$ for $t \ge 1$. Also by (6) and the fact that $k_t = l_t$ for $k_t > T$, the difference between the present values that π_0 and π_1 earn from project n is $\mathrm{E}\sum_1^\infty (\beta^{k_t} - \beta^{l_t}) r^n_{s^n_t}$ $= \mathrm{E}\sum_1^T (\beta^{k_t} - \beta^{l_t}) r^n_{s^n_t}$, which is $\mathrm{E}R^1_{\sigma^1} - \mathrm{E}R^1_\tau$ if $n = 1$ and $\mathrm{E}R^n_{\sigma^n}$ if $n > 1$. Hence,

$$V_{\pi_1} - V_{\pi_0} = \mathrm{E}R^1_\tau - \sum_{n=1}^N \mathrm{E}R^n_{\sigma^n}. \tag{8}$$

If the reward that each active project earns in each state is instead $(1 - \beta)/\beta$, the left-hand side of (8) vanishes, so because $\sum_1^t \beta^j (1 - \beta)/\beta = 1 - \beta^t$, (8) reduces to

$$0 = \mathrm{E}B_\tau - \sum_{n=1}^N \mathrm{E}B_{\sigma^n}. \tag{9}$$

[6]For π_1 to be a policy, it must be nonanticipative. That this is so may be shown by noting that π_1 essentially consists of activating project one for the first τ periods and then using π_0 afresh thereafter, being careful to skip the first τ times that π_0 activates project one.

Now for each n, $\sigma^n{+}1 \geq 1$ is a stopping time for project n since $\Pr(\sigma^n{+}1 \leq t \,|\, \boldsymbol{s}) = 1 - \alpha_t$ is independent of $s^n_{t+1}, s^n_{t+2}, \ldots$. Hence, from (7)-(9) and the definition of m^1,

$$V_{\pi_1} - V_{\pi_0} \geq m^1(\mathrm{E}B_\tau - \sum_{n=1}^{N} \mathrm{E}B_{\sigma^n}) = 0.$$

Using the above construction, it follows by induction on $t \geq 1$ that there exist policies π_t that agree with π^* through period $\tau_t \geq t$ and have the property that $V_{\pi_t} - V_{\pi_{t-1}} \geq 0$. Thus, $V_{\pi_t} - V_{\pi_0} \geq 0$ and $V_{\pi^*} = \lim_{t\to\infty} V_{\pi_t} \geq V_{\pi_0} \geq V_{\pi'} - \epsilon$. Since ϵ is arbitrary, $V_{\pi^*} \geq V_{\pi'}$. ∎

Restart Problem. To facilitate computation, it is useful to characterize the index of a project in state s, say, as the value of the project in an associated "restart-in-s" problem. For notational simplicity, drop the superscript designating the project in the sequel without further mention. Now the *restart-in-s* problem entails choosing a maximum-present-value stopping time $\tau{+}1$ in state s for the problem in which one starts in state s, chooses the stopping time $\tau{+}1$ and restarts afresh in state s in period $\tau{+}1$. If v_s has maximum present value for this problem, then $v_s = \max_{\tau\geq 1} \mathrm{E}(R_\tau + \beta^\tau v_s)$. Now if $v_{\tau s}$ is the (expected) present value of rewards that the stopping time $\tau{+}1$ earns, then $v_{\tau s} = \mathrm{E}R_\tau + (\mathrm{E}\beta^\tau)v_{\tau s}$, so $v_{\tau s} = \mathrm{E}R_\tau/\mathrm{E}B_\tau$. Hence, $v_s = \max_{\tau\geq 1} v_{\tau s} = m_s$ is the index of the project in state s.

In order to examine the computational efficiency of this formulation, suppose that the sequence of states that the project enters is a Markov chain on the finite state-space $\mathcal{S}$ of size S. The above approach to the restart-in-s problem looks at the system only when it is in state s and chooses a stopping time. Because the system is Markovian, it suffices to limit attention to a nonrandomized Markovian stopping time, or equivalently a restarting set, i.e., a set of states in which to restart the system in state s. Though this approach requires only one state, the number of actions (restarting sets) in that state is 2^S, and so grows geometrically with S.

A better approach to this problem is to retain the original state-space of the Markov chain. Then there are only two actions in each state j, viz., continue or restart in s. To implement this approach, let P_j be the j^{th} row of the transition matrix of the Markov chain on $\mathcal{S}$. Then the maximum-present-value vector for this form of the restart-in-s problem is the unique solution $v = (v_j)$ of

$$v_j = \beta\max(r_j + P_j v, r_s + P_s v),\ j \in \mathcal{S}. \tag{10}$$

Of course, $m_s = v_s$ is the desired index of the project in state s.

Size Comparison. Suppose that the sizes of the state-spaces of the N Markov chains are each S. Then the original N-dimensional problem has S^N states, N actions in each state and N^{S^N} state-action pairs. By contrast, the largest-index rule requires solving SN restarting problems, one for each state of each project, with each such problem having S states, two actions and $2S$

state-action pairs. For example, if $S = N = 4$, the original problem would have 256 states and 2^{512} state-action pairs. By contrast the largest-index rule would require solving 16 restarting problems each with 4 states and 8 state-action pairs.

On-Line Computation. The "discounted" computational effort required to solve the restarting problems to find the needed indices can be reduced significantly if one computes *on-line*. Initially one must solve S restarting problems. Thereafter, in order to implement the largest-index rule, it suffices to wait until the state of the project that was active in the previous period is observed to be in the restart set and then solve the restart problem for that project and state only. This entails solving at most one restarting problem in each period. Thus the "discounted" number of restarting problems that must be solved is at most $\beta(N + \beta/(1-\beta))$, and so is independent of S.

8 MAXIMUM PRESENT VALUE WITH SMALL INTEREST RATES IN BOUNDED SYSTEMS

[Bl62], [Ve66], [MV69], [Ve69a], [Ve74], [Ro75c], [RV92]

When a decision maker is indifferent between receiving income in different periods, attention shifts to problems in which the interest rate is zero. Problems of this type are quite important. Some examples appear below. In the first, the inflation-adjusted after-tax interest rate is zero. In all but the first two, there is no interest rate. However, it is useful to introduce small interest rates to address these problems over long time horizons.

- **Zero Inflation-Adjusted After-Tax Interest Rate.** Suppose that the interest rate is 5%, the marginal tax rate is 40% and the rate of inflation is 3%. Then the after-tax rate of interest is 3% which coincides with the rate of inflation. Consequently, the inflation-adjusted after-tax interest rate is zero.

- **Maximum Present Value for an Interval of Interest Rates.** A decision maker may wish to find a policy that simultaneously has maximum present value for all small enough interest rates majorizing a given nonnegative interest rate.

- **Maximize Sustainable Yield of a Renewable Resource.** One may wish to maximize the long-term sustainable yield of fish, crops, lumber, etc.

- **Maximize Expected Instantaneous Rate of Return.** This problem arises in infinite-horizon versions of Example 5 on p. 8.

- **Maximize Throughput of a Factory.** The general manager of a factory may wish to maximize its throughput.

- **Maximize Yield of a Production Process.** A production engineer may wish to maximize the yield of chips from a fab plant.

- **Maximize Expected Number of Patients Treated Successfully.** A hospital may wish to maximize the expected number of patients treated successfully.

- **Maximize Expected Number of On-Time Deliveries.** A trucking, shipping, or air-cargo company may wish to maximize the expected number on-time deliveries.

- **Maximize the Expected Number of Customers Served.** A firm may wish to maximize the expected number of satisfied customers.

- **Minimize Expected Number of Fatalities.** A public agency may wish to minimize the expected number of fatalities from various hazards.

For transient systems, the maximum-present-value problem with zero interest rates is precisely the maximum-value problem discussed in §1.7. But for nontransient systems, the value of a policy is typically not well defined or is infinite. For this reason, generalizing the notion of maximum value from transient to nontransient systems requires a different formulation.

Bounded System

Assume throughout this section that the system is *bounded*, i.e., its system degree d is zero or one. For example, substochastic systems are bounded because then P_π^N is substochastic for each N and so is bounded. However, not all bounded systems are substochastic. In bounded systems, the present value V_π^ρ of a policy π is well defined and finite for all $\rho > 0$ even though that is not generally the case where $\rho = 0$. However, for nontransient bounded systems with $\rho = 0$, it is not generally possible or useful to compare policies by means of their present values because those values may not be well defined or finite.

Strong Maximum Present Value

One way of resolving this dilemma is to consider present-value preference relations for small (positive) interest rates. Such preference relations are of special interest when decisions are made frequently, say every week or day. As will be seen subsequently, the strongest preference relation of this type is the following. A policy λ has *strong maximum present value* if λ has simultaneous maximum present value for all sufficiently small positive interest rates, i.e., there is a $\rho^* > 0$ such that

$$V_\lambda^\rho - V_\pi^\rho \geq 0 \text{ for all } 0 < \rho < \rho^* \text{ and all } \pi. \tag{1}$$

This preference relation will play a fundamental role in what follows, especially for systems in which there is exogenous immigration. The development to follow shows that stationary strong maximum-present-value policies exist and generalizes the policy improvement method to find such a policy.

Cesàro Limits and Neumann Series

First, however, it is useful to generalize the Neumann series expansion of Lemma 2 from transient to bounded systems. If $P_0, P_1, \ldots$ and P are $S \times S$ complex matrices and $\widehat{P}_N \equiv \sum_{i=0}^N P_i$, write $P_N \to P$ $(\mathrm{C},1)$ if $N^{-1}\widehat{P}_{N-1} \to P$. Similarly, write $\sum_{N=0}^\infty P_N = P$ $(\mathrm{C},1)$ and say that the first series *converges* $(\mathrm{C},1)$ if $\widehat{P}_N \to P$ $(\mathrm{C},1)$, or equivalently, $N^{-1}\sum_{i=0}^{N-1} \widehat{P}_i \to P$. On the other hand, write $P_N \to P$ $(\mathrm{C},0)$ if $P_N \to P$. Incidentally, $(\mathrm{C},0)$ (resp., $(\mathrm{C},1)$) stands for *Cesàro limit of order zero* (resp., *one*). Thus Cesàro limits of order zero are ordinary limits.

It is easy to see that if a sequence converges $(\mathrm{C},0)$, the sequence also converges $(\mathrm{C},1)$, and the two limits coincide. However the converse is generally false. For example, every periodic sequence converges $(\mathrm{C},1)$. But such a sequence does not converge $(\mathrm{C},0)$ except in the trivial case

of a constant sequence. Thus, one important feature of $(\mathrm{C},1)$ convergence is that it assures that a broader class of sequences converges than does ordinary $(\mathrm{C},0)$ convergence.

In the above terminology, Lemma 2 asserts that if P is a square complex matrix and $P^N \to 0$ $(\mathrm{C},0)$, then $I-P$ is nonsingular and has a $(\mathrm{C},0)$ Neumann series expansion. The next lemma asserts that this result remains valid for $(\mathrm{C},1)$ limits as well.

Lemma 18. Cesàro Neumann Series. *If P is a square complex matrix and $P^N \to 0$ $(\mathrm{C},1)$, then $I-P$ is nonsingular and*

$$(I-P)^{-1} = \sum_{N=0}^{\infty} P^N \ (\mathrm{C},1). \tag{2}$$

Proof. Recall as in the proof of Lemma 2 that

$$I - P^j = \sum_{i=0}^{j-1}(P^i - P^{i+1}) = (I-P)\sum_{i=0}^{j-1} P^i = (I-P)\widehat{P}_{j-1}$$

where $\widehat{P}_j \equiv \sum_{i=0}^{j} P^i$. Summing the above equation over j from 1 to N and dividing by N, yields

$$I - N^{-1}\widehat{P}_{N-1}P = (I-P)(N^{-1}\sum_{i=0}^{N-1}\widehat{P}_i),\ N=1,2,\ldots. \tag{3}$$

Thus if $P^N \to 0$ $(\mathrm{C},1)$, then the left hand side of (3) is nearly I for large enough N and so is nonsingular. Hence $I-P$ is nonsingular. Now premultiply (3) by $(I-P)^{-1}$ and let $N \to \infty$. ∎

In order to appreciate the significance of this result, observe that since $P^N \to 0$ $(\mathrm{C},0)$ implies $P^N \to 0$ $(\mathrm{C},1)$, it follows that (2) will certainly hold whenever $P^N \to 0$ $(\mathrm{C},0)$. But (2) holds in other cases as well. For example, if $P=-I$, then $P^N \to 0$ $(\mathrm{C},1)$. In that event (2) asserts that $(\frac{1}{2})I = \sum_{N=0}^{\infty}(-I)^N$ $(\mathrm{C},1)$. Moreover, the last two $(\mathrm{C},1)$ limits cannot be replaced by $(\mathrm{C},0)$ limits.

Stationary and Deviation Matrices

The next two results characterize two important matrices associated with a square matrix P.

Theorem 19. Stationary Matrix. *If P is a square complex matrix with $d_P \le 1$, then there is a stationary matrix P^* for P such that*

$$P^N \to P^* \ (\mathrm{C}, d_P). \tag{4}$$

Moreover,

$$PP^* = P^*P = P^*P^* = P^*. \tag{5}$$

If P is real (resp., nonnegative, substochastic, stochastic), then so is P^*. Also, $P^*=0$ if, and provided that P is nonnegative, only if P is transient. Finally, $P=I$ if and only if $P^*=I$.

Proof. If $d_P = 0$, then $P^* = 0$ and the result is clear. Suppose $d_P = 1$. Put $B_N \equiv N^{-1}\sum_{i=0}^{N-1} P^i$. Then $\{P^N\}$ and $\{B_N\}$ are bounded. Hence, to prove (4), it suffices to show that the limit points of $\{B_N\}$ coincide. (A *limit point* of a sequence of matrices is, of course, the limit of a subsequence of the matrices). Let J be any limit point of $\{B_N\}$. Then $B_N + N^{-1}(P^N - I) = PB_N = B_N P$ have the common limit point $J = PJ = JP$. Thus $J = P^N J = JP^N$ and so $J = B_N J = JB_N$. Now let K be any limit point of $\{B_N\}$. Then the preceding equalities have the common limit point $J = KJ = JK$. Since J and K are arbitrary limit points, they can be interchanged, giving $K = JK = KJ$. Thus $J = K$, so (4) holds. Moreover, this also establishes (5). We claim that if P is nonnegative and $P^* = 0$, then P is transient. To show this, observe from Lemma 18 that $D \equiv (I - P)^{-1}$ exists and is nonnegative, so $D = I + PD = \cdots = \sum_0^N P^i + P^{N+1}D \geq \sum_0^N P^i$ for all N. Thus, since P is nonnegative, it is transient as claimed. The remaining assertions follow from (4) and (5). ∎

In view of (4), when P is nonnegative, the st^{th} element of P^* can be interpreted as the long run expected average number of individuals in state t per period generated by one individual starting in state s. Notice from $P^*P = P^*$ in (5) that each row of P^* is left stationary when postmultiplied by P. Thus, by (4), the s^{th} row of P^* is the *stationary distribution* of the associated Markov population chain to which the N-step transition rates from state s converge (C, d_P).

Theorem 20. Deviation Matrix. *If P is an $S \times S$ complex matrix with $d_P \leq 1$, then $P^* - Q$ is nonsingular where P^* is the stationary matrix for P and $Q \equiv P - I$. Also the deviation matrix D for P, defined by*

$$D \equiv (P^* - Q)^{-1}(I - P^*) = (I - P^*)(P^* - Q)^{-1}, \tag{6}$$

and P^* satisfy

$$D = \sum_{N=0}^{\infty} (P^N - P^*) \ (\mathrm{C}, d_P) \tag{7}$$

$$DP^* = P^*D = 0 \tag{8}$$

and

$$(P^* - Q)^{-1} = P^* + D. \tag{9}$$

Moreover, $\mathbb{C}^S$ ($\mathbb{C}$ is the complex numbers) is the direct sum of the ranges of P^* and D. The sum of the ranks of P^* and D is S. Also $P^* = 0$ if and only if $D = (I - P)^{-1}$. And $P^* = I$ if and only if $D = 0$. If further P is real, then so is D.

Proof. It follows readily from (5) that

$$(P - P^*)^N = P^N - P^* \text{ for } N \geq 1.$$

On setting $B = P - P^*$, it is apparent from (4) and the above equation that $B^N \to 0$ (C, d_P). Hence, $P^* - Q = I - B$ is nonsingular by Lemmas 2 and 18, so by the above facts,

$$(10) \qquad (P^* - Q)^{-1} = \sum_{N=0}^{\infty} (P - P^*)^N = \sum_{N=0}^{\infty} (P^N - P^*) + P^* \ (\mathrm{C}, d_P).$$

Now postmultiply (resp., premultiply) this equation by $I - P^*$ and use (5) to establish (6), (7) and the fact that $I - P^*$ and $(P^* - Q)^{-1}$ commute. Next postmultiply (resp., premultiply) (6) by P^* and use (5) to establish (8). Also, (9) follows from (7) and (10). In view of (9), $\mathbb{C}^S$ is the direct sum of the ranges of P^* and D provided $x = P^*y = Dz$ implies that $x = 0$. To see that the last is so, observe from (8) that $(P^* + D)x = P^*Dz + DP^*y = 0$, whence $x = 0$ by (9). The characterizations of $P^* = 0$ and $P^* = I$ follow from (6) and (8). The fact that P is real implies that D is real follows from (4) and (6). ∎

Notice from (5) and (6) that if P is nonnegative, the st^{th} element of D is the (C, d_P) limit of the amount by which the sum of the expected N-period sojourn times of all individuals in state t generated by one individual starting in state s exceeds that when the system starts instead with the stationary population distribution for state s. Thus D measures the deviation from stationarity.

Laurent Expansion of Resolvent

Suppose that P is a square complex matrix and that $d_P \leq 1$, so $\sigma_P \leq 1$. Then $\rho I - Q = (1 + \rho)I - P$ is nonsingular for $\rho > 0$ because $(1 + \rho)^{-1}P$ has spectral radius less than one. Call $R^\rho \equiv (\rho I - Q)^{-1}$ the *resolvent* of Q for $\rho > 0$. Also $R^\rho = \beta(I - \beta P)^{-1} = \sum_{N=0}^{\infty} \beta^{N+1} P^N$ where $\beta \equiv \dfrac{1}{1+\rho}$. Therefore, if P is real and nonnegative, then R^ρ is the matrix whose st^{th} element is the sum of the expected discounted sojourn times of all individuals in state t generated by one individual in state s initially where β is the discount factor and $100\rho\%$ is the interest rate.

In order to study policies with maximum present value for small interest rates $\rho > 0$, it is useful to develop an expansion of R^ρ in ρ. It turns out that the desired expansion is a Laurent series about the origin whose coefficients are the stationary matrix P^* and powers of the deviation matrix D.

Theorem 21. Laurent Expansion of Resolvent. *If P is a square complex matrix with $d_P \leq 1$ and if P^* and D are respectively the stationary and deviation matrices for P, then for all $\rho > 0$ with $\rho\sigma_D < 1$,*

$$(11) \qquad R^\rho = \rho^{-1}P^* + \sum_{n=0}^{\infty} (-\rho)^n D^{n+1}.$$

If $P^* = 0$, then $R^\rho = \sum_{n=0}^{\infty} (-\rho)^n (I - P)^{-n-1}$. If $D = 0$, then $R^\rho = \rho^{-1}I$. If P is real, so is R^ρ.

Proof. It follows from Theorem 19 that $QP^* = 0 = P^*Q$, so $(\rho I - Q)P^* = \rho P^* = P^*(\rho I - Q)$. Premultiplying the next-to-last equation by R^ρ and postmultiplying the last equation by R^ρ yields

$$(12) \qquad R^\rho P^* = \rho^{-1}P^* = P^*R^\rho.$$

Also, from Theorems 19 and 20,

$$(I - P^*)(I + \rho D) = (P^* - Q)D + \rho D = (\rho I - Q)D. \tag{13}$$

Now for all $\rho > 0$ with $\rho\sigma_D < 1$, $I + \rho D$ is nonsingular, so its inverse exists and has a Neumann series expansion. Thus, premultiplying (13) by R^ρ and postmultiplying (13) by $(I + \rho D)^{-1}$ yields

$$R^\rho(I - P^*) = D(I + \rho D)^{-1} = \sum_{n=0}^{\infty} (-\rho)^n D^{n+1}. \tag{14}$$

Now add the first equation in (12) to (14). The rest follows from (10). ∎

Laurent Expansion of Present Value in Interest Rate

Suppose that the system is bounded. Then on putting $Q_\delta \equiv P_\delta - I$ and $R^\rho_\delta \equiv (\rho I - Q_\delta)^{-1}$, it follows that the present value of δ is

$$V^\rho_\delta = \Big(\sum_{N=0}^{\infty} \beta^{N+1} P^N_\delta \Big) r_\delta = \beta(I - \beta P_\delta)^{-1} r_\delta = R^\rho_\delta r_\delta.$$

This fact and Theorem 21 imply that for all $\rho > 0$ with $\rho\sigma_D < 1$,

$$V^\rho_\delta = \sum_{n=-d}^{\infty} \rho^n v^n_\delta \tag{15}$$

where $v^{-1}_\delta \equiv P^*_\delta r_\delta$ and $v^n_\delta \equiv (-1)^n D^{n+1}_\delta r_\delta$ for $n = 0, 1, \ldots,$ with P^*_δ and D_δ being respectively the stationary and deviation matrices for P_δ, and d is the system degree. Evidently (15) gives the desired Laurent expansion of V^ρ_δ in ρ.

In order to obtain a clearer understanding of the fundamentally important Laurent expansion (15) of the present value of a stationary policy δ^∞ for small interest rates, it is useful to interpret briefly the quantities appearing in the expansion. To that end, observe first that the present value V^ρ_δ of δ is the product of the expected present value R^ρ_δ of the sojourn times of all individuals in each state and the unit rewards per period r_δ that individuals earn in those states.

If $d = 1$, the first coefficient in the Laurent expansion of V^ρ_δ is v^{-1}_δ. In order to interpret this term, observe that depositing V^ρ_δ in a bank that pays interest at the rate $100\rho\%$ earns the interest payment ρV^ρ_δ each period in perpetuity. Thus the present value V^ρ_δ is equivalent to the infinite stream of equal interest payments ρV^ρ_δ in each period. Hence, it follows from (15) that the limit of the interest payments in each period as the interest rate approaches zero is v^{-1}_δ.

There is a related interpretation of v^{-1}_δ that may be described as follows. Observe from $V^\rho_\delta = R^\rho_\delta r_\delta$, (12) and (15) that $P^*_\delta V^\rho_\delta = \rho^{-1} v^{-1}_\delta$ is the present value of δ starting with its stationary distribution in each state. Thus, v^{-1}_δ is also the interest payment in each period resulting from depositing $P^*_\delta V^\rho_\delta$ in a bank that pays interest at the rate $100\rho\%$.

If $d = 1$, the second coefficient in the expansion (15) of V^ρ_δ is v^0_δ. It can be interpreted by observing from (15) that $V^\rho_\delta = \rho^{-1} v^{-1}_\delta + v^0_\delta + o(1)$ where $o(1)$ *is a function of* ρ *that converges to zero as* $\rho \downarrow 0$. Thus v^0_δ is the limit as $\rho \downarrow 0$ of the difference $V^\rho_\delta - \rho^{-1} v^{-1}_\delta$ between the present values of δ starting in each state and starting with the stationary distribution for that state.

Characterization of Stationary Strong Maximum-Present-Value Policies as ∞-Optimal

The Laurent expansion of the present value of a stationary policy in the interest rate plays a key role in characterizing stationary strong maximum-present-value policies in a bounded system. To see this requires a definition. Call a real matrix B *lexicographically nonnegative*, written $B \succeq 0$, if the first nonvanishing element (if any) of each row of B is positive. Similarly, write $B \succ 0$ if $B \succeq 0$ and $B \neq 0$. Also write $B \preceq 0$ (resp., $B \prec 0$) if $-B \succeq 0$ (resp., $-B \succ 0$).

Let $V_\delta \equiv (v_\delta^{-d}\ v_\delta^{-d+1} \cdots)$ be the matrix of coefficients of the Laurent expansion (15) of V_δ^ρ. This matrix has S rows and infinitely many columns. The reader should note that V_δ differs from its earlier interpretation as the value of a transient policy δ^∞. Indeed, it follows from (15) that in the present notation, the value of δ^∞ would then be $\lim_{\rho \downarrow 0} V_\delta^\rho = v_\delta^0$ because in that event $v_\delta^{-1} = 0$. Thus in the sequel the reader will need to be careful to remember which of the two interpretations of V_δ is intended. The reason for the twin interpretations of V_δ is that many of the following results regarding strong present-value optimality for bounded systems have the same form as for maximum-value optimality for transient systems.

Now observe from (15) that

$$V_\delta^\rho - V_\gamma^\rho = \sum_{n=-d}^{\infty} \rho^n (v_\delta^n - v_\gamma^n). \tag{16}$$

This difference is nonnegative, i.e., the present value of δ is at least as large as that for γ, for all $\rho > 0$ with $\rho \max_{\eta \in \Delta} \sigma_{D_\eta} < 1$ if and only if $V_\delta - V_\gamma \succeq 0$. The reason for this is that for all such positive ρ, the sign of each row of (16) is simply the sign of the first nonvanishing coefficient of the same row of the Laurent expansion. Thus since there is a stationary maximum-present-value policy for each $\rho > 0$, it follows that δ^∞ *has strong maximum present value if and only if* δ^∞ *is* ∞*-optimal*, i.e., $V_\delta \succeq V_\gamma$ for all $\gamma \in \Delta$. Let Δ_∞ be the set of $\delta \in \Delta$ for which δ^∞ is ∞-optimal.

Application to Controlling Service and Rework Rates in a $G/M/\infty$ Queue

As an illustration of the above ideas, consider the following application to controlled $G/M/\infty$ queues. Suppose jobs arrive for service at a machine shop at the beginning of each hour according to an arbitrary stochastic process. Each job may be processed on a machine during the hour following arrival by a skilled or unskilled worker since the numbers of machines and of workers of each type is adequate to serve all arriving jobs (this amounts to assuming that there are infinitely many channels in the queue). Unskilled workers earn w per hour while working on a job and skilled workers earn twice that much. A job completed by a worker is *shipped* and earns the shop r if the job is acceptable and is *reworked* (i.e., reprocessed during the next hour) if it is not. Each job completed by an unskilled worker requires rework with probability p, $\frac{1}{2} < p < 1$, whether or not the job was previously reworked. By contrast each job completed by a skilled worker requires rework with probability $2p - 1$. Thus, the probability of shipping a job

completed by a skilled worker is $2(1-p)$, which is double the corresponding probability $1-p$ for an unskilled worker. The problem, which is illustrated in Figure 8, is to decide whether to use skilled or unskilled workers to process jobs when the goal is to achieve strong maximum present value. In short, the issue is whether it is better to use a slow worker with low pay or a fast worker with high pay.

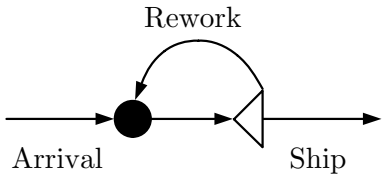


Figure 8. Optimal Control of Service and Rework Rates in a $G/M/\infty$ Queue

Initially it is convenient to ignore the arrival process and instead assume that there is a single job awaiting service. Observe that this is a single-state system in which there are two decisions υ and σ corresponding respectively to using an unskilled or skilled worker to process the job. Put $\Delta = \{\upsilon, \sigma\}$. Then the one-period reward and transition rates are

$$r_\upsilon = (1-p)r - w,\ P_\upsilon = p,\ r_\sigma = 2r_\upsilon \text{ and } P_\sigma = 2p-1.$$

For each $\delta \in \Delta$, denote by V_δ^ρ the expected present value of net income (revenue minus wages) received from a job when a worker of type δ processes the job until it is shipped and the hourly interest rate is $\rho > 0$.

Notice that since $p \vee (2p-1) < 1$, the system is transient and so $d = 0$. Thus, $P_\delta^* = 0$ and $D_\delta = (1-P_\delta)^{-1}$ for $\delta \in \Delta$, so $D_\upsilon = \dfrac{1}{1-p}$ and $D_\sigma = \dfrac{1}{2(1-p)}$. Also, $v_\delta^0 = D_\delta r_\delta$ and $v_\delta^1 = -D_\delta v_\delta^0$ for $\delta \in \Delta$, so $v_\upsilon^0 = v_\sigma^0 \equiv v^0 \equiv r - \dfrac{w}{1-p}$, $v_\upsilon^1 = -D_\upsilon v^0$ and $v_\sigma^1 = -D_\sigma v^0$. Hence

$$V_\sigma^\rho - V_\upsilon^\rho = (D_\upsilon - D_\sigma)v^0\rho + o(\rho) = \frac{v^0\rho}{2(1-p)} + o(\rho). \tag{17}$$

Observe that the expected profit v^0 when a skilled worker processes a job is the same as that for an unskilled worker, i.e., a job is profitable or unprofitable independently of the type of worker who processes the job. Thus if the goal of the shop is to use a maximum-value policy (with $\rho = 0$), the shop will be indifferent between using a skilled or an unskilled worker. However, if instead the shop seeks a strong maximum-present-value policy, it follows from (17) that it is optimal to use a skilled worker if a job is profitable and to use an unskilled worker if the job is unprofitable! The explanation for this is that if a job is profitable, it is better to use a skilled worker who earns the profit earlier when the expected discount factor is larger, while if a job is unprofitable, it is better to use an unskilled worker who incurs the loss later when the expected discount factor is smaller.

But what if a job is neither profitable nor unprofitable, i.e., $v^0 = 0$? In that event, $v^n_\delta = (-1)^n D^n_\delta v^0 = 0$ for all $n \geq 0$, whence $V^\rho_\sigma - V^\rho_v = 0$ for all $\rho > 0$, i.e., the shop is indifferent between using a skilled or an unskilled worker for all positive interest rates.

Incidentally, so far no consideration has been given to the effect of the arrival process into the queue. Does it matter? The answer is it does as the developments in the sequel will show. But for now, return to the general theory.

Strong Policy-Improvement Method

It is now possible to generalize the policy-improvement method to find a strong maximum present-value policy for bounded systems. To that end, it is useful first to adapt the Comparison Lemma to the present-value problem.

Comparison Lemma for Present Values. Observe that the maximum-present-value problem for a bounded system is equivalent to a maximum-value problem for a transient system with one-period reward vectors βr_δ and transition matrices βP_δ for each $\delta \in \Delta$. Then on applying the Comparison Lemma for transient systems to the maximum-present-value problem, the resulting Comparison Lemma for stationary policies with $\rho > 0$ becomes

$$V^\rho_\gamma - V^\rho_\delta = R^\rho_\gamma G^\rho_{\gamma\delta}, \tag{18}$$

where $G^\rho_{\gamma\delta} \equiv r_\gamma + Q_\gamma V^\rho_\delta - \rho V^\rho_\delta$ is the *present-value comparison function*. For a direct proof of this fact, observe that $V^\rho_\gamma - V^\rho_\delta = R^\rho_\gamma[r_\gamma - (\rho I - Q_\gamma)V^\rho_\delta] = R^\rho_\gamma G^\rho_{\gamma\delta}$. Also, note that in the special case in which γ and δ are transient, setting the interest rate $\rho = 0$ reduces (18) to the Comparison Lemma for stationary transient policies.

Laurent Expansion of Present-Value Comparison Function. It is now possible to give the desired generalization of the policy-improvement method. Observe that if the goal were merely seeking an improvement of δ for a single fixed positive interest rate $\rho > 0$, it would be enough to choose γ so that $G^\rho_{\gamma\delta} > 0$ for that single value of ρ. For then by (18), $V^\rho_\gamma > V^\rho_\delta$. However, the goal now is to be sure that γ is an improvement of δ simultaneously for all sufficiently small positive ρ. One way to achieve this is to choose γ so that $G^\rho_{\gamma\delta} > 0$ simultaneously for all sufficiently small positive $\rho > 0$. To find such a γ, it is useful to develop the Laurent expansion of $G^\rho_{\gamma\delta}$ in $\rho > 0$. To that end, substitute the Laurent expansion (15) of V^ρ_δ into $G^\rho_{\gamma\delta}$ and collect terms of like powers of ρ. The result is that

$$G^\rho_{\gamma\delta} = \sum_{n=-d}^{\infty} \rho^n g^n_{\gamma\delta} \tag{19}$$

for all small enough $\rho > 0$ where $g^n_{\gamma\delta} \equiv r^n_\gamma + Q_\gamma v^n_\delta - v^{n-1}_\delta$ for $n = -d, -d+1, \ldots,\ r^n_\gamma \equiv 0$ for $n \neq 0$, $r^0_\gamma \equiv r_\gamma$, $v^{-d-1}_\delta \equiv 0$ and d is the system degree. Let $G_{\gamma\delta} \equiv (g^{-d}_{\gamma\delta}\ g^{-d+1}_{\gamma\delta}\ g^{-d+2}_{\gamma\delta}\ \cdots)$ be the matrix of coefficients of the Laurent expansion (19). Observe that $G_{\gamma\delta}$ is a matrix with S rows and infin-

itely many columns. The reader is warned again that $G_{\gamma\delta}$ generalizes its earlier interpretation as comparing the values of the transient policies (γ, δ^∞) and δ^∞ to comparing the present values of those policies for all small enough positive interest rates. Moreover, in the present notation and under the hypothesis that the system is transient, $\lim_{\rho\downarrow 0} G^\rho_{\gamma\delta} = g^0_{\gamma\delta} = r_\gamma + P_\gamma v^0_\delta - v^0_\delta$ which, because v^0_δ is then the value of δ, is the difference between the values of the transient policies (γ, δ^∞) and δ^∞.

Strong Policy Improvement. It follows from (19) that $G^\rho_{\gamma\delta} > 0$ (resp., ≤ 0) for all small enough $\rho > 0$ if and only if $G_{\gamma\delta} \succ 0$ (resp., $G_{\gamma\delta} \preceq 0$). Since $R^\rho_\delta \geq \beta I$ for $\rho > 0$, it follows from (16), (18) and (19) that $V_\gamma - V_\delta \succ 0$ (resp., $\preceq 0$) if $G_{\gamma\delta} \succ 0$ (resp., $\preceq 0$). For this reason, say that γ *strongly improves* δ if $G_{\gamma\delta} \succ 0$. Now observe from (18) that since $V^\rho_\delta - V^\rho_\delta = 0$ and R^ρ_δ is nonsingular, $G^\rho_{\delta\delta} = 0$ for all $\rho > 0$. Thus, $G_{\delta\delta} = 0$ by (19). Hence, if no decision strongly improves δ, then, analogously with the corresponding claim for the transient case, $G_{\gamma\delta} \preceq 0$ for all γ. To see why this is so, observe that it is possible to improve the action in one state without affecting the others. This is because the s^{th} row $G_{\gamma\delta s}$ of $G_{\gamma\delta}$ depends on the action γ^s that γ takes in state s, but not on the action γ^t that γ takes in any other state t. Thus, if there is any state s and action $\gamma^s \in A_s$ for which $G_{\gamma\delta s}$ is lexicographically positive, then choosing $\gamma^t = \delta^t$ for all $t \neq s$ will assure that $G_{\gamma\delta} \succ 0$ since then $G_{\gamma\delta t} = G_{\delta\delta t} = 0$ for $t \neq s$. It follows from the above discussion that δ^∞ is ∞-optimal if and only if $G_{\gamma\delta} \preceq 0$ for all γ.

Existence and Characterization of Stationary Strong Maximum-Present-Value Policies

Using these facts, it is now possible to prove the main result of this section.

Theorem 22. Existence and Characterization of Stationary Strong Maximum-Present-Value Policies. *In a bounded system, there is a stationary strong maximum-present-value policy. If* $\delta \in \Delta$*, the following are equivalent*: 1° δ^∞ *has strong maximum present value*; 2° δ^∞ *is* ∞*-optimal*; *and* 3° $G_{\gamma\delta} \preceq 0$ *for all* γ.

Proof. The equivalence of 2° and 3° (resp., 1°) was shown above (resp., following (16)). Thus, since 3° implies 1°, it suffices to show that there is a δ for which 3° holds. To that end, let $\delta_0 \in \Delta$ be arbitrary and choose $\delta_1, \delta_2, \ldots$ in Δ inductively as follows. Given δ_N, let δ_{N+1} strongly improve δ_N, if such a decision exists, and terminate with δ_N otherwise. Since V_{δ_N} increases lexicographically with N, no decision can appear twice in the sequence. Thus because there are only finitely many decisions, the sequence must terminate with a $\delta = \delta_N$ that no decision strongly improves. Then 3° holds from the discussion preceding the Theorem. ∎

Proof of System-Degree Theorem 9 where $d = 1$. Theorem 22 provides the tool necessary to prove Theorem 9 where $d = 1$, i.e., the sequence $\max_{\pi\in\Delta^\infty} \|P^N_\pi\|$ has degree one. To that end,

put $r_\delta = 1$ for all $\delta \in \Delta$. Then by Theorem 22, there is a δ such that $G_{\gamma\delta} \preceq 0$ for all γ, so that $g_{\gamma\delta}^{-1} \leq 0$ for all γ and $V_\delta^\rho \geq 0$ for all small $\rho > 0$, so $V_\delta \succeq 0$. On setting $v = v_\delta^{-1}$, it follows that $v = P_\delta^* r_\delta \geq 0$. Let J, K be a partition of the states for which $v_J \gg 0$ and $v_K = 0$. Now rewrite the inequality $g_{\gamma\delta}^{-1} \leq 0$ as $v \geq P_\gamma v$ for all $\gamma \in \Delta$. Iterating this inequality shows that $v \geq P_\pi^N v$ for all π, so $v_J \geq P_{\pi JJ}^N v_J$ and $0 = P_{\pi KJ}^N v_J$ for all π and $N \geq 1$. Because $v_J \gg 0$, the first inequality implies that $\max_{\pi\in\Delta^\infty} \|P_{\pi JJ}^N\| = O(1)$ and the second implies $P_{\pi KJ}^N = 0$ for all π and N, i.e., individuals in K cannot generate individuals in J. Thus, since $v_K = 0$ and $V_\delta \succeq 0$, $u \equiv v_{\delta K}^0 \geq 0$. Also $g_{\gamma\delta K}^{-1} = 0$ and so $g_{\gamma\delta K}^0 \leq 0$ for all γ, i.e., $u \geq 1 + P_{\gamma KK} u$ for all γ. Hence the restriction of the problem to states in K is a transient system, so by Theorem 9 for that case, $\max_{\pi\in\Delta^\infty} \|P_{\pi KK}^N\| = O(\alpha^N)$ for some $0 \leq \alpha < 1$. Now on setting $\pi = (\gamma_1, \gamma_2, \dots)$ and $\pi_i = (\gamma_{i+1}, \gamma_{i+2}, \dots)$, it follows that

$$P_{\pi JK}^N = \sum_{i=1}^N P_{\pi JJ}^{i-1} P_{\gamma_i JK} P_{\pi_i KK}^{N-i}.$$

To see this, consider the jk^{th} element of the i^{th} term of the summand on the right. That element is the expected number of ancestors in state $k \in K$ in period $N{+}1$ of an individual in state $j \in J$ in period one for which some descendant of j in J in period i generates an ancestor of k in K in period $i{+}1$. Thus, since $\max_{\pi\in\Delta^\infty} \|P_{\pi JJ}^N\| = O(1)$ and $\max_{\pi\in\Delta^\infty} \|P_{\pi KK}^N\| = O(\alpha^N)$, it follows that $\max_{\pi\in\Delta^\infty} \|P_{\pi JK}^N\| = O(1)$ and so $\max_{\pi\in\Delta^\infty} \|P_\pi^N\| = O(1)$. ∎

Call the algorithm given in the proof of Theorem 22 for finding a stationary strong maximum-present-value policy the *strong policy-improvement method*. The method requires only finitely many iterations, each involving two steps. For a given δ, first compute V_δ. Then seek a γ with $G_{\gamma\delta} \succ 0$. The last two steps do not appear to be executable in finite time since they seem to require computing the respective infinite matrices V_δ and $G_{\gamma\delta}$. However, it is possible to refine the algorithm to run in finite time as the development below shows.

Truncation of Infinite Matrices

In a bounded system, it turns out to be sufficient to compute the finite truncated matrices $V_\delta^n \equiv (v_\delta^{-d} \cdots v_\delta^n)$ and $G_{\gamma\delta}^n \equiv (g_{\gamma\delta}^{-d} \cdots g_{\gamma\delta}^n)$ for $n = S - T$ where T is the rank of P_δ^*. As examples, observe that if P_δ is transient, then $T = 0$; if P_δ is stochastic and irreducible, then $T = 1$; if $P_\delta = I$, then $T = S$. In order to establish the claim, a preliminary result is necessary.

Lemma 23. *If B is a real $S \times S$ matrix with rank l, L is a subspace of $\Re^S$, $x \in \Re^S$ and $B^i x \in L$ for $1 \leq i \leq l$, then $B^i x \in L$ for $i = 1, 2, \dots$.*

Proof. The result is trivial if $l = 0$ since then $B = 0$. Thus suppose that $l \geq 1$. The first step is to show that there is a positive integer $j \leq l$ such that $B^{j+1} x$ is a linear combination of

$B^1x, \ldots, B^jx$. Since the dimension of the subspace spanned by the columns of B is l, the vectors $B^1x, \ldots, B^{l+1}x$ are linearly dependent, from which the assertion follows.

Next show by induction on k that B^kx is a linear combination of $B^1x, \ldots, B^jx$ for all $k \geq 1$, which, because L is a subspace, will complete the proof. This is so by construction for $1 \leq k \leq j{+}1$. Suppose it holds for $k{-}1$ ($\geq j{+}1$), so $B^{k-1}x = \sum_{i=1}^{j} \lambda_i B^i x$. Premultiplying this equation by B gives $B^kx = \sum_{i=1}^{j} \lambda_i B^{i+1}x$. Since $B^{j+1}x$ is a linear combination of $B^1x, \ldots, B^jx$, so is B^kx. ∎

Theorem 24. Truncation. *Suppose the system is bounded, $\gamma, \delta \in \Delta$ and P_δ^* has rank T. Then*
1° $G_{\gamma\delta} = 0$ if and only if $G_{\gamma\delta}^{S-T} = 0$.
2° $V_\gamma = V_\delta$ if and only if $V_\gamma^{S-T} = V_\delta^{S-T}$.

Proof. For part 1°, it suffices to show the "if" part. Observe that since $G_{\delta\delta}^n = 0$,

$$g_{\gamma\delta}^n = g_{\gamma\delta}^n - g_{\delta\delta}^n = (P_\gamma - P_\delta)v_\delta^n, \; n = 1, 2, \ldots. \tag{20}$$

Because $G_{\gamma\delta}^{S-T} = 0$, it follows from (20) that

$$(P_\gamma - P_\delta)v_\delta^n = 0, \text{ for } 1 \leq n \leq S{-}T. \tag{21}$$

In view of (20), it suffices to show that (21) holds for $n > S{-}T$. That this is so follows from Lemma 23 with $B = -D_\delta$, L the null space of $P_\gamma - P_\delta$, and $x = v_\delta^0$, on noting from Theorem 20 that because the rank of P_δ^* is T, the rank of D_δ is $S{-}T$.

For part 2°, it suffices to show the "if" part. Now $V_\gamma^{S-T} = V_\delta^{S-T}$ implies that

$$(D_\gamma - D_\delta)v_\delta^n = 0, \text{ for } 0 \leq n \leq S{-}T{-}1. \tag{22}$$

To show that $V_\gamma = V_\delta$, it suffices to establish (22) for $n \geq S - T$. The last follows from Lemma 23 with $B = -D_\delta$, L the null space of $D_\gamma - D_\delta$, and $x = -r_\delta$, on noting, as above, that the rank of D_δ is $S{-}T$. ∎

Remark 1. If the rank T of P_δ^* is S, then $P_\delta^* = I$ and $D_\delta = 0$ by Theorem 20. Then $v_\delta^{-1} = r_\delta$ and $v_\delta^n = 0$ for $n = 0, 1, \ldots$.

Remark 2. Part 1° of Theorem 24 implies that if the s^{th}, say, row of $G_{\gamma\delta}^{S-T}$ vanishes, the same is so of the s^{th} row of $G_{\gamma\delta}$. As a consequence, if γ strongly improves δ, then $G_{\gamma\delta}^{S-T} \succ 0$. Thus, in searching for a decision that strongly improves δ, it is never necessary to look beyond the matrices V_δ^S and $G_{\gamma\delta}^S$, or if T is known, V_δ^{S-T} and $G_{\gamma\delta}^{S-T}$, for $\gamma \in \Delta$. Observe also that $T = 0$ if and only if P_δ is transient. Otherwise $T \geq 1$.

n-Optimality: Efficient Implementation of the Strong Policy-Improvement Method

A more efficient way to execute the strong policy-improvement method for bounded S-state systems is to lexicographically maximize V_δ by first maximizing V_δ^{-d}, then V_δ^{-d+1}, then V_δ^{-d+2}, etc. Indeed, as the sequel shows, for each $n \geq d$, it is often of interest to seek a policy δ^∞ that is *n-optimal*, i.e., $V_\delta^n \succeq V_\gamma^n$, all $\gamma \in \Delta$. Let Δ_n be the set of decisions δ for which δ^∞ is n-optimal.

Theorem 25. Selectivity. *In a bounded S-state system with system degree d, $\Delta \supseteq \Delta_{-d} \supseteq \Delta_{-d+1} \supseteq \cdots \supseteq \Delta_{S-T} = \Delta_{S-T+1} = \cdots = \Delta_\infty$ where T is the minimum of the ranks of the stationary matrices over all decisions.*

Proof. Immediate from the definitions and the Truncation Theorem 24.

n-Improvements. It is useful now to develop a method of finding an n-optimal policy. To that end, put $G_{\gamma\delta}^n \equiv (g_{\gamma\delta}^{-d} \cdots g_{\gamma\delta}^n)$. Call $\gamma \in \Delta$ an *n-improvement* of $\delta \in \Delta$ if $G_{\gamma\delta}^n \succ 0$ and if $\gamma^s = \delta^s$ whenever the s^{th} row of $G_{\gamma\delta}^n$ vanishes. Clearly γ strongly improves δ.

Theorem 26. n-improvements and n-optimal policies. *In a bounded system with system degree d, if γ is an n-improvement of δ, then $V_\gamma^n \succ V_\delta^n$. If δ has no $(n+d)$-improvement, then δ^∞ is n-optimal.*

Proof. If γ is an n-improvement of δ, then γ is a strong improvement of δ because $G_{\gamma\delta s}^n = 0$ implies $G_{\gamma\delta s} = G_{\delta\delta s} = 0$. Thus $G_{\gamma\delta}^\rho > 0$ for all small enough $\rho > 0$. Moreover, since $R_\gamma^\rho \geq \beta I$,

$$\lim_{\rho\downarrow 0} \sum_{k=-d}^{n} \rho^{k-n}(v_\gamma^k - v_\delta^k) = \lim_{\rho\downarrow 0} \rho^{-n}(V_\gamma^\rho - V_\delta^\rho) = \lim_{\rho\downarrow 0} \rho^{-n} R_\gamma^\rho G_{\gamma\delta}^\rho \geq \lim_{\rho\downarrow 0} \beta\rho^{-n} G_{\gamma\delta}^\rho = \lim_{\rho\downarrow 0} \sum_{k=-d}^{n} \rho^{k-n} g_{\gamma\delta}^k > 0,$$

so $V_\gamma^n \succ V_\delta^n$ as claimed.

Suppose δ has no $(n+d)$-improvement. Then for each γ, $\rho^{-n-d} G_{\gamma\delta}^\rho \leq o(1)$. Also $\lim_{\rho\downarrow 0} \rho^d R_\gamma^\rho$ exists and is finite by Theorem 20 and R_γ^ρ is nonnegative for all small enough $\rho > 0$. Thus

$$\lim_{\rho\downarrow 0} \sum_{k=-d}^{n} \rho^{k-n}(v_\gamma^k - v_\delta^k) = \lim_{\rho\downarrow 0} \rho^{-n}(V_\gamma^\rho - V_\delta^\rho) = \lim_{\rho\downarrow 0} \rho^d R_\gamma^\rho \rho^{-n-d} G_{\gamma\delta}^\rho \leq 0,$$

whence $V_\gamma^n \preceq V_\delta^n$ for each γ, i.e., δ is n-optimal. ∎

Example. $d = 1$, δ Has No n-Improvement, and $\delta \notin \Delta_n$. It is natural to hope that if $d = 1$ and δ has no n-improvement, then $\delta \in \Delta_n$. However, that is not so. For example, suppose that there is a single state and two decisions γ, δ with $P_\gamma = P_\delta = 1$, so $Q_\gamma = Q_\delta = 0$, and $r_\gamma = 1$, $r_\delta = 0$. Then $v_\gamma^{-1} = 1$ and $v_\delta^{-1} = 0$, so $g_{\gamma\delta}^{-1} = Q_\gamma v_\delta^{-1} = 0$, whence δ has no -1-improvement. But γ is a 0-improvement of δ because $g_{\gamma\delta}^0 = r_\gamma + Q_\gamma v_\delta^0 - v_\delta^{-1} = 1 > 0$. Also, γ is -1-optimal.

Example. $d = 1$, δ Has a -1-Improvement. Suppose there are two states and two decisions γ and δ with $\gamma^2 = \delta^2$. Let $r_\gamma = r_\delta = (0 \quad 1)^T$. Let $P_\delta = I$, so $Q_\delta = 0$. Let $P_{\gamma 1} = (0 \quad 1)$, so $Q_{\gamma 1} = (-1 \quad 1)$. Then $v_{\delta 1}^{-1} = (0 \quad 1)^T$, $G_{\gamma\delta 1}^{-1} = Q_{\gamma 1} v_{\delta 1}^{-1} = 1$ and $G_{\gamma\delta 2}^{-1} = G_{\delta\delta 2}^{-1} = 0$, so γ -1-improves δ.

Computing V_δ^n. The only method discussed so far for computing V_δ^n requires first calculating the stationary and deviation matrices for δ. It turns out that this is not necessary. Indeed, it suffices to solve a system of linear equations for V_δ^n.

Theorem 27. Computation of V_δ^n. *Suppose the system is bounded with system degree d. Let $V^n = (v^{-d} \cdots v^n)$ where the v^j are S-vectors and $v^{-d-1} \equiv 0$. Then $V^{n+d} = V_\delta^{n+d}$ satisfies the linear equations*

$$r_\delta^j + Q_\delta v^j = v^{j-1} \text{ for } -d \le j \le n+d. \tag{23}$$

Conversely, if V^{n+d} satisfies these equations, then $V^n = V_\delta^n$.

Proof. Observe first that since $G_{\delta\delta} = 0$, then $G_{\delta\delta}^{n+d} = 0$, or equivalently, $V^{n+d} = V_\delta^{n+d}$ satisfies (23). Conversely, suppose that V^{n+d} satisfies (23). Since V_δ^{n+d} also satisfies (23), it follows that $U^{n+d} \equiv V^{n+d} - V_\delta^{n+d}$ satisfies

$$Q_\delta u^j = u^{j-1} \text{ for } -d \le j \le n+d \tag{24}$$

where $U^{n+d} = (u^{-d}, \ldots, u^{n+d})$ and $u^{-d-1} \equiv 0$. Now (24) implies that

$$Q_\delta u^j = u^{j-1} \text{ for } -d \le j \le n \tag{25}$$

and, on premultiplying (24) by P_δ^* and using the facts that $P_\delta^* Q_\delta = 0$ and, if $d = 0$, $P_\delta^* = 0$,

$$P_\delta^* u^j = 0 \text{ for } -d \le j \le n. \tag{26}$$

Subtracting (26) from (25) gives

$$(Q_\delta - P_\delta^*) u^j = u^{j-1} \text{ for } -d \le j \le n. \tag{27}$$

Since $P_\delta^* - Q_\delta$ is nonsingular and $u^{-d-1} = 0$, it follows from (27) that $u^{-d} = 0$. Proceeding inductively, it follows that $U^n = 0$, i.e., $V^n = V_\delta^n$ as claimed. ∎

Example. Why V^{n+d} Must Satisfy (23) to Assure $V^n = V_\delta^n$ with $d = 1$. Suppose $P_\delta = I$, so $Q_\delta = 0$, and $n = -2$,. Then the first equation in (23) is $0v^{-1} = 0$, and so does not constrain v^{-1} at all. However, the second equation in (23) is $r_\delta + 0v^0 = v^{-1}$, which shows that $v^{-1} = r_\delta$.

Policy Improvement in Bounded Systems. The above results lead to the following policy-improvement method for finding an n-optimal policy for a bounded system with $d = 1$ that requires only $(n+1)$-improvements of decisions. Let δ_0 be given. Choose $\delta_1, \ldots, \delta_N$ inductively so that δ_k is an $(n+1)$-improvement of δ_{k-1} for given $k = 1, \ldots, N$ where N is the first integer (which exists) for which δ_N has no $(n+1)$-improvement. Then δ_N is n-optimal. In order to check whether a decision δ has an $(n+1)$-improvement, it is necessary to compute V^{n+2} satisfying (23) with $n+1$ replacing n there. Then from Theorem 27, $V^{n+1} = V_\delta^{n+1}$, which is needed to begin the search for an $(n+1)$-improvement of δ.

Sequential Implementation. The above *generic implementation* of policy improvement to find an element of Δ_n can be made more efficient by *sequential implementation*, i.e., by first seeking a $\delta_{-1} \in \Delta_{-1}$, then a $\delta_0 \in \Delta_0$, then a $\delta_1 \in \Delta_1$, etc., until a $\delta_n \in \Delta_n$ is found. More precisely, suppose one has found a $\delta_{k-1} \in \Delta_{k-1}$. Then repeatedly seek $(k+1)$-improvements until a decision δ_k is found with no $(k+1)$-improvement. Then by Theorem 26, $\delta_k \in \Delta_k$.

Sequential implementation is more efficient than generic implementation for two reasons. First, sequential implementation generally requires fewer iterations because it focuses on maximizing the coefficients V^n_δ of the Laurent expansion of V^ρ_δ in order of declining importance. Thus, sequential implementation first finds $\delta = \delta_{-1}$ maximizing v^{-1}_δ, then from among all such δ a $\delta = \delta_0$ maximizing v^0_δ, then from among all such δ a $\delta = \delta_1$ maximizing v^1_δ, etc.

Second, sequential implementation requires much less computation at each iteration than generic implementation. The implementation below explains why by showing how to use Theorems 26 and 27 to process inputs $\delta \in \Delta_{k-1}$ and V^k_δ to produce outputs $\gamma \in \Delta_k$ and V^{k+1}_γ.

(*a*) Find v^{k+1}_δ. Solve $r^{k+1}_\delta + Q_\delta v^{k+1} = v^k_\delta$ and $r^{k+2}_\delta + Q_\delta v^{k+2} = v^{k+1}$ for $v^{k+1} = v^{k+1}_\delta$ and v^{k+2}.

(*b*) Find a $(k+1)$-improvement γ of δ if one exists. Since $\delta \in \Delta_{k-1}$, δ has no $(k-1)$-improvement. Consider the γ for which $G^{k-1}_{\gamma\delta} = G^{k-1}_{\delta\delta} = 0$, so $G^{k+1}_{\gamma\delta} = (0\ g^k_{\gamma\delta}\ g^{k+1}_{\gamma\delta})$. Choose from among those γ one for which $(g^k_{\gamma\delta}\ g^{k+1}_{\gamma\delta}) \succ 0$ and $\gamma^s = \delta^s$ whenever $(g^k_{\gamma\delta s}\ g^{k+1}_{\gamma\delta s}) = 0$. If there is a γ that $(k+1)$-improves δ, go to (c). Otherwise set $\gamma = \delta$ and go to (d). In both cases $V^{k-1}_\gamma = V^{k-1}_\delta$ and $\gamma \in \Delta_{k-1}$.

(*c*) Find v^k_γ. Solve $r^k_\gamma + Q_\gamma v^k = v^{k-1}_\delta$ and $r^{k+1}_\gamma + Q_\gamma v^{k+1} = v^k$ for $v^k = v^k_\gamma$ and v^{k+1}. Reset $\delta = \gamma$ and go to (a).

(*d*) Terminate with $\gamma \in \Delta_k$ and V^{k+1}_γ. Since γ has no $(k+1)$-improvement, $\gamma = \delta \in \Delta_k$. Solve $r^{k+1}_\gamma + Q_\gamma v^{k+1} = v^k_\delta$ and $r^{k+2}_\gamma + Q_\gamma v^{k+2} = v^{k+1}$ for $v^{k+1} = v^{k+1}_\gamma$ and v^{k+2}.

Note that steps (a), (c) and (d) each require solving $2S$ linear equations. And finding the *best* $(k+1)$-improvement γ of δ in step (b) entails lexicographically maximizing $(g^k_{\gamma\delta}\ g^{k+1}_{\gamma\delta})$, which requires up to $2A$ operations.[7]

Uniqueness. In practice, it is usually so that for relatively small k and δ^∞ k-optimal, only $\gamma = \delta$ satisfies $G^k_{\gamma\delta} = G^k_{\delta\delta} = 0$. Then, by the Selectivity Theorem 25, δ^∞ is n-optimal for all $n \geq k$. Hence one can terminate sequential implementation early at iteration k.

Policy Improvement in Transient Systems. It is particularly easy to describe the amount of work required to find a decision in Δ_n in transient systems, or equivalently, $d = 0$. In that event, let $V^n = (v^0\ v^1\ \cdots\ v^n) = V^n_\delta$ for any $\delta \in \Delta_n$. Now from Theorem 26, $\delta \in \Delta_n$ if and only if $G^n_{\gamma\delta} \preceq 0$ for all $\gamma \in \Delta$, or equivalently from Theorem 27 ($v^{-1} = 0$, $\Delta_{-1} = \Delta$),

$$v^k = \max_{\gamma \in \Delta_{k-1}} [r^k_\gamma - v^{k-1} + P_\gamma v^k],\ k = 0, \ldots, n. \tag{28}$$

[7]By contrast with sequential implementation, the generic implementation entails solving a system of $(k+4)S$ linear equations at each iteration. And seeking the best $(k+1)$-improvement at each iteration requires up to $(k+3)A$ operations to do each lexicographic maximization.

Moreover, the set of decisions γ that maximize the right-hand side of the k^{th} equation in (28) is $\Delta_k \subseteq \Delta_{k-1}$. Notice that these equations can be solved for v^0, $v^1, \ldots, v^n$ in that order. Once v^{k-1} has been found, the work of finding v^k is precisely that of finding a stationary maximum-value policy for a transient system with a restricted decision set Δ_{k-1} and a new reward vector $r^k_\gamma - v^{k-1}$. Thus the total amount of work required to find a $\delta \in \Delta_n$ is at most $n+1$ times that required to find a stationary maximum-value policy in a transient system—and usually much less. In particular, the work required to find a strong maximum-present-value policy is at most $S+1$ times that required to find a stationary maximum-value policy in a transient system. Although we do not discuss it in detail here, we remark that the situation described above is similar for bounded systems. In short, one should expect to be able to find strong maximum-present-value policies on PCs for transient or bounded systems with thousands of states.

9 CESÀRO OVERTAKING OPTIMALITY WITH IMMIGRATION IN BOUNDED SYSTEMS

[Ve66], [Ve68], [DM68], [Li69], [Ve74], [RV92]

The preceding section generalizes the maximum-value criterion from transient to bounded systems by means of maximum-present-value preference relations for small interest rates. Another approach to bounded-systems is instead to consider preference relations that make a suitable function of the rewards large over long horizons. For example, it is often of interest to consider preference relations that focus on making the long-run expected reward rate, total reward, total float, future value with a negative interest rate, instantaneous rate of return, or expected multiplicative utility large. It is possible to encompass these and other preference relations in a uniform and simple way by generalizing our model to allow exogenous (possibly negative) immigration of individuals into the population in each period. Then, each of the above preference relations reduces to Cesàro overtaking optimality for a suitable choice of the immigration stream.

Call a policy Cesàro overtaking optimal if the difference between the expected N-period rewards earned by that policy and any other policy has a nonnegative Cesàro limit inferior of appropriate order. Cesàro limits are needed to smooth out oscillations in the differences of expected N-period rewards that would otherwise generally prevent the existence of overtaking-optimal policies from being assured. It might be supposed that a policy that is Cesàro overtaking optimal for each individual immigrant would also be Cesàro overtaking optimal for the immigration stream itself. But that is not so! The reason is fundamentally that each individual immigrant would be indifferent between two policies that produce the same permanent income even though one policy produces more float, i.e., temporary income, than the other. But the system would prefer the policy that produces more float if, as is possible, the aggregate of the float that each individual immigrant produces provides extra permanent income for the system. That is the reason why the set of Cesàro-overtaking-optimal policies depends on the immigration stream and permits several preference relations to be reduced thereto.

Allowing immigration into the population is also interesting in its own right in the study of controlled populations. For example, exogenous arrivals into a queueing network comprise an immigration stream therefor.

Examples

Controlled Queueing Network with Proportional Service Rates. Consider a controlled S-station queueing network. Customers arrive exogenously at the various stations according to a general stochastic process that is independent of the service process, with w_s^N being the (finite) expected number of such arrivals at station s in period $N+1$. Each station has infinitely many independent and identical servers, each with geometrically-distributed services times. A customer arriving exogenously or endogenously at a station at the beginning of a period begins service there immediately because there are infinitely many servers available. In each period and at each station, the network manager chooses an action from a finite set that determines the common service rate of customers at the station and the common rates at which customers leave the station in the period to exit the network or to go to various stations in the next period. The cost of servicing a customer and the revenue earned when the customer completes service at a station depends on the action taken by the manager. This is a Markov decision chain with immigration in which the states are the S stations, $r(s,a)$ is the reward earned by a customer at station s when the manager takes action a at that station, $p(t \mid s,a)$ is the probability that a customer being served at station s completes service at that station during the period and is sent to station t at the beginning of the following period, and $1-\sum_t p(t \mid s,a)$ is the probability that a customer completes service at station s and leaves the system.

Keep in mind that the assumption that a station has infinitely many identical servers has an alternate interpretation, viz., that the station has a single server whose service rate is proportional to the number of individuals waiting to be served at the station. In this event, selection of an action by the manager amounts to choosing the proportionality constant in the proportional service rate there.

Figure 9 illustrates a four-station network. Exogenous arrivals occur only at stations 1 and 2. Stations 1 and 2 generate output at both stations 3 and 4. Stations 2 and 3 also generate output that must be reworked respectively by starting afresh at stations 2 and 1 respectively. This model of controlled queueing networks has several strengths. One is that only one state is required for each station (it is not necessary to keep track of the number of customers in a station because there are infinitely many servers), so an S-station network leads to a problem with only S states. For this reason, it is computationally tractable to find optimal policies for controlling queueing networks with as many as several thousand stations! Two other strengths are that the exogenous arrival process is arbitrary and the manager is allowed to control the service rates and routing of customers.

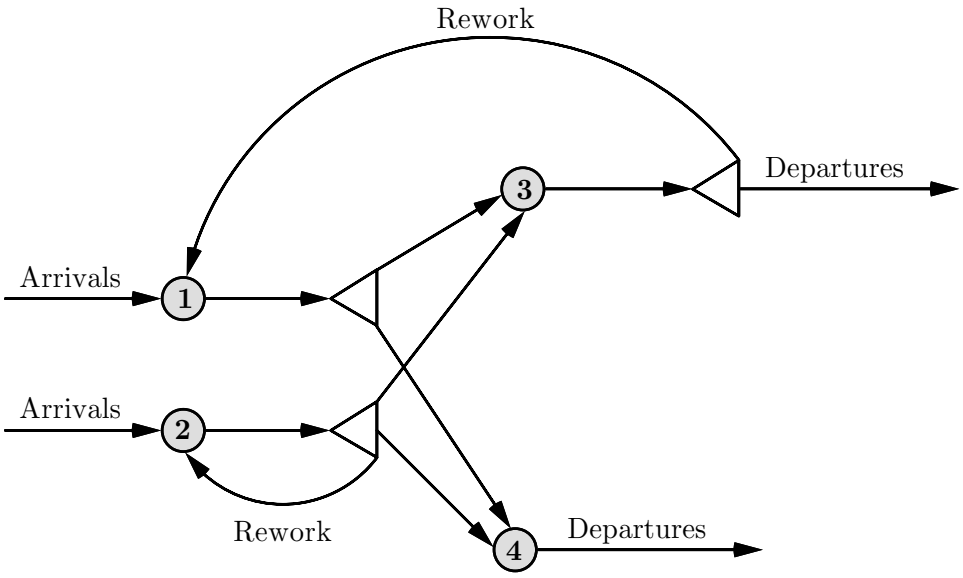


Figure 9. Optimal Control of Service/Rework Rates in a Four-Station Network of Queues

Of course the model also has limitations. The service rate at each station must be proportional to the number of customers there; the number of such customers cannot be restricted; and rewards are linear in the expected number of customers at each station.

Cash Management. Bills arrive daily at a firm for payment from all over the country. The firm pays the bills by writing checks drawn on several banks throughout the country at which it has accounts. The length of time for a check to clear depends on the location of the payee and of the bank on which the check is drawn. Generally, the more distant the payee from the bank on which the check is written, the longer the check takes to clear. From past experience, the firm knows the distribution of the random time to clear checks drawn on each bank to pay bills from each payee. The problem is to choose the bank on which to write a check to pay bills from each payee.

Manpower Planning. A personnel director manages a work force in which individuals are hired, move between various jobs and skill-levels over time so as to meet an organization's needs, and eventually leave. An individual's state may be differentiated by factors like age, compensation, current assignment, education, skill, etc. Actions may include policies on recruitment, hiring, training, compensation, job assignment, promotion, layoff, firing, retirement, etc.

Insurance Management. An insurance company insures its policy holders against risks of various types. A policy holder's state could include the type and terms of his policy, age, sex, health, safety record, occupation, prior claims, etc. The company's actions might include rules for accepting or rejecting new policy holders, premiums charged, dividends paid, coverage, cancellation provisions, etc.

Asset Management. Managers of assets like vehicles, machines, equipment, computers, buildings, etc., maintain them to provide service to internal and/or external customers. An asset's state may include its type, condition, current status, reliability, age, capabilities, market value, location, etc. Actions may include buying, renting, leasing, operating, pricing, maintaining, selling, storing, moving, etc.

Immigration Stream

As several of the above examples suggest, it is useful to generalize the systems considered heretofore to allow individuals to enter the population through exogenous uncontrolled immigration. In particular, the (possibly negative) number of immigrants in state s in period $N+1 \geq 1$ is w_s^N. Let $w^N = \text{diag}(w_s^N)$ be the $S \times S$ diagonal *immigration matrix* whose s^{th} diagonal element is w_s^N. Call immigrants in a period *positive* or *negative* according as the number of immigrants in the period is positive or negative. Positive (resp., negative) immigrants add (resp., subtract) their rewards and those of their descendants to (resp., from) that of the system. Call the sequence $w = (w^0 \;\; w^1 \;\; \cdots \;\;)$ of immigration matrices the *immigration stream*.

An alternate approach is to represent an immigration stream by appending immigrant-generating states. This device can be useful for studying certain well behaved immigration streams. Indeed this viewpoint is necessary when it is possible to control the immigration stream. However, it is usually better not to append additional states to reflect immigration for two reasons. One is that this approach facilitates the study of the effect of changes in the immigration stream on the set of Cesàro-overtaking-optimal policies. The other is that this method does not enlarge the state-space and leads to more efficient computational methods.

Cohort and Markov Policies

Each immigrant in a period generates populations in subsequent periods. Call the collection of all immigrants in a given period and their descendants a *cohort*. The N^{th} cohort is the one that the immigrants in period N generate. Of course each individual in the population in a given period belongs to exactly one cohort, so there is a partition of the population in any given period into cohorts as Figure 10 illustrates.

In the presence of immigration, a policy $\pi = (\delta_N)$ has at least two different interpretations. One is as a *Markov policy*, i.e., every individual in period N uses δ_N. The other is as a *cohort policy*, i.e., all members of cohort i use δ_N in period $N+i-1$. Equivalently, a cohort policy starts afresh for each cohort. Figure 11 illustrates these two types of policies. Of course, the distinction between Markov and cohort policies vanishes when they are stationary.

Nonstationary Markov policies might seem more natural than nonstationary cohort policies because the action an individual takes in a given state intuitively should depend only on the state

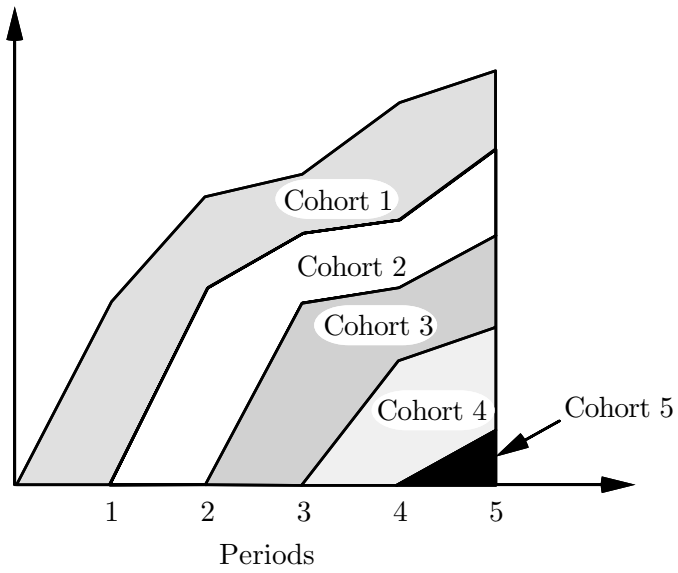


Figure 10. Cohorts

Markov Policy

	Period			
Cohort	1	2	3	$\cdots$
1	δ_1	δ_2	δ_3	
2		δ_2	δ_3	$\cdots$
3			δ_3	
$\vdots$		$\vdots$		

Cohort Policy

	Period			
Cohort	1	2	3	$\cdots$
1	δ_1	δ_2	δ_3	
2		δ_1	δ_2	$\cdots$
3			δ_1	
$\vdots$		$\vdots$		

Figure 11. Interpretations of (δ_N) as Markov and Cohort Policies

and not the cohort to which the individual belongs. However, perhaps surprisingly, it turns out that cohort policies are more convenient for two reasons. One is that they are easier to work with because the formulas for the relevant quantities are nicer, and results about Markov policies can easily be derived from them in any case. The second reason is that cumulative immigration streams have an alternate interpretation in terms of unit values of rewards earned in different periods. In that event, when there is no exogenous immigration, all individuals are in the same cohort so cohort policies are the most general that need to be considered. For these reasons, *in the remainder of this section policies mean cohort policies unless explicitly stated to the contrary.*

Let V_π^{Nw} be the N-period value when w is the immigration stream and π is a (cohort) policy. Observe that since $N-k$ periods remain when the immigrants in cohort $k+1$ arrive, then, as Figure 12 illustrates,

$$V_\pi^{Nw} = \sum_{k=0}^{N} w^k V_\pi^{N-k}.$$

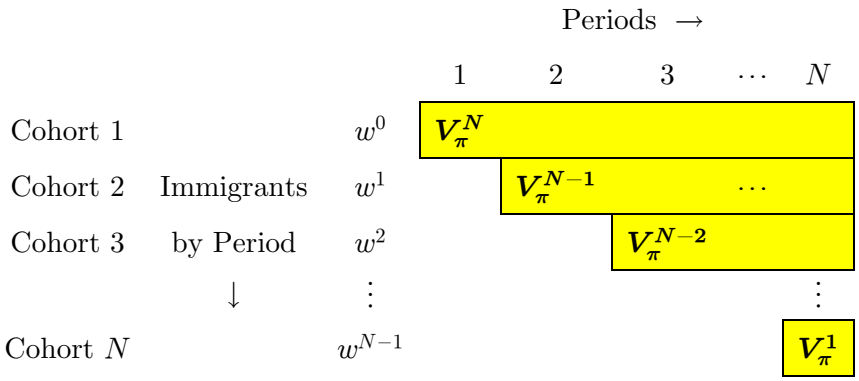


Figure 12. N-Period Value with Immigration Stream w

Overtaking Optimality

It might be expected that the analog of finding a strong maximum-present-value policy, i.e., one that simultaneously maximizes the present value for all small enough positive interest rates, would be to find a policy λ that maximizes the N-period value V_λ^{Nw} simultaneously for all sufficiently large N. Unfortunately, as the example below illustrates, such policies do not generally exist because the set of maximum-N-period-value policies usually depends on N for large N, even for transient systems. Of course, there are some exceptions to this general rule, e.g., circuitless systems with no immigration after the first period, minimum-cost-chain problems, etc.

But a satisfactory treatment of bounded systems requires a weaker concept. To motivate one such concept, observe that in a transient system, $\lim_{N\to\infty} V_\pi^N = V_\pi$ for each π. Thus a policy λ has maximum value, i.e., $V_\lambda - V_\pi \geq 0$ for all π, if and only if $\liminf_{N\to\infty}(V_\lambda^N - V_\pi^N) \geq 0$ for all π. Both definitions are well defined in a transient system, but only the latter is generally well defined in a bounded system. This suggests the following definition in a bounded system.

Call a policy λ *overtaking optimal* for w if

$$\liminf_{N\to\infty}(V_\lambda^{Nw} - V_\pi^{Nw}) \geq 0 \text{ for all } \pi. \tag{1}$$

Call λ *overtaking optimal* if $V_\lambda^N - V_\pi^N$ replaces $V_\lambda^{Nw} - V_\pi^{Nw}$ in (1). Observe that instead of maximizing a criterion function, overtaking optimality is a binary *preference relation* on the set of policies. Moreover, that relation is a *quasiorder*, i.e., is reflexive and transitive, but not necessarily antisymmetric.

Stationary overtaking-optimal policies exist for transient systems and, when the *stopping value* is finite, for *stopping problems*. Indeed, they exist provided that the limit inferior in (1) can be replaced by a finite limit, as occurs in *convergent systems*, and often in *positive* (all rewards are nonnegative) and *negative* (all rewards are nonpositive) *systems*. Unfortunately, overtaking-optimal policies do not exist in general because $V_{\delta s}^N - V_{\gamma s}^N$ can oscillate about zero for some states s as the next example illustrates.

Example. Nonexistence of Overtaking-Optimal Policies when $d = 1$ and Dependence of Maximum-N-Period-Value Policies on N. Consider the three-state deterministic system in Figure 13. There are two decisions γ and δ, and the rewards are as indicated on each arc. Note that γ^∞ and δ^∞ are effectively the only policies and $d = 1$. Now $V^N_{\delta 0} - V^N_{\gamma 0} = 2(-1)^{N+1} + 1$ alternates between 3 and -1 so δ^∞ has maximum N-period value for odd N and γ^∞ has maximum N-period value for even N. Thus, no policy has maximum N-period value for all large enough N. Also,

$$\liminf_{N\to\infty}(V^N_{\delta 0} - V^N_{\gamma 0}) = -1 \text{ and } \liminf_{N\to\infty}(V^N_{\gamma 0} - V^N_{\delta 0}) = -3,$$

so no policy is overtaking optimal.

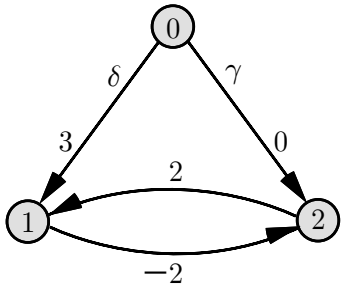


Figure 13. Nonexistence of Overtaking-Optimal Policies

Cesàro Overtaking Optimality

One way to address this problem is to smooth out the fluctuations in $V^{Nw}_\lambda - V^{Nw}_\pi$ by substituting a (C, d) limit inferior for the limit inferior in (1). If $b_0, b_1, \ldots$ is a sequence, the (C, d) limit inferior of the sequence is the ordinary limit inferior of the sequence if $d = 0$ and the limit inferior of the sequence of arithmetic averages, viz., $\liminf_{N\to\infty} N^{-1}\sum_{i=0}^{N-1} b_i$, if $d = 1$.

Call λ *Cesàro overtaking optimal* for w if

$$(1)' \qquad \liminf_{N\to\infty}(V^{Nw}_\lambda - V^{Nw}_\pi) \geq 0 \ (\mathrm{C}, d) \text{ for all } \pi.$$

Call λ *Cesàro overtaking optimal* if $V^N_\lambda - V^N_\pi$ replaces $V^{Nw}_\lambda - V^{Nw}_\pi$ in $(1)'$. Observe that in the example above, δ^∞ is the unique Cesàro-overtaking-optimal policy, though it is not overtaking optimal. It turns out that stationary Cesàro-overtaking-optimal policies always exist for bounded systems under natural assumptions on the immigration stream. In order to show this, it is necessary to develop some preliminary concepts.

Convolutions

It is convenient to represent (V^{Nw}_δ) as a "convolution" of the two sequences (V^N_δ) and (w^N). To define this notion, suppose $B = (B^N)$ and $C = (C^N)$ are sequences for which $N \geq 0$ and one or both are either scalars or matrices of common size and for which the products $B^i C^j$ are well defined for all i, j. The last is always so if one or both of the sequences are scalars or if both are

matrices and the number of columns of each B^i equals the number of rows of each C^j. Denote by $B*C$ the sequence whose N^{th} element $(B*C)^N$ is $\sum_0^N B^i C^{N-i}$; call $B*C$ the *convolution* of B and C. The convolution operation $*$ is *associative*, i.e., $(B*C)*E = B*(C*E)$, provided that $B*C$ and $C*E$ are well defined. If also, $B^iC^j = C^jB^i$ for all i,j, then the convolution operation is also *commutative*, i.e., $B*C = C*B$. In particular, that is so if B or C is a sequences of scalars. If B and C are each finite sequences with $e \geq 1$ elements, denote by $B \bullet C$ the middle element of the convolution $B*C$ of $2e-1$ elements. Thus, $B \bullet C$ is the sum over $k = 1, \ldots, e$ of the product of the element of B with k^{th} smallest index and that of C with k^{th} largest index.

Binomial Coefficients and Sequences

For any positive integer j, set $j! \equiv (j)(j-1)\cdots(1)$, and for any number k, define k *choose* j, denoted $\binom{k}{j}$, by

$$\binom{k}{j} \equiv \frac{k(k-1)\cdots(k-j+1)}{j!}.$$

For $j = 0$, put $0! \equiv 1$ and $\binom{k}{0} \equiv 1$. Of course, if also $k \geq j$ are nonnegative integers, then

$$\binom{k}{j} = \frac{k!}{j!(k-j)!} = \binom{k}{k-j}.$$

It is easy to verify for any k and nonnegative integer j that $\left(\binom{k}{-1} \equiv 0\right)$

$$\binom{k}{j-1} + \binom{k}{j} = \binom{k+1}{j}. \tag{2}$$

It turns out that the immigration streams in which the numbers of immigrants in a state in a period are independent of the state and in which the common sequence is "binomial" play a fundamental role in the analysis. Define the *binomial sequence of order* n, denoted 1_n, by induction on n. Call $1_0 \equiv (1 \quad 0 \quad 0 \quad \cdots)$ the *identity sequence* because $1_0 * B = B$ for any sequence $B = (B^0 \quad B^1 \cdots)$. Call $1_1 \equiv (1 \quad 1 \quad 1 \quad \cdots)$ the *summing sequence* because the N^{th} element of $1_1 * B$ is $\sum_{i=0}^N B^i$. The inverse of summing is differencing. Call $1_{-1} \equiv (1 \quad -1 \quad 0 \quad 0 \quad \cdots)$ the *differencing sequence* because the N^{th} element of $1_{-1} * B$ is $B^N - B^{N-1}$ where $B^{-1} \equiv 0$.

Now define $1_n \equiv (1_n^N)$ inductively by the rule:

$$1_n = \begin{cases} 1_1 * 1_{n-1}, \, n > 1 \\ 1_{-1} * 1_{n+1}, \, n < -1. \end{cases}$$

Call $1_n * B$ and $1_{-n} * B$ respectively the *n-fold sum* and *n-fold difference* of B. Observe that

$$1_m * 1_n = 1_{m+n}, \, m, n = 0, \pm 1, \pm 2, \ldots.$$

It is easy to show from (2) by induction on n that

$$1_n^N = \binom{N+n-1}{N},\ N, |n| = 0, 1, 2, \ldots.$$

Thus,

$$1_n^N = \binom{N+n-1}{n-1},\ n > 0,$$

which is a polynomial in N of degree $n-1$, whence 1_n has degree n. Also

$$1_n^N = (-1)^N \binom{-n}{N},\ n \leq 0,$$

which alternates in sign for $0 \leq N \leq -n$ and is zero for $N > -n$.

The table below gives the first five elements of the binomial sequences 1_n for $|n| \leq 2$.

n	1_n					
-2	1	-2	1	0	0	$\cdots$
-1	1	-1	0	0	0	
0	1	0	0	0	0	$\cdots$
1	1	1	1	1	1	
2	1	2	3	4	5	$\cdots$

Figure 14. Some Binomial Sequences

Polynomial Expansion of Expected Population Sizes with Binomial Immigration

Our goal now is to give a polynomial expansion in N of the N-period value V_π^{Nn} of a policy π for a *binomial immigration stream* 1_n *of order* n, i.e., $1_n^{N-1}I$ is the immigration matrix in each period $N \geq 1$. Note that the symbol 1_n refers both to the binomial sequence and immigration stream of order n. The context will make the meaning clear, though either interpretation will usually do.

For each policy π, let $\mathbb{V}_\pi \equiv (V_\pi^0 \ \ V_\pi^1 \ \ \cdots)$ and $\mathbb{V}_\pi^n \equiv (V_\pi^{0n} \ \ V_\pi^{1n} \ \ \cdots) \equiv 1_n * \mathbb{V}_\pi$. If $\pi = \delta^\infty$, replace π by δ in these definitions and let $\mathbb{P}_\delta \equiv (0 \ \ P_\delta^0 \ \ P_\delta^1 \ \ \cdots)$. It follows that $\mathbb{V}_\delta = (1_1 * \mathbb{P}_\delta) r_\delta$ and $\mathbb{V}_\delta^n = (1_{n+1} * \mathbb{P}_\delta) r_\delta$.

There are two interpretations of the st^{th} element of $(1_{n+1} * \mathbb{P}_\delta)^N = (1_n * (1_1 * \mathbb{P}_\delta))^N$. One is the expected population size in state t in period $N \geq 1$ generated by binomial immigration of order $n+1$ into state s. The other is the sum of the expected sojourn times in state t in the first $N \geq 1$ periods of the immigrants into state s and their descendants when there is binomial immigration of order n into state s.

Thus, in order to obtain a polynomial expansion of V_δ^{Nn} in N, it is necessary to find a polynomial expansion of the N^{th} element of $1_{n+1} * \mathbb{P}_\delta$ in N. The polynomial expansion is analogous to the Laurent expansion of the resolvent of Q_δ that was needed to study the problem with small interest rates. The desired polynomial expansion is provided in the next theorem.

Theorem 28. Polynomial Expansion of Convolutions of Binomial Sequences with Matrix Powers. *Suppose that P is a square complex matrix with $d_P \leq 1$ and with stationary and deviation matrices P^* and D respectively. Then for $N \geq 1$ and $n \geq -1$,*

$$(3) \qquad (1_{n+1} * \mathbb{P})^N = \binom{N+n}{n+1} P^* + \sum_{j=0}^{n} \binom{N+n}{n-j} (-1)^j D^{j+1} + P^{N+n}(-D)^{n+1}(I - P^*)$$

where $\mathbb{P} = (0 \quad P^0 \quad P^1 \quad \cdots)$.

Before proving this Theorem, it is useful to discuss it. Notice that on letting $\mathbb{D}_{n+1}$ be the sequence of matrices $(P^* \quad D^1 \quad -D^2 \cdots (-1)^n D^{n+1})$ and b_n^N be the sequence of binomial coefficients $\binom{N+n}{0},\ldots,\binom{N+n}{n+1}$, the expansion (3) can be rewritten more compactly using convolutions as

$$(3)' \qquad (1_{n+1} * \mathbb{P})^N = b_n^N \bullet \mathbb{D}_{n+1} + \epsilon_n^N$$

where $\epsilon_n^N \equiv P^{N+n}(-D)^{n+1}(I - P^*)$. Observe that $\epsilon_n^N = o_{d_P}(1)$, i.e., $\epsilon_n^N \to 0$ (C, d_P), since P and D commute, $P^N \to P^*$ (C, d_P) and $P^*P^* = P^*$. Thus, $(3)'$ provides an expansion that is a sum of a polynomial in N of degree $n+1$ and an error term that converges to 0 (C, d_P).

The case in which $n = 0$ is of special interest. In that event, $(3)'$ becomes

$$(3)'' \qquad \sum_{i=0}^{N-1} P^i = NP^* + D + o_{d_P}(1).$$

Notice that this representation is simply a restatement of the representation (7) of the deviation matrix in §1.8 as may be seen by subtracting NP^* from both sides of the above equation and taking the (C, d_P) limit as $N \to \infty$.

The expansion $(3)'$ is the sum of a polynomial in N and a small error term, and is analogous to a corresponding expansion of $\rho^{-n}R^\rho$ as the sum of a polynomial in ρ^{-1} and a small error term. The last expansion is obtained from the Laurent expansion of R^ρ and is

$$\rho^{-n}R^\rho = \rho^{-n-1}P^* + \rho^{-n}D + \cdots + (-1)^n D^{n+1} + o(1).$$

Observe that the polynomial in this expansion has the same degree $n+1$ and the same coefficient matrix $\mathbb{D}_{n+1}$ as that in $(3)'$. Incidentally, the multiplication of R^ρ by ρ^{-n} in the above expression arises because the present value of 1_n used in $(3)'$, when discounted to period $-n$, is ρ^{-n}. In the special case in which $n = 0$, this last expansion becomes

$$R^\rho = \rho^{-1}P^* + D + o(1).$$

The right-hand side of this expansion is similar to $(3)''$ with ρ^{-1} replacing N.

Proof of Theorem 28. To prove (3), it suffices to show for $N \geq 1$ and $n \geq -1$ that

$$(1_{n+1} * \mathbb{P})^N P^* = \binom{N+n}{n+1} P^* \tag{4}$$

and

$$(1_{n+1} * \mathbb{P})^N (I - P^*) = \sum_{j=0}^{n} \binom{N+n}{n-j} (-1)^j D^{j+1} + P^{N+n}(-D)^{n+1}(I - P^*), \tag{5}$$

since (3) follows by adding (4) and (5). To prove (4), observe that

$$(1_{n+1} * \mathbb{P})^N P^* = (1_{n+1} * 1_1)^{N-1} P^* = 1_{n+2}^{N-1} P^* = \binom{N+n}{n+1} P^*.$$

The next step is to establish (5) by induction on n. The result is trivial for $n = -1$. Now observe that $(1_1 * \mathbb{P})(I - P) = (1_1 - 1_0)I - P\mathbb{P}$. Postmultiplying this equation by D and using the facts that $DP^* = P^*D = 0$ and $-DQ = I - P^* = -QD$ from Theorem 20 yields

$$(1_1 * \mathbb{P})(I - P^*) = (1_1 - 1_0)D - P\mathbb{P}(I - P^*)D, \tag{6}$$

establishing (5) for $n = 0$. Suppose now that (5) holds for n (≥ 0) and consider $n + 1$. By taking the convolution of (6) with 1_n and then using the bilinearity and associativity of convolution, the induction hypothesis, the fact that $-D^{n+1}Q = D^n = -QD^{n+1}$, and (2), it follows that for $N \geq 1$,

$$\begin{aligned}
(1_{n+1} * \mathbb{P})^N (I - P^*) &= (1_{n+1}^N - 1_n^N)D - P(1_n * \mathbb{P})^N (I - P^*)D \\
&= \binom{N+n-1}{n} D + P\Big(\sum_{j=1}^{n} \binom{N+n-1}{n-j} (-1)^j D^j - P^{N+n-1}(-D)^n (I - P^*)\Big) D \\
&= \binom{N+n-1}{n} D + \sum_{j=1}^{n} \binom{N+n-1}{n-j} (-1)^j D^j (D - I) + P^{N+n}(-D)^{n+1}(I - P^*) \\
&= \sum_{j=0}^{n} \Big(\binom{N+n-1}{n-j} + \binom{N+n-1}{n-j-1} \Big) (-1)^j D^{j+1} + P^{N+n}(-D)^{n+1}(I - P^*) \\
&= \sum_{j=0}^{n} \binom{N+n}{n-j} (-1)^j D^{j+1} + P^{N+n}(-D)^{n+1}(I - P^*),
\end{aligned}$$

which establishes (5). ∎

Polynomial Expansion of N-Period Values

Recall that $\mathbb{V}_\delta^m = 1_m * \mathbb{V}_\delta = (1_m * 1_1) * (1_{-1} * \mathbb{V}_\delta) = 1_{m+1} * \mathbb{V}_\delta^{-1} = (1_{m+1} * \mathbb{P}_\delta) r_\delta$. Thus, it follows from the polynomial expansion (3) and (3)′ of the sum of the expected sojourn times of individuals in the first N periods with binomial immigration of order $n = m$ when using δ^∞ that

$$V_\delta^{Nm} = \sum_{j=-d}^{m} \binom{N+m}{m-j} v_\delta^j - P_\delta^N u_\delta^m = b_m^N \bullet V_\delta^m - P_\delta^N u_\delta^m \tag{7}$$

for $N \geq 1$ and $m \geq 0$ where $u_\delta^m \equiv P_\delta^m v_\delta^m$. Observe that this formula gives an expansion of V_δ^{Nm} as the sum of a polynomial in N of degree $m+d$ with coefficient matrix V_δ^m and a term that is $o_d(1)$.

Although the error term is small, it is convenient to eliminate it entirely by including an appropriate terminal reward. This important idea will play a fundamental role in the sequel. To obtain the desired exact formula, let $V_\delta^{Nm}(u) = V_\delta^{Nm} + P_\delta^N u$ where u is the terminal reward received at the end of period N. It follows from (7) that $V_\delta^{Nm}(u_\delta^m)$ is a polynomial in N of degree $m+d$ with

$$V_\delta^{Nm}(u_\delta^m) = \sum_{j=-d}^{m} \binom{N+m}{m-j} v_\delta^j = b_m^N \bullet V_\delta^m \tag{8}$$

for $N \geq 1$ and $m \geq 0$. Observe that since $d \leq 1$, it follows from (7), (8) and $P_\delta^* D_\delta = 0$ that $\lim_{N\to\infty}[V_\delta^{Nm} - V_\delta^{Nm}(u_\delta^m)] = 0$ (C, d) for $m \geq 0$. Notice from (8) that δ maximizes $V_\delta^{Nm}(u_\delta^m)$ over Δ for all large enough N if and only if δ maximizes V_δ^m lexicographically. This is a strong form of overtaking optimality for stationary policies except that the terminal value u_δ^m depends on δ.

The case in which $m = 0$ is of special interest. In that event, (8) expresses the fact that with the proper choice of the terminal reward, viz., u_δ^0, the N-period value

$$V_\delta^N(u_\delta^0) = Nv_\delta^{-1} + v_\delta^0$$

is affine in N. Moreover, the slope

$$v_\delta^{-1} = P_\delta^* r_\delta = \lim_{N\to\infty} N^{-1} V_\delta^N$$

is the *stationary reward* or *reward rate* and the intercept

$$v_\delta^0 = D_\delta r_\delta = \lim_{N\to\infty} (V_\delta^N - Nv_\delta^{-1}) \ (\mathrm{C}, d)$$

is the (C, d) limit of the difference between the N-period values starting from each state and starting with the stationary distribution therein.

For every policy $\pi = (\gamma_i)$, decision δ, $u \in \Re^S$, $m, N \geq 0$, and $L \geq 1$, let $\pi^L \equiv (\gamma_1, \dots, \gamma_L)$, $\pi_L \equiv (\gamma_{L+1}, \gamma_{L+2}, \dots)$ and $\pi^L\delta \equiv (\pi^L, \delta^\infty)$. If $\gamma, \delta \in \Delta$, let $\gamma^L \equiv (\gamma^\infty)^L$, $\gamma\delta \equiv (\gamma, \delta^\infty)$ and $\gamma^L\delta \equiv (\gamma^L, \delta^\infty)$. Subscripts on decisions will be reserved for enumerating them.

The next lemma expresses $V_\pi^{Nm}(u)$ as the sum of the expected rewards earned in the first L periods and those earned in the subsequent $M \equiv N - L \geq 0$ periods when $m \geq 0$. The result follows readily from the facts that $\mathbb{V}_\pi^m = 1_{m+1} * \mathbb{V}_\pi^{-1}$, $1_{m+1}^N = \binom{N+m}{m}$ and $\mathbb{V}_\pi^{-1} = (0 \ P_\pi^0 r_{\gamma_1} \ P_\pi^1 r_{\gamma_2} \cdots)$.

Lemma 29. *If $\pi = (\gamma_i) \in \Delta^\infty$, $u \in \Re^S$, $L, M, m \geq 0$, and $N = L + M \geq 1$, then*

$$V_\pi^{Nm}(u) = \sum_{i=1}^{L} \binom{N-i+m}{m} P_\pi^{i-1} r_{\gamma_i} + P_\pi^L V_{\pi_L}^{Mm}(u).$$

Proof. Use the fact that $V_\pi^{Nm}(u) = \sum_{i=1}^{L}\binom{N-i+m}{m}P_\pi^{i-1}r_{\gamma_i} + \sum_{i=L+1}^{N}\binom{N-i+m}{m}P_\pi^{i-1}r_{\gamma_i} + P_\pi^N u$. ∎

Comparison Lemma for N-Period Values

Unfortunately, there is no exact comparison lemma for the difference of the N-period values of two policies in terms of a comparison function. However, it is possible to give an exact comparison lemma for "eventually stationary" policies with a suitable terminal reward. Subsequently, we show that the error made in altering the problem in this way is comparatively small.

The following lemma gives an exact representation for the difference between the N-period value $V_{\pi^L\delta}^{Nm}(u_\delta^m)$ of the eventually stationary policy $\pi^L\delta$ and the N-period value $V_\delta^{Nm}(u_\delta^m)$ of the corresponding stationary policy δ^∞, both with the terminal reward u_δ^m. The exactness of the representation is a consequence of this choice of the terminal reward. The representation is in terms of the *time-domain comparison function*

$$G_{\gamma\delta}^{km} \equiv \sum_{j=-d}^{m}\binom{k+m}{m-j}g_{\gamma\delta}^j = b_m^k \bullet G_{\gamma\delta}^m \tag{9}$$

where $k, m \geq 0$ and $\gamma,\delta \in \Delta$.

Lemma 30. *If $\pi = (\gamma_i) \in \Delta^\infty$, $\delta \in \Delta$, $d_\delta \leq 1$, $0 \leq L \leq N$, and $m \geq 0$, then*

$$V_{\pi^L\delta}^{Nm}(u_\delta^m) - V_\delta^{Nm}(u_\delta^m) = \sum_{i=1}^{L}P_\pi^{i-1}G_{\gamma_i\delta}^{N-i,m}.$$

Proof. First establish the result for $L = 1$. Put $\gamma = \gamma_1$. By Lemma 29 with $L = 1$, $V_{\gamma\delta}^{Nm}(u_\delta^m) = \binom{N+m-1}{m}r_\gamma + P_\gamma V_\delta^{N-1,m}(u_\delta^m)$. Using this fact, (8), the identity (2) for binomial coefficients, and $v_\delta^{-d-1} = 0$, one finds that

$$\begin{aligned}
V_{\gamma\delta}^{Nm}(u_\delta^m) - V_\delta^{Nm}(u_\delta^m) &= \binom{N+m-1}{m}r_\gamma + P_\gamma\sum_{j=-d}^{m}\binom{N+m-1}{m-j}v_\delta^j - \sum_{j=-d}^{m}\binom{N+m}{m-j}v_\delta^j \\
&= \binom{N+m-1}{m}r_\gamma + Q_\gamma\sum_{j=-d}^{m}\binom{N+m-1}{m-j}v_\delta^j - \sum_{j=-d}^{m-1}\binom{N+m-1}{m-j-1}v_\delta^j \\
&= G_{\gamma\delta}^{N-1,m},
\end{aligned}$$

which establishes the result for $L = 1$. This fact and Lemma 29 imply that for $L \geq 1$,

$$\begin{aligned}
V_{\pi^L\delta}^{Nm}(u_\delta^m) - V_\delta^{Nm}(u_\delta^m) &= \sum_{i=1}^{L}[V_{\pi^i\delta}^{Nm}(u_\delta^m) - V_{\pi^{i-1}\delta}^{Nm}(u_\delta^m)] \\
&= \sum_{i=1}^{L}P_\pi^{i-1}[V_{\gamma_i\delta}^{N-i+1,m}(u_\delta^m) - V_\delta^{N-i+1,m}(u_\delta^m)] = \sum_{i=1}^{L}P_\pi^{i-1}G_{\gamma_i\delta}^{N-i,m}.
\end{aligned}$$ ∎

The above result permits us to deduce the following important Comparison Lemma which gives an approximate representation for the difference between the N-period values of an arbitrary policy and a stationary policy—both with a binomial immigration stream of order m.

Lemma 31. Comparison Lemma. *If $d \leq 1$, $\pi = (\gamma_i) \in \Delta^\infty$, $\delta \in \Delta$, $N \geq M \geq 0$ and $m \geq 0$, then*

$$V_\pi^{Nm} - V_\delta^{Nm} = \sum_{i=1}^{N-M}P_\pi^{i-1}G_{\gamma_i\delta}^{N-i,m} + O(N^{d-1}) \text{ uniformly in } \pi.$$

Proof. It follows from Lemma 29 and the System-Degree Theorem 9 that

$$V_\pi^{Nm} - V_{\pi^{N-M}\delta}^{Nm}(u_\delta^m) = P_\pi^{N-M}[V_{\pi_{N-M}}^{Mm} - V_\delta^{Mm}(u_\delta^m)] = O(N^{d-1}) \text{ uniformly in } \pi$$

for $N \geq M$, and the same is so when δ^∞ replaces π. The result then follows by observing that the error from replacing $V_{\pi^{N-M}\delta}^{Nm}(u_\delta^m)$ by V_π^{Nm} and $V_\delta^{Nm}(u_\delta^m)$ by V_δ^{Nm} in Lemma 30 is $O(N^{d-1})$. ∎

Cesàro Overtaking Optimality with Binomial Immigration Streams

Now a policy λ is Cesàro overtaking optimal for the binomial immigration stream of order n if

$$\liminf_{N\to\infty} (V_\lambda^{Nn} - V_\pi^{Nn}) \geq 0 \text{ (C, } d\text{) for all } \pi \tag{10}$$

or, equivalently, since $\sum_{i=0}^N V_\pi^{in} = (1_1 * (1_n * \mathbb{V}_\pi))^N = (1_{n+1} * \mathbb{V}_\pi)^N = V_\pi^{N,n+1}$ and $\lim_{N\to\infty} \frac{N}{N+1} = 1$,

$$\liminf_{N\to\infty} N^{-d}(V_\lambda^{N,n+d} - V_\pi^{N,n+d}) \geq 0 \text{ for all } \pi. \tag{10}'$$

If $d = 0$, Cesàro overtaking optimality for the binomial immigration stream of order n is the same as overtaking optimality therefor. If $d = 1$, the (C, d) limits are needed to damp out the effect of the (bounded) error term in the Comparison Lemma 31.

Call a policy *strong Cesàro overtaking optimal* if it is Cesàro overtaking optimal for binomial immigration streams of all orders. Now apply Lemma 31 to establish the next result.

Theorem 32. Existence and Characterization of Stationary Cesàro-Overtaking-Optimal Policies. *In a bounded system with system degree d, there is a stationary strong Cesàro-overtaking-optimal policy. Also, δ^∞ is Cesàro overtaking optimal for the binomial immigration stream of order $n \geq -d$ if and only if δ^∞ is n-optimal. Further, δ^∞ is strong Cesàro overtaking optimal if and only if δ^∞ is ∞-optimal.*

Proof. Since the system is bounded, it follows from Theorem 22 that there is a stationary policy δ^∞ that has strong maximum present value and for which $G_{\gamma\delta} \preceq 0$ for all $\gamma \in \Delta$. Thus, if $M > 0$ is large enough, it follows from (9) that $G_{\gamma\delta}^{k,n+d} \leq 0$ for all $\gamma \in \Delta$ and $k \geq M$. Also since the system is bounded, $d_\pi \leq 1$ for all π by the System-Degree Theorem 9. Hence, by the Comparison Lemma 31 (with $m = n + d$ on noting that $N - i \geq M$ for $i \leq N - M$), $\lambda = \delta^\infty$ satisfies (10)′, i.e., δ^∞ is Cesàro overtaking optimal for the binomial immigration stream of order n. Since this is so for all n, δ^∞ is strong Cesàro overtaking optimal. The two characterizations follow from (10)′ and (7). ∎

Special Cases. Several examples of Cesàro overtaking optimality for binomial immigration streams of order n appear below. Each of these examples provides an alternate interpretation of Cesàro overtaking optimality as a natural concept of optimality for the system without exogenous immigration after the first period.

Reward-Rate Optimality. Since $1_{-1} = (I \ \ -I \ \ 0 \ \ 0 \ \cdots)$, $V_\pi^{N,-1} = P_\pi^{N-1} r_{\gamma_N}$ is the expected reward that $\pi = (\gamma_i)$ earns in period N. Thus, it is natural to call a policy *reward-rate optimal* if it is Cesàro overtaking optimal for the binomial immigration stream of order -1.

As an example, consider the two-state substochastic deterministic system in Figure 15. There are two decisions γ, δ with rewards the numbers on each arc. Observe that both γ^∞ and δ^∞ have common unit reward rate starting from either state, and so both are reward-rate optimal. By contrast, $V^N_{\delta 1} - V^N_{\gamma 1} = 2$ for $N \geq 1$, so only δ^∞ is overtaking optimal.

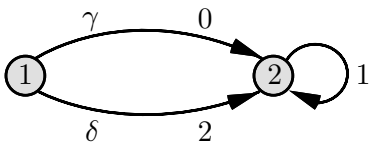


Figure 15

Cesàro Overtaking Optimality. Since $1_0 = (I \quad 0 \quad 0 \quad \cdots)$, $V^{N0}_\pi = V^N_\pi$. Thus, Cesàro overtaking optimality for the binomial stream of order zero is ordinary Cesàro overtaking optimality. In a transient system, a policy is Cesàro overtaking optimal if and only if it has maximum value. Moreover, Cesàro overtaking optimality is well defined for bounded systems. By contrast, for such systems, values are often undefined or equal $\pm\infty$ (as in Figure 15), rendering the maximum-value criterion insufficiently selective. Thus Cesàro overtaking optimality is the proper way of extending the concept of maximum value from transient to bounded systems.

Now consider the two-state substochastic deterministic system that Figure 16 illustrates. There are two decisions γ and δ with rewards the numbers on each arc. Then both γ and δ have common value zero starting from state one, and so are Cesàro overtaking optimal.

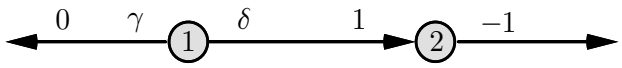


Figure 16

However, if the immigration stream is binomial of order one, each individual who arrives in state one in period N and uses δ^∞ earns a unit *temporary* reward in that period. Moreover, as Figure 17 illustrates, the aggregate of the temporary rewards that all individuals generate when they use δ^∞ provides a *permanent* unit reward for the system, i.e., $V^{N1}_{\delta 1} = 1$ for all $N \geq 1$. Thus only δ^∞ maximizes the total reward for the system with binomial immigration of order one.

Float Optimality. Since $1_1 = (I \quad I \quad I \quad \cdots)$, V^{N1}_π is the N-period *float*, i.e., the sum $\sum_{i=1}^N V^i_\pi$ of the cash positions in each period $1 \leq i \leq N$. For this reason, call a policy *float optimal* if it is Cesàro overtaking optimal for the binomial immigration stream of order one. As an example, consider the S-station queueing network discussed at the beginning of this section in which the expected exogenous arrivals in each period into each station depends on the station, but not on the period. Then the stream of expected arrivals is a nonnegative diagonal matrix times the binomial stream of order one. Thus if the goal is to find a policy that is Cesàro overtaking optimal

Immigration		**Periods N**				
Cohort	$\boldsymbol{w_1^{N-1}}$	**1**	**2**	**3**	**4**	$\cdots$
1	1	1	-1			
2	1		1	-1		
3	1			1	-1	
$\vdots$	$\vdots$					$\cdots$
Total Rewards		**1**	**0**	**0**	**0**	$\cdots$

Figure 17. Rewards each Immigrant Earns in each Period with $w = \text{diag}(1 \quad 0) \cdot 1_1$ when Using δ^∞

for the queueing network, it suffices to find a policy that is float optimal with no exogenous immigration after the first period.

Value Interpretation of Immigration Stream

In each of the above examples, an immigration stream w, or more precisely the *cumulative immigration stream* $W = (W^N) \equiv w * 1_1$, has an alternate interpretation as values of rewards an arbitrary policy π earns with no immigration after period one. To see this, let $\mathbb{V}_\pi^w = (V_\pi^{Nw})$ be the sequence of N-period values that π earns when the immigration stream is w. Then it follows that $\mathbb{V}_\pi^w = w * \mathbb{V}_\pi = W * \mathbb{V}_\pi^{-1}$, so that $V_\pi^{Nw} = \sum_{i=0}^N W^{N-i} V_\pi^{i,-1} = \sum_{i=0}^N W^i V_\pi^{N-i,-1}$. Moreover, since $V_\pi^{i,-1} = P_\pi^{i-1} r_{\gamma_i}$ is the expected reward that $\pi = (\gamma_i)$ earns in period i, it follows that W^i can be thought of alternately as the unit value of rewards in period $N-i$, i.e., with $i+1$ periods remaining.

As examples, observe that when w is the binomial stream of order -1, then $W = 1_0$ so $W^N = 0$ for $N > 0$, i.e., the weight is on rewards only in the last period. In this event, Cesàro overtaking optimality is equivalent to maximizing the reward rate. By contrast, if w is the binomial stream of order 0, then $W = 1_1$, so $W^N = I$ for all $N \geq 0$, i.e., there is equal weight on rewards in each period. In this event, Cesàro overtaking optimality is equivalent to ordinary Cesàro overtaking optimality. If instead, $w = 1_n$ and $n > 0$, then $W = 1_{n+1}$, i.e., the weight on rewards with i periods remaining is a polynomial in i of degree n. Thus as n increases, the absolute and relative weights on rewards in early periods increases.

Evidently, this *value interpretation of* (*cumulative*) *immigration streams* encompasses most standard optimality concepts (in which there is no immigration after the first period). Moreover, in this setting, the *cohort and Markov policies coincide*.

Combining Physical and Value Immigration Streams

So far there are two distinct interpretations of immigration streams, viz., *physical* and *value*. Can they be combined as would be desirable, for example, in the study of queueing networks where both are needed? The answer is 'yes'. Indeed, Cesàro overtaking optimality for a system in which there is a physical immigration stream w and a value one w' is equivalent to Cesàro

overtaking optimality for the system with the single combined immigration stream $w' * w$. To see this, observe that the stream of N-period values with the physical stream w and policy π is simply $w * \mathbb{V}_\pi$. Now if it is desirable to impose the value stream w' on this system, the new value-weighted stream of N-period values is $w' * (w * \mathbb{V}_\pi) = (w' * w) * \mathbb{V}_\pi$, which justifies our claim.

Thus, for example, if the vector of expected arrivals into the stations of the queueing network is the diagonal matrix w^0 in each period and the goal is to maximize the float of the network, then $w = w^0 \cdot 1_1$ and $w' = 1_1$, so the combined immigration stream is $w' * w = w^0 \cdot 1_2$. On the other hand if the goal is merely to maximize the network's reward rate, then $w' = 1_{-1}$ and the combined stream is $w' * w = w^0 \cdot 1_0$.

Cesàro Overtaking Optimality with More General Immigration Streams

It is now time to generalize the above results concerning the existence of stationary Cesàro-overtaking-optimal policies from binomial to a more general class of immigration streams which includes positive linear combinations of binomial streams. The basic idea is to reduce Cesàro overtaking optimality for the more general class of immigration streams to that for corresponding binomial immigration streams. To see how this can be done, it is useful first to establish a preliminary lemma.

Lemma 33. Nonnegativity of Cesàro Limits of Convolutions. *If* $B = (B^0 \;\; B^1 \;\cdots)$ *and* $C = (C^0 \;\; C^1 \;\cdots)$ *are sequences of real matrices or scalars for which the convolution* $B * C$ *is defined,* B *is nonnegative and has degree zero, and* $\liminf_{N\to\infty} C^N \geq 0$ (C, d) *with* $0 \leq d \leq 1$*, then necessarily* $\liminf_{N\to\infty} (B * C)^N \geq 0$ (C, d).

Proof. Since $\liminf_{N\to\infty} C^N \geq 0$ (C, d), $\liminf_{N\to\infty} N^{-d}\widehat{C}^N \geq 0$ where $\widehat{C} \equiv C * 1_d$. Let $\epsilon \gg 0$ be a given matrix. Choose M so that $i^{-d}\widehat{C}^i \geq -\epsilon$ for $i \geq M$. Then since $B \geq 0$ has degree zero, there is a matrix K not depending on M or ϵ such that

$$(B * \widehat{C})^N = \sum_{i=0}^{M-1} B^{N-i}\widehat{C}^i + \sum_{i=M}^{N} B^{N-i}\widehat{C}^i \geq O(N^{d-1}) - \epsilon K N^d.$$

Therefore, $\liminf_{N\to\infty} N^{-d}(B * \widehat{C})^N \geq -\epsilon K$, whence $\liminf_{N\to\infty} (B * C)^N \geq -\epsilon K$ (C, d). Thus because ϵ is arbitrary, $\liminf_{N\to\infty} (B * C)^N \geq 0$ (C, d). ∎

In many practical settings, interest centers on immigration streams that are more general than binomial. One such class is the set $\underline{\mathcal{W}}_n$ of immigration streams w for which $w * 1_{-m}$ is nonnegative and has degree zero for some $m \leq n$. For example, the binomial streams 1_m of order $m \leq n$ are in $\underline{\mathcal{W}}_n$ because $1_m * 1_{-m} = 1_0$ is nonnegative and has degree zero.

Denote by $\mathcal{W}_n$ the convex cone that consists of the nonnegative linear combinations of elements of $\underline{\mathcal{W}}_n$. It is of interest that $\mathcal{W}_n \supset \underline{\mathcal{W}}_n$. To see why, let $w = 1_n + 2 \cdot 1_{n-1}$, so $w \in \mathcal{W}_n$. However, $w * 1_{-n} = 1_0 + 2 \cdot 1_{-1} = (3I \;\; -2I \;\; 0 \;\; 0 \;\; \cdots)$ is not nonnegative and $w * 1_{-n+1} = 1_1 + 2 \cdot 1_0 =$

$(3I \;\; I \;\; I \cdots)$ does not have degree zero. Thus $w * 1_{-m}$ is not nonnegative for $m \geq n$ and does not have degree zero for $m < n$. Hence $w \notin \underline{\mathcal{W}}_n$.

Theorem 34. Reduction of Immigration Streams to Binomial Ones. *In a bounded system for which $w \in \mathcal{W}_n$ with $n \geq -1$, there is a stationary policy that is Cesàro overtaking optimal for w, viz., any stationary policy that is Cesàro overtaking optimal for the binomial immigration stream of order n.*

Proof. Suppose $\delta \in \Delta_n$ and the system degree is d. By hypothesis, $w = \sum_{m=k}^{n} w_m$ for some $k \leq n$ and immigration streams $w_k, \ldots, w_n$ such that $w_m * 1_{-m}$ is nonnegative and has degree zero for $k \leq m \leq n$. For each policy π,

$$\mathbb{V}_\delta^w - \mathbb{V}_\pi^w = w * (\mathbb{V}_\delta - \mathbb{V}_\pi) = \sum_{m=k}^{n} [(w_m * 1_{-m}) * (\mathbb{V}_\delta^m - \mathbb{V}_\pi^m)].$$

Let $\mathcal{L}$ (resp., $\mathcal{L}_m$) denote the (C, d) limit inferior of the sequence on the left-hand side (resp., sequence in brackets on the right-hand side) of the above equation. Since $\delta \in \Delta_n \subseteq \Delta_m$ for each $m \leq n$, it follows that $\mathcal{L}_m \geq 0$ by Lemma 33. Hence $\mathcal{L} \geq \sum_{m=k}^{n} \mathcal{L}_m \geq 0$. ∎

One may ask whether the hypothesis $w \in \mathcal{W}_n$ in Theorem 34 can be dropped. The answer is that it cannot as the following example illustrates.

Example. Nonexistence of a Stationary Cesàro-Overtaking-Optimal Policy When $w \notin \mathcal{W}_n$ for all $n \geq -1$. Consider the two-state substochastic system that Figure 18 illustrates. The rewards and actions available in each state appear on the arcs. Let $w^0 = \mathrm{diag}(1 \;\; -1)$ and $w^i = 0$ for $i \geq 1$. The nonstationary policy π that takes action a in the first period and action b in the second is the unique (Cesàro-) overtaking-optimal policy because it earns the system 1 starting from each state. By contrast, all stationary policies earn the system -1 or 0 in some state and so are not *Cesàro overtaking optimal for w*. Thus, by Theorem 34, it must be so that $w = (w^i) \notin \mathcal{W}_n$ for all $n \geq -1$.

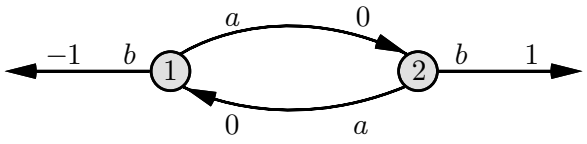


Figure 18

It can be shown that the condition that $w * 1_{-m}$ has degree zero is not essential to Theorem 34. But in that event it is necessary to replace Cesàro limits of order one by higher-order Cesàro limits. For that reason we omit a discussion of the more general case.

Future-Value Optimality

The above result has a useful application to problems in which the interest rate $100\rho\%$ is negative and $-1 < \rho$. Recall from §1.7 that the last assumption means that one cannot lose more than one invests. Also, negative interest rates arise when the decision maker is interested in inflation-adjusted rewards and in which the after-tax rate of interest is less than the rate of inflation, a not uncommon situation. In this circumstance, the inflation-adjusted present value of a policy is often a divergent series, and so is not well defined. The way out of this difficulty is to instead *discount income to the future*.

The *future value* $\widehat{V}_\pi^{N\rho}$ in period N of a policy $\pi = (\gamma_i)$ is

$$\widehat{V}_\pi^{N\rho} \equiv \sum_{i=1}^{N} \alpha^{N-i} P_\pi^{i-1} r_{\gamma_i} = (\mathbb{A} * \mathbb{V}_\pi^{-1})^N = ((\mathbb{A} * 1_{-1}) * \mathbb{V}_\pi)^N$$

where $\alpha \equiv 1 + \rho$ and $\mathbb{A} \equiv (\alpha^0 I \;\; \alpha^1 I \;\; \alpha^2 I \;\; \cdots)$. Observe that $W = \mathbb{A}$ can be thought of alternately as a cumulative immigration stream. The hypotheses on ρ imply that $0 < \alpha < 1$. Call a policy λ *future-value optimal* if

$$\liminf_{N\to\infty} (\widehat{V}_\lambda^{N\rho} - \widehat{V}_\pi^{N\rho}) \geq 0 \;\; (\mathrm{C}, d) \text{ for all } \pi.$$

Theorem 35. Reduction of Future-Value to Reward-Rate Optimality. *In a bounded system, there is a stationary policy that is future-value optimal for all $-1 < \rho < 0$, viz., any stationary reward-rate optimal policy.*

Proof. Set $w \equiv \mathbb{A} * 1_{-1}$. Then $w * 1_1 = \mathbb{A}$ is nonnegative and has degree zero, so $w \in \mathcal{W}_{-1}$. Now apply Theorem 34 with $n = -1$. ∎

10 SUMMARY

Figures 19 and 20 below provide directed *assertion-implication* graphs that summarize many of the results of this chapter, the homework problems and a few others. The graphs describe relations among the various optimality concepts for transient and bounded systems.

The nodes of the graphs are *assertions* and the arcs are *implications.* An arc (arrow) from one assertion to another signifies that the first assertion implies the second. Similarly, a double arrow between two assertions signifies that the two assertions are equivalent. A chain from one assertion to another assures that the first assertion implies the second. Each strong component of a graph contains a maximal collection of equivalent assertions. For example, the assertions above the dashed line are equivalent.

In Figure 20, some arcs have text labels. When this is the case, an arc's implication is valid when the statement in the arc's text label is valid. For example, "δ^∞ overtaking optimal" implies "δ^∞ Cesàro overtaking optimal". The converse is generally false as the example of Figure 13 illustrates. However, it can be shown that "δ^∞ Cesàro overtaking optimal" implies "δ^∞ overtaking optimal" *provided* that a stationary overtaking-optimal policy exists.

Since there exists a stationary strong present-value optimal policy, there is a stationary policy that is optimal for each of the concepts considered here except, as discussed above, overtaking optimality. However, *there is a stationary Cesàro-overtaking-optimal policy, and each such policy is overtaking optimal provided only that a stationary overtaking-optimal policy exists.*

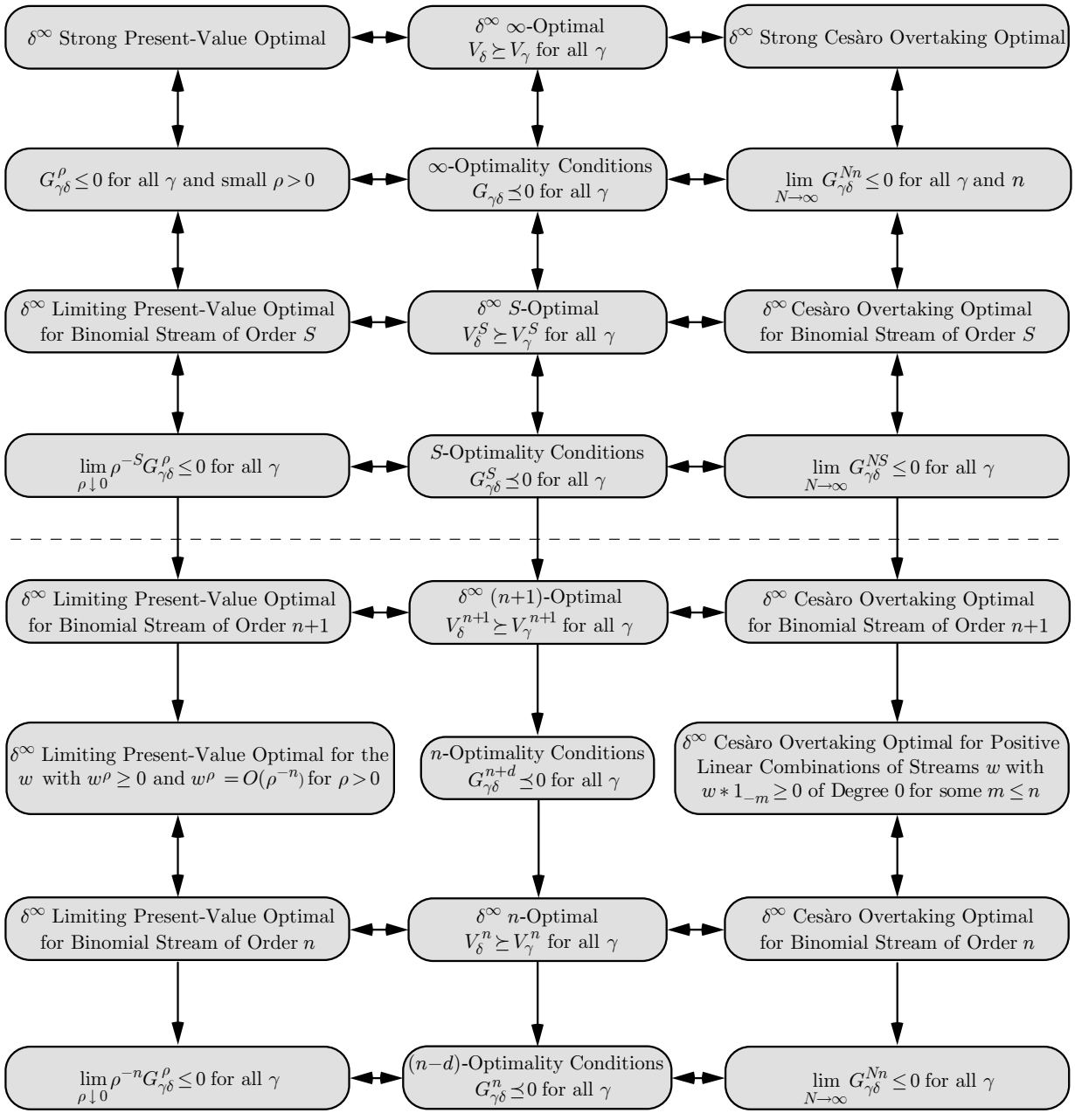


Figure 19. Optimality of δ^∞ for a Bounded System with Immigration

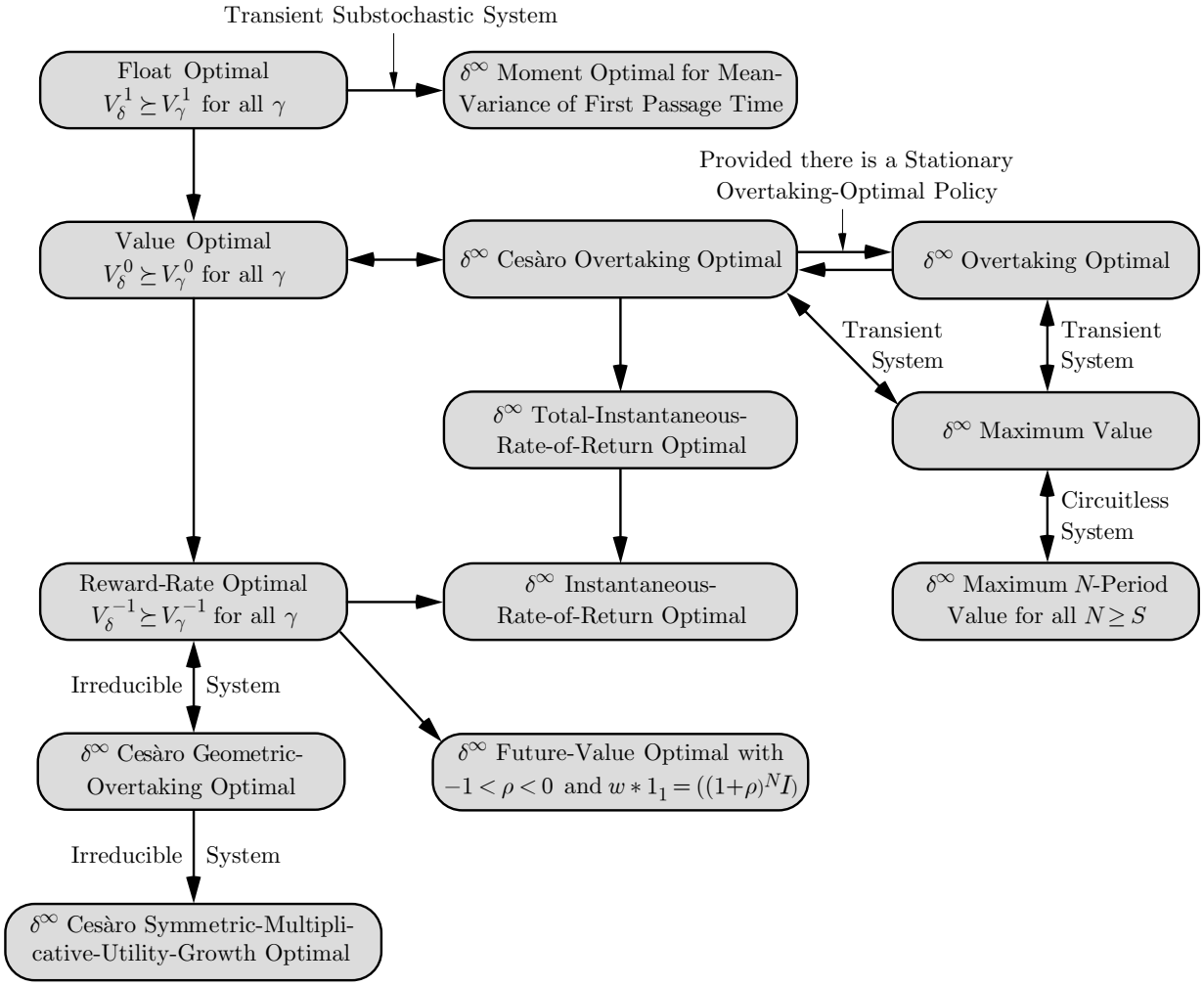


Figure 20. Optimality of δ^∞ for Some Special Bounded Systems

2

Team Decisions, Certainty Equivalents and Stochastic Programming

1 FORMULATION AND EXAMPLES [Be55], [Da55], [Ra55], [Ve65], [MR72]

Team decision theory is concerned with a *team*, i.e., collection, of individuals that each

- control different decision variables,
- base their decisions on possibly different information, and
- share a common objective.

In the sequel, we shall formulate such problems as *stochastic programs*.

A team is composed of n members labeled $1, \ldots, n$. The i^{th} team member observes a random vector Z_i assuming values in a set $\mathbb{Z}_i$ and then uses a real-valued *decision* $X_i : \mathbb{Z}_i \to \Re$ to select an action $X_i(Z_i)$. This formulation encompasses team members that have vector-valued decisions since each such member can be considered to be a group of different team members each with common information and real-valued decisions. Let $Z = (Z_1, \ldots, Z_{n+1})$ be the vector whose elements are the n random vectors $Z_1, \ldots, Z_n$ observed by the n team members and a random

vector Z_{n+1} that contains a (possibly empty) subset of those random vectors as well as relevant unobserved information. The distribution of Z is known to all members of the team. Denote by $\mathbb{Z}$ the range of Z, which we take to be finite. Call $X \equiv (X_1, \ldots, X_n)$ a *team decision*. Let $\mathbb{X}$ denote the set of all team decisions. Let $r(\,\cdot\,,\,\cdot\,) : \Re^n \times \mathbb{Z} \to \Re$ be the given *reward function* and $g(\,\cdot\,,\,\cdot\,) : \Re^n \times \mathbb{Z} \to \Re^m$ be the given *constraint function*. The problem is to choose X to maximize

$$R(X) \equiv \mathrm{E}r(X(Z), Z) \tag{1}$$

subject to

$$g(X(Z), Z) \geq 0 \tag{2}$$

and

$$X \in \mathbb{X}. \tag{3}$$

One of the principal goals of team decision theory is to study the effect of different information structures on the quality of optimal decisions as measured by the maximum expected reward. Of course, the more information that each team member receives, the greater the maximum expected reward, but generally so also is the cost of collecting and distributing the information. Thus one goal of team decision theory is to quantify the value of additional information to facilitate comparison with the costs of providing and using such information. This enables the team to choose the proper amount of information to provide each member. From this viewpoint, team decision theory can be considered to be a branch of the field of *management information systems*.

Example 1. Airline Reservations with Uncertain Demand. An airline wishes to give its agents in n cities rules for deciding how many seats to book on a flight of capacity $K > 0$ so as to maximize the expected net revenue from the flight where $Z_{n+1} \equiv (D_1, \ldots, D_n) \geq 0$ is the random vector of demands experienced by agents $1, \ldots, n$. Let X_i be the number of tickets sold by agent i. Then

$$r(X, Z) = r\sum_{i=1}^{n} X_i - p(\sum_{i=1}^{n} X_i - K)^+$$

where r is the ticket price and p is the penalty for each oversold ticket. Also (2) becomes

$$0 \leq X_i \leq D_i,\ 1 \leq i \leq n.$$

Information Structures. The information on ticket demands can be shared among the agents in many ways. We mention four examples.

(a) Complete Communication. Let $Z_i = Z_{n+1}$ for all i, i.e., every agent gets complete information about tickets demanded at other agents.

(b) Decentralization. Let $Z_i = D_i$ for all i, i.e., each agent receives information only on his own ticket demand.

(c) Partial Communication. Let $Z_i = (D_i, \sum_{j=1}^n D_j)$, i.e., each agent receives information on his own demand and the total demand at all agents.

(d) Management by Exception. Let $Z_i = (D_i, T)$ where

$$T = \begin{cases} \sum_{j=1}^n D_j, \text{ if } \sum_{j=1}^n D_j > M \\ 0 \qquad\quad \text{, otherwise,} \end{cases}$$

i.e., each agent receives information on his demand and, if the total demand at all agents exceeds a threshold M, the total demand at all agents. In this event, M can be chosen to be at least K. For if an agent is informed that the total ticket demand at all agents does not exceed the capacity of the plane, there is nothing to be gained by telling the agent the precise demands at the other agents.

Example 2. Inventory Control with Uncertain Demand. An inventory manager seeks to minimize the expected cost of storage and shortage of a single product over n weeks. Let X_i and D_i be respectively the cumulative orders and demands in weeks $1, \ldots, i$ where $i = 1, \ldots, n$. Let $Z_{n+1} = (D_1, \ldots, D_n)$. Then the storage and shortage cost is

$$\sum_{i=1}^n [h(X_i - D_i)^+ + p(D_i - X_i)^+]$$

where h and p are respectively the unit costs of storage and shortage. The constraints are

$$X_1 \le X_2 \le \cdots \le X_n.$$

Information Structures. Among the information structures that may be appropriate in this problem are the following.

(a) Complete Information. Let $Z_i = (D_1, \ldots, D_{i-1}) = (Z_{i-1}, D_{i-1})$, i.e., decisions are made sequentially with the ordering decision in week i being based on all previously observed demands.

(b) Delayed Reporting. Let $Z_i = (D_1, \ldots, D_{i-1-l})$, i.e., ordering decisions in week i are based on all previously reported demands where there is an l-week delay in reporting.

(c) Forecasting. Let $Z_i = (D_1, \ldots, D_{i-1+l})$, i.e., ordering decisions are based on all previously observed demands as well as perfect forecasts of the demands in the next l weeks.

(d) Reporting Errors. Let $Z_i = (D_1, \ldots, D_{i-1}) + (\epsilon_1, \ldots, \epsilon_{i-1})$, i.e., ordering decisions in each period are based on demands observed plus a random error in reporting.

(e) Current Cumulative Demand. Let $Z_i = D_{i-1}$, i.e., ordering decisions in week i depend only on cumulative observed demand to date. Although this assumption is a natural one and reduces the amount of information needed to make decisions, rules of this type are not generally optimal—even when demands are independent in each week. For in the latter case, the state of the system at the beginning of week i is the difference between cumulative orders and demands through week $i-1$ and so depends on demands in all prior weeks through the cumulative orders.

Example 3. Transportation Problem with Uncertain Demand. Consider the stochastic version of the transportation problem in which there are s_i aircraft available at city $i = 1, \ldots, n$ one evening, there is a random demand D_j at city $j = 1, \ldots, m$ for aircraft the following day, and the goal is to minimize the expected costs of *deadheading*, i.e., sending empty aircraft from one city to another, and lost revenue. There is a cost c_{ij} of sending an empty aircraft from city i to city j during the night. The aircraft scheduler must decide the number x_{ij} of empty aircraft to send from city i to city j before observing the demands at the various cities. Denote by $y_j(D_j)$ the difference between the number of aircraft needed at city j and the number sent there. The expected revenue lost from each aircraft needed at city j that is not available there is r_j. Then the cost of deadheading and lost revenues is

$$\sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} x_{ij} + \sum_{j=1}^{m} r_j y_j(D_j)^+.$$

The constraints are

$$\begin{aligned} \sum_{j=1}^{m} x_{ij} &\leq s_i\,, \ i = 1, \ldots, n. \\ \sum_{i=1}^{n} x_{ij} + y_j(D_j) &= D_j,\ j = 1, \ldots, m \\ x_{ij} &\geq 0, \text{ all } i, j. \end{aligned}$$

Information Structure. In this case, the decision functions x_{ij} and y_j are based respectively on no information, i.e., $Z_{ij} = \emptyset$, and the number of aircraft $Z_j = D_j$ needed at city j.

Example 4. Capacity Planning with Uncertain Demand. A firm seeks a minimum-expected-cost policy for leasing and renting autos to meet independently distributed employee demands $D_1, \ldots, D_n$ therefor in weeks $1, \ldots, n$. There is a unit cost l_{ij} of leasing an auto at the

beginning of week i for use during weeks $i,\dots,j$. When the firm does not have a leased auto available for an employee who needs one in a week, the firm rents one at a cost $r>0$. Decisions on leases that begin in week i are made after observing demands in prior weeks, but before observing the demand in week i. But rental decisions for week i are made after also observing the demand in that week. Let x_{ij} be the number of autos leased at the beginning of week i for use during weeks $i,\dots,j$, $1\le i\le j\le n$. Let $y_i(D_i)$ be the difference between the numbers of autos demanded and leased in week i, $1\le i\le n$. The total cost of leasing and renting is

$$\sum_{i\le j} l_{ij}x_{ij}+\sum_i r_j y_j(D_j)^+.$$

The constraints are

$$\sum_{i\le j\le k} x_{ik}+y_j(D_j)=D_j,\ j=1,\dots,n$$

and

$$x_{ij}\ge 0,\text{ all } i,j.$$

Information Structure. In this case, the decision functions x_{ij} and y_i are based respectively on no information, i.e., $Z_{ij}=\emptyset$, and the auto demand $Z_i=D_i$ in week i. In order to see why it is not necessary to allow leasing decisions to depend on any of the demands, notice that the state of the system at the beginning of a week is the vector of numbers of autos leased in prior weeks for use in the week and thereafter. Thus since the demands are independent, the state in a week is a function only of leasing decisions in prior weeks and not the demands in prior weeks themselves. Hence the transition law for the system is deterministic and only the rental decision in a week depends on the demand in the week. Incidentally, this conclusion would not be valid if the demands were dependent, e.g., if they formed a Markov chain. In that event, leasing decisions in a week should depend on the demands in all prior weeks—even when the demands are Markovian.

2 REDUCTION OF STOCHASTIC TO ORDINARY MATHEMATICAL PROGRAMS

Since $\mathbb{Z}$ is finite, the stochastic program (1)-(3) reduces to that of choosing $X(z)\equiv(X_1(z_1),\dots,X_n(z_n))$ (where $z=(z_1,\dots,z_{n+1})\in\mathbb{Z}$ and $p(z)=P(Z=z)>0$), that maximizes

$$(1)'\qquad \sum_{z\in\mathbb{Z}} r(X(z),z)p(z)$$

subject to

$$(2)'\qquad g(X(z),z)\ge 0,\ z\in\mathbb{Z},$$

which is an ordinary mathematical program. The variables of this mathematical program are the values of the decision for each team member corresponding to each value of the random vector the member observes. The constraints $(2)'$ replicate the original constraints (2) once for each value of the random vector Z. In applications the number of redundant replications in $(2)'$ of an inequality in (2) can be significantly reduced by replicating the inequality only for values of the subvector of Z that appears in the inequality rather than also so doing for the complementary subvector of Z.

Linear and Quadratic Programs

It is of interest to examine when the above mathematical program is a linear or concave quadratic program. If $r(\,\cdot\,, z)$ and $g(\,\cdot\,, z)$ are affine functions for each $z \in \mathbb{Z}$, then $(1)'$-$(2)'$ is a linear program. If instead $r(\,\cdot\,, z)$ is quadratic and the associated form is negative semidefinite for all $z \in \mathbb{Z}$ and negative definite for some $z \in \mathbb{Z}$, and if $g(\,\cdot\,, z)$ is affine for each $z \in \mathbb{Z}$, then $(1)'$-$(2)'$ is a strictly-concave quadratic program. Thus, if it is feasible, it has a unique solution. In particular if $g \equiv \emptyset$, i.e., there are no inequality constraints $(2)'$, then the optimal team decision X^* is the unique solution to a system of linear equations obtained by setting the partial derivative of $(1)'$ with respect to each decision variable $X_i(z_i)$ equal to zero.

In order to make the above ideas more concrete, we formulate Examples 3 and 4 above as linear programs. Consider first the *Transportation Problem with Uncertain Demand* in which there are N values $d_{j1}, \ldots, d_{jN}$ of the random demand D_j and $p_{jk} \equiv P(D_j = d_{jk})$ for $k = 1, \ldots, N$ and each j. Set $y_{jk} = y_j(d_{jk})$ and write $y_{jk} = y^+_{jk} - y^-_{jk}$ as a difference of two nonnegative variables. Then the problem becomes the ordinary linear program of choosing x_{ij} and y_{jk} that minimize

$$\sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij} x_{ij} + \sum_{j=1}^{m} \sum_{k=1}^{N} r_j p_{jk} y^+_{jk}$$

subject to

$$\begin{aligned} \sum_{j=1}^{m} x_{ij} \qquad\qquad &\leq s_i \ , i = 1, \ldots, n \\ \sum_{i=1}^{n} x_{ij} + y^+_{jk} - y^-_{jk} &= d_{jk}, \, j = 1, \ldots, m \text{ and } k = 1, \ldots, N, \\ x_{ij},\, y^+_{jk},\, y^-_{jk} &\geq 0, \text{ all } i, j, k. \end{aligned}$$

Consider now the problem of *Capacity-Planning-with-Uncertain-Demand* in which there are at most N values $d_{j1}, \ldots, d_{jN}$ of the demand D_j and $p_{jk} \equiv P(D_j = d_{jk})$ for $k = \ 1, \ldots, N$ and each j. Set $y_{jk} = y_j(d_{jk})$ and write $y_{jk} = y^+_{jk} - y^-_{jk}$ as a difference of two nonnegative variables. Then the problem becomes the ordinary linear program of choosing x_{ij} and y_{jk} that minimize

$$\sum_{i\le j} l_{ij}x_{ij} + \sum_{i,k} rp_{ik}y^+_{ik}.$$

subject to

$$\sum_{i\le j\le k} x_{ik} + y^+_{jk} - y^-_{jk} = d_{jk},\ j = 1,\dots,n \text{ and } k = 1,\dots,N,$$

and

$$x_{ij},\ y^+_{jk},\ y^-_{jk} \ge 0, \text{ all } i,j,k.$$

Computations

It is useful to examine the size of the mathematical program that results from the above approach to stochastic programming. To that end, let $N_j \equiv |\mathbb{Z}_j|$ and let $N \equiv \max_j N_j$ be the maximum number of values of any Z_j. Also, let S_i be the set of indices j of nonconstant random variables Z_j that appear in the i^{th} constraint in (2). Let $M \equiv \max_i |S_i|$ be the maximum number of nonconstant Z_j that appear in an inequality. Then the mathematical program may have up to $\sum_{i=1}^n N_i \le nN$ variables and up to $\sum_{i=1}^m \prod_{j\in S_i} N_j \le mN^M$ inequality constraints. This problem is generally tractable only if N is not too large and if either M is very small or $m = 0$. These two conditions are often fulfilled in a problem that either has a very small number of time periods, e.g., say three or less, or has a deterministic transition law. The *Transportation Problem with Uncertain Demand* (Example 3) is an example of the former type and involves $m(n + 2N)$ variables and $n + mN$ linear equations or inequalities beyond nonnegativity. The problem of *Capacity Planning with Uncertain Demand* (Example 4) is an example of the latter type and involves up to $\frac{1}{2}n^2 + 2nN$ variables and nN linear equations as well as nonnegativity of the variables. In both cases N is generally not too large and $M = 1$. As a third example, suppose $m = 0$ and the objective is quadratic and strictly concave. Then one simply sets the partial derivatives with respect to each of the nN variables equal to zero and solves the resulting system of nN linear equations in a like number of unknowns. This last problem is tractable if N is not too large since $m = 0$.

The stochastic programming approach is not tractable for many multiperiod stochastic sequential decision problems because N is large. To illustrate, consider the problem of *Inventory Control with Uncertain Demand* (Example 2) and complete information. If each D_i has $K \ge 2$ values, then $N = |\mathbb{Z}_n| = K^{n-1}$ grows exponentially in the number of periods n. Thus the number of variables, and hence inequalities, in the mathematical programming formulation of this stochastic program is usually impossibly large. This is so even if the demands in successive periods are independent.

Comparison of Dynamic and Stochastic Programming

The dynamic- and stochastic-programming approaches to solving sequential decision problems under uncertainty are generally useful for computational purposes in different circumstances.

The dynamic-programming approach is useful if the number of state-action pairs is not too large. The stochastic-programming approach is useful if the number of possible values of the information that one observes is not too large. To illustrate, the dynamic-programming approach is useful in Examples 2 where the demands in successive periods are independent and there is complete information, whereas the stochastic-programming approach to this problem is generally intractable. On the other hand, the stochastic-programing approach is useful in Example 3 where the demands for capacity in successive periods are independent, whereas the dynamic-programming approach to this problem is generally intractable.

3 QUADRATIC UNCONSTRAINED TEAM DECISION PROBLEMS [Ra61, 62], [MR72]

As discussed above, a team decision problem that is especially easy to solve is one that is strictly concave and quadratic, and unconstrained. In particular, assume that $g = \emptyset$ and

$$r(X, Z) = X^{\mathrm{T}}\delta(Z) - \frac{1}{2}X^{\mathrm{T}}QX \tag{1}$$

where $\delta = \delta(Z) = (\delta_i)$, $Q = (q_{ij})$, $q_i = (q_{i1}, \ldots, q_{in})$ and $\mathrm{E}^{Z_i}\delta_i \equiv \mathrm{E}(\delta_i \mid Z_i)$.

Theorem 1. Quadratic Unconstrained Team Decision Problems. *In a quadratic unconstrained team decision problem with symmetric positive-definite Q, the team decision X^* is optimal if and only if*

$$\mathrm{E}^{Z_i}(\delta_i - q_i X^*) = 0,\ i = 1, \ldots, n. \tag{2}$$

Also there is a unique team decision X^* satisfying (2).

Proof. The function $r(\,\cdot\,, z)$ is strictly concave for each fixed $z \in \mathbb{Z}$. Therefore it follows that $\sum_{z\in\mathbb{Z}} r(X(z), z)p(z)$ is strictly concave in $\{X_i(z_i)\colon z_i \in \mathbb{Z}_i \text{ for all } i\}$ and quadratic. Thus the $X_i^*(z_i)$ are the unique solution to

$$\frac{\partial}{\partial X_i(z_i)} R(X^*) = 0,\ z_i \in \mathbb{Z}_i,\ i = 1, \ldots, n.$$

Now use (1) to rewrite this system as

$$\begin{aligned} 0 = \frac{\partial}{\partial X_i(z_i)} R(X^*) &= \frac{\partial}{\partial X_i(z_i)} \mathrm{E}[\mathrm{E}^{Z_i} r(X^*(Z), Z)] \\ &= \frac{\partial}{\partial X_i(z_i)} [\mathrm{E}^{Z_i = z_i} r(X^*(Z), Z)] P(Z_i = z_i) \\ &= \mathrm{E}^{Z_i = z_i}[\delta_i(Z) - q_i X^*(Z)] P(Z_i = z_i), \end{aligned}$$

so because $P(Z_i = z_i) > 0$, (2) holds as claimed. ∎

Theorem 1 requires that a system of up to nN linear equations in a like number of variables be solved and so is not tractable if N is large. In the special case in which δ_i is affine in Z_i for each i and Z has a joint normal distribution, a homework problem shows that the X_i are affine in the Z_i. The coefficients can be found by solving two systems of linear equations, one $n \times n$ and the other $m \times m$ where m is the number of random variables among $Z_1, \ldots, Z_n$.

4 SEQUENTIAL QUADRATIC UNCONSTRAINED TEAMS: CERTAINTY EQUIVALENTS

[HMSM60], [MR72]

On the other hand, even without the above assumptions that δ_i is affine in Z_i for each i and Z has a joint normal distribution, but with the added assumption that the problem is "sequential", we now show how to reduce the number of linear equations that need to be solved in such problems to about two systems of n linear equations in a like number of variables—a dramatic reduction indeed.

In order to motivate the idea, consider the following example.

Example 5. Single-Person Quadratic Unconstrained Team. Suppose $n = 1$, Z_1 is the empty vector, Z_2 is real valued, X does not depend on Z, and $r(X, Z) = -\mathrm{E}(Z_2 - X)^2$. Then

$$\min_X \mathrm{E}(Z_2 - X)^2 = \mathrm{E}[(Z_2 - \mathrm{E}Z_2) + (\mathrm{E}Z_2 - X^*)]^2 = \mathrm{E}(Z_2 - \mathrm{E}Z_2)^2 + (\mathrm{E}Z_2 - X^*)^2 = \mathrm{Var}Z_2$$

where $X^* = \mathrm{E}Z_2$. Since X^* depends only on the expected value of Z_2, we would have found the same solution if we had initially replaced Z_2 by $\mathrm{E}Z_2$ and solved the "equivalent" deterministic problem $\min_X (\mathrm{E}Z_2 - X)^2$, which equals 0, so again $X^* = \mathrm{E}Z_2$. Thus $\mathrm{E}Z_2$ is a "certainty equivalent" for Z_2 with respect to the problem of finding X^*.

If U and V are finite-valued random vectors, we say that U *determines* V if $V = f(U)$ for some function f. In that event, for any finite-valued random variable W, $\mathrm{E}^V\mathrm{E}^U W = \mathrm{E}^V\mathrm{E}^{U,V}W = \mathrm{E}^V W$. Also, if U determines V and V determines W, then U determines W, i.e., the determines relation is transitive.

Call the team decision problem *sequential* if Z_{i+1} determines Z_i for $i = 1, \ldots, n-1$. The next result is important because it allows one *to solve sequential quadratic unconstrained team decision problems by replacing all random variables by their conditional expectations given the then current state of information and solving the equivalent deterministic problem.*

Theorem 2. Certainty Equivalents for Sequential Quadratic Unconstrained Team Decision Problems. *In a sequential quadratic unconstrained team decision problem with symmetric positive-definite Q, there is a unique optimal team decision X^* that satisfies*

$$\mathrm{E}^{Z_i}(\delta_j - q_j X^*) = 0,\ 1 \le i \le j \le n. \tag{3}$$

Moreover, on setting $H = \{1, \ldots, i-1\}$ and $F = \{i, \ldots, n\}$ for given i,

$$\mathrm{E}^{Z_i} X^*_F = Q^{-1}_{FF}[\mathrm{E}^{Z_i}\delta_F - Q_{FH}X^*_H]. \tag{4}$$

Proof. Recall from Theorem 1 that

$$\mathrm{E}^{Z_j}(\delta_j - q_j X^*) = 0$$

so because Z_j determines Z_i for $i \leq j$,

$$0 = \mathrm{E}^{Z_i}\mathrm{E}^{Z_j}(\delta_j - q_j X^*) = \mathrm{E}^{Z_i}(\delta_j - q_j X^*).$$

For fixed i, these equations can be rewritten as

$$\mathrm{E}^{Z_i}(\delta_F - Q_{FH} X_H^* - Q_{FF} X_F^*) = 0.$$

Now premultiply this equation by Q_{FF}^{-1} and use the fact that $\mathrm{E}^{Z_i} X_H^* = X_H^*$. ∎

Interpretation of Solution

Remark 1. Solution as a Sequence of Deterministic Problems. The formula (4) permits $X_i^*(Z_i) = \mathrm{E}^{Z_i} X_i^*$ to be computed once the variables $X_1^*(Z_1), \ldots, X_{i-1}^*(Z_{i-1})$ have been found. Moreover, since $X_i^*(Z_i)$ depends on δ_F only through its conditional expectation $\mathrm{E}^{Z_i}\delta_F$ given the then current information Z_i, it follows that $X_i^*(Z_i)$ has the same value that it would have for the corresponding deterministic problem in which the random vector δ_F is replaced by its conditional expectation $\mathrm{E}^{Z_i}\delta_F$!

Remark 2. On-Line Implementation or Wait-and-See Solution. It is not necessary to tabulate $X_i^*(\,\cdot\,)$ in order to choose $X_i^*(Z_i)$ optimally. All that is required is to wait until Z_i is observed —in which case $Z_1, \ldots, Z_{i-1}$ are determined—and then choose $X_i^*(Z_i)$ as in Remark 1 above.

Remark 3. Linearity of Optimal Decision Function. Observe from (4) that the optimal decision function $X_i^*(Z_i)$ for person i is linear in the *history* X_H^* of decisions of prior persons and the conditional expected value $\mathrm{E}^{Z_i}\delta_F$ of the *future* random vector δ_F (yet to be observed except for $\delta_i(Z_i)$) given the state of information Z_i of person i at the time he chooses his decision.

Computations

If we use Remark 2 above, then except for the work in computing the needed conditional expected values $\mathrm{E}^{Z_i}\delta_F$, one for each i, the work in solving (4) is independent of N_j for all j . If one first finds the Cholesky factorization of the matrix Q, i.e., the upper triangular matrix R for which $Q = R^{\mathrm{T}}R$, then $\frac{n^3}{2}$ operations suffice to find all optimal decisions.

Quadratic Control Problem [Be87], [Ku71]

It is often the case in practice that sequential quadratic team decision problems discussed in Theorem 2 have additional structure that leads to computational simplifications. To illustrate these ideas, consider the following *quadratic control* problem. Choose *controls* $u_1, \ldots, u_T$ and, given an *initial state vector* x_1, *state vectors* $x_2, \ldots, x_T$ that minimize the expected value of the quadratic form

$$\sum_{t=1}^{T-1}(x_t^{\mathrm{T}} Q_t x_t + u_t^{\mathrm{T}} R_t u_t) + x_T^{\mathrm{T}} Q_T x_T \tag{5}$$

subject to the linear dynamical equations

$$x_{t+1} = A_t x_t + B_t u_t + \alpha_t,\ t = 1, \ldots, T-1 \tag{6}$$

where the x_t and α_t are column vectors of common dimension, the u_t are column vectors of common dimension, the matrices Q_t are symmetric and positive semidefinite, the matrices R_t are symmetric and positive definite, the matrices A_t and B_t have appropriate dimension, the α_t are random errors whose distributions are independent of the controls applied, and $(\alpha_1, \ldots, \alpha_{t-1})$ is the information available to choose the control u_t for each t. It is natural in most applications to think of t as one of T periods, e.g., days, weeks, months, etc., and we shall use this terminology in what follows. Problems of this type arise in a variety of settings of which we mention two.

Example 6. Rocket Control. Consider the problem of moving a rocket efficiently from its initial position x_1 to a terminal position that is close to a target position at time T. Let x_t be the position of the rocket at time t and u_t be the "control" exerted at that time. The position of a rocket might be a vector including at least its six location and velocity coordinates. The desired target location and velocity vector at time T is the null vector (the distance and velocity should be zero at time T). Assume that the motion of the rocket satisfies the linear dynamical equations (6). The objective function might have $Q_t = 0$ for all $t < T$. Then the objective function would entail minimizing the sum of two terms, the first representing the sum of the costs of the controls at the first $T-1$ times and the second $x_T^{\mathrm{T}} Q_T x_T$ the cost of missing the target position at time T. Each cost reflects the fact that small quantities have small costs and large quantities have large costs.

Example 7. Multiproduct Supply Management. Consider the problem of managing inventories of several products so as to minimize the sum of the expected resource and storage/shortage costs in the presence of uncertain demands for the products. In this setting, x_t would be the possibly negative vector of inventories of the products in period t, u_t the possibly negative vector of several resources (perhaps labor, materials, capital, etc.) applied to production in period t, and

α_t the vector of demands for the products in period t. Negative inventories represent backorders and negative resource consumption represents returns of resources. The matrix A_t in period t would be the identity matrix if stocks of products are merely held in storage without transformation. That matrix might differ from the identity matrix if stocks move from one state to another over time, e.g., reflecting age, deterioration, location, etc. The ij^{th} element of the matrix B_t would be the rate of production of product i in period t per unit of resource j applied in that period. The costs reflect the desirability of low inventories and low resource costs.

Reduction to Unconstrained Problem. It might seem that this problem is more general than the unconstrained sequential decision problem considered to date because of the linear equality constraints (6). But that is not the case because one can use (6) to eliminate the x_t leaving (5) depending only on the u_t. Then Theorem 3 applies at once. As a consequence, in order to determine the optimal choice of u_1, it suffices to replace the random vectors α_t by their conditional expectations given the state of information at the time the vector u_1 is chosen. For this reason we can and do assume without loss of generality that the α_t are constant vectors.

In fact it is possible to eliminate the constant vectors entirely. This may be accomplished by appending an additional state variable y_t in each period t and appending to (6) the equations

$$y_{t+1} = y_t,\ t = 1, \ldots, T \tag{7}$$

where $y_1 \equiv 1$. This assures that $y_t = 1$ for all t, so we can replace α_t in (6) by the product $\alpha_t y_t$. The *augmented system* with one additional state variable and the same control variables then has the desired form with zero random errors. If we denote the various quantities in the augmented system with overbars, those quantities have the following relations to the corresponding quantities of the original system: $\overline{\alpha}_t = 0$, $\overline{u}_t = u_t$, $\overline{R}_t = R_t$,

$$\overline{Q}_t = \begin{pmatrix} Q_t & 0 \\ 0 & 0 \end{pmatrix},\ \overline{A}_t = \begin{pmatrix} A_t & \alpha_t \\ 0 & 1 \end{pmatrix},\ \overline{B}_t = \begin{pmatrix} B_t \\ 0 \end{pmatrix} \text{ and } \overline{x}_t = \begin{pmatrix} x_t \\ y_t \end{pmatrix}, \tag{8}$$

where in each case the second matrix in a column (resp., row) partition has only one column (resp., row).

Dynamic-Programming Solution with Zero Random Errors. It is natural to solve the zero-random-error problem by dynamic programming. This approach is tractable because even though the state vector x_t in period t may have many components, the minimum cost is a positive semidefinite quadratic form that can be calculated explicitly. To see this, let $\mathcal{C}_t(x)$ be the minimum cost in periods $t, \ldots, T$ given x is the state in period t, and let $u_t(x)$ be the corresponding optimal control in that state and period. Evidently

$$\mathcal{C}_t(x) = \min_u [x^{\mathrm{T}} Q_t x + u^{\mathrm{T}} R_t u + \mathcal{C}_{t+1}(A_t x + B_t u)],\ t = 1, \ldots, T-1 \tag{9}$$

and $\mathcal{C}_T(x) = x^{\mathrm{T}} Q_T x$ for all x.

Theorem 3. Optimal Quadratic Control. *If the quadratic control problem* (5)-(6) *has zero random errors, each Q_t is symmetric and positive semidefinite, and each R_t is symmetric and positive definite, then the minimum cost $\mathcal{C}_t$ and optimal control u_t are*

$$\mathcal{C}_t(x) = x^T K_{t-1} x,\ t = 1, \ldots, T, \tag{10}$$

$$u_t(x) = -S_t^{-1} U_t A_t x,\ t = 1, \ldots, T-1 \tag{11}$$

with symmetric positive semidefinite matrices K_t given from the Riccati recursion

$$K_{t-1} = Q_t + A_t^T W_t A_t,\ t = 1, \ldots, T-1 \tag{12}$$

and $K_{T-1} = Q_T$ where $W_t \equiv K_t - U_t^T S_t^{-1} U_t$, $S_t \equiv R_t + B_t^T K_t B_t$ and $U_t \equiv B_t^T K_t$ for $t = 1, \ldots, T-1$.

Proof. The proof of (10)-(11) is by induction on t. This claim is trivially true for $t = T$. Suppose it holds for $1 < t+1 \le T$ and consider t. Then S_t is symmetric and positive definite since that is so of R_t and since K_t is symmetric and positive semidefinite from the induction hypothesis. Thus S_t is nonsingular, whence by (10) for $t+1$ and completing the square, the expression in brackets on the right-hand side of (9) is

$$x^{\mathrm{T}} Q_t x + u^{\mathrm{T}} R_t u + \mathcal{C}_{t+1}(A_t x + B_t u) = x^{\mathrm{T}}(Q_t + A_t^{\mathrm{T}} K_t A_t) x + (u^{\mathrm{T}} S_t u + 2u^{\mathrm{T}} U_t A_t x)$$

$$= x^{\mathrm{T}}(Q_t + A_t^{\mathrm{T}} W_t A_t) x + (u + S_t^{-1} U_t A_t x)^{\mathrm{T}} S_t (u + S_t^{-1} U_t A_t x).$$

Now since S_t is symmetric and positive definite, $u = -S_t^{-1} U_t A_t x$ is the unique minimizer of the above function, so (10) and (11) follow from (9). Furthermore, it is immediate from (12) and the induction hypothesis that K_{t-1} is symmetric. Finally, K_{t-1} is positive semidefinite because $\mathcal{C}_t$ is nonnegative. The last follows from the fact that the expression in brackets on the right-hand side of (9) is nonnegative for all x because Q_t, R_t and K_t are symmetric and positive semidefinite. ∎

Remark 1. Linearity of Optimal Control. Observe from (11) that the optimal control u_t in period t is linear in the state vector. Thus when there is a random error, u_t is linear in the state vector in period t and the conditional expected value of the random error in that period given the random errors in prior periods.

Remark 2. Linearity of Computational Effort in T. Observe that the matrices K_t can be calculated recursively from (12). Thus the computational effort to determine those matrices and the optimal controls is linear in T.

Remark 3. Quadratic Plus Linear Costs. It is also possible to apply the above results to the case in which a linear function of the state and control vectors is added to the objective function (5). To do this, complete the square, thereby expressing the generalization of (5), apart from a constant, as

$$(5)' \qquad \sum_{t=1}^{T-1}[(x_t-\xi_t)^{\mathrm{T}}Q_t(x_t-\xi_t)+(u_t-\upsilon_t)^{\mathrm{T}}R_t(u_t-\upsilon_t)]+(x_T-\xi_T)^{\mathrm{T}}Q_T(x_T-\xi_T)$$

for some translation vectors ξ_t and υ_t for each t. Then substitute $\overline{x}_t \equiv x_t-\xi_t$, $\overline{u}_t \equiv u_t-\upsilon_t$ and $\overline{\alpha}_t \equiv A_t\xi_t+B_t\upsilon_t+\alpha_t-\xi_{t+1}$ in both (5)′ and (6) for each t. The resulting problem assumes the form (5)-(6) with $\overline{x}_t$, $\overline{u}_t$ and $\overline{\alpha}_t$ replacing x_t, u_t and α_t respectively.

Solution with Independent Random Errors. If the random errors in different periods are independent, then the conditional expectations of the random errors in period t or later given the random errors in periods prior to t are independent of the latter. Consequently the optimal control functions for the corresponding augmented zero-random-error system are optimal for the original system with independent random errors.

Solution with Dependent Random Errors. If, however, the random errors in different periods are dependent, then the optimal control function in period $t>1$ determined from the corresponding augmented zero-random-error system is not generally optimal for the original system (with random errors) in period t. The reason is that the augmented matrix $\overline{A}_t$ depends on the conditional expected error α_t, whose value depends on the values of the conditioning variables. Consequently, the matrix $\overline{K}_t$ depends on $\alpha_{tT} \equiv (\alpha_t,\dots,\alpha_{T-1})$. To examine this dependence, partition $\overline{K}_t$ like (8) in the form

$$\overline{K}_t = \begin{pmatrix} K_t & k_t \\ k_t^{\mathrm{T}} & \kappa_t \end{pmatrix}.$$

Then a straightforward, but tedious, computation using (12) for the corresponding augmented system shows by induction on t that the matrices K_t are independent of α_{tT}, the vectors k_t satisfy the recursion ($k_{T-1} \equiv 0$)

$$(13) \qquad k_{t-1} = A_t^{\mathrm{T}}[W_t\alpha_t+(I-U_t^{\mathrm{T}}S_t^{-1}B_t^{\mathrm{T}})k_t],\ t=1,\dots,T-1,$$

and the optimal control $u_t(x,\alpha_{tT})$ takes the form

$$u_t(x, \alpha_{tT}) = S_t^{-1}[U_t(A_t x + \alpha_t) + B_t^{\mathrm{T}} k_t],\ t = 1, \ldots, T-1, \tag{14}$$

Indeed, the matrices K_t coincide with the corresponding matrices satisfying the Riccati recursion for the unaugmented zero-random-error problem. Also, the matrices defining the coefficients of x, α_t and k_t in (13) and (6) are all independent of α_{tT}. Consequently, to compute the optimal controls u_t from (6) when α_{tT} is altered because of new information, simply recompute the k_t using the recursion (13). Incidentally, it also follows by iterating (13) that $k_t = k_t(\alpha_{t+1,T})$ is linear in $\alpha_{t+1,T}$. Thus, in view of (6), $u_t(x, \alpha_{tT})$ is linear in x and α_{tT}, confirming Remark 3 following Theorem 2 in the present instance.

Strengths and Weaknesses

The Certainty-Equivalence Theorem 2 permits sequential decision problems under uncertainty with thousands of variables and constraints to be solved nearly as easily as corresponding deterministic problems provided that the objective function is convex and quadratic, and the constraints consist only of linear equations—not inequalities. For example, a multiproduct multiperiod inventory problem with thousands of products that are tied together by joint convex quadratic production and storage/shortage costs, uncertain demand for each product, and no inequality constraints can be solved easily by the Certainty-Equivalence Theorem and its implementation as a quadratic control problem discussed above.

Because of the Certainty-Equivalence Theorem, it is tempting to use its conclusion, viz., replacing all random variables by their expectations and solving the resulting deterministic problem, in circumstances where it cannot be shown to lead to optimal solutions. The reason for so doing is the hope that the error from so doing will not be great. Indeed, that general approach is used implicitly in practice for many large sequential decision problems under uncertainty that could not otherwise be solved.

Nevertheless, it must be recognized that the assumption that the costs are convex and quadratic may not be a reasonable approximation in practice. And the absence of inequality constraints does not, for example, permit one to require production in each period to be nonnegative, as is usually the case in practice. Thus, if the optimal production schedule is not nonnegative, the optimal solution may not be meaningful. On the other hand, in practice one can usually modify the optimal solution in reasonable ways to overcome these limitations—though not without losing optimality.

Though the certainty-equivalence approach has limitations, it is one of very few approaches that is useful for finding optimal policies for large sequential decision problems under uncertainty. For this reason, it is worthy of serious consideration in many circumstances.

This page is intentionally left blank.

3

Continuous-Time-Parameter Markov Population Decision Processes

1 FINITE CHAINS: FORMULATION

Consider a finite population of individuals that is observed continuously and indefinitely beginning at time zero. At each time, each individual will be in some *state* in a finite set $\mathcal{S}$ of $S < \infty$ states. Each time $t \geq 0$ an individual is observed in state s, an *action* a is chosen from a finite set A_s of possible actions and a *reward rate* $r(s, a)$ is earned per unit time. Also the individual generates a finite expected number $q(u|s, a)$ of additional individuals in state u per unit time. Assume that $q(u|s, a) \geq 0$ for $u \neq s$, since otherwise positive populations may produce negative ones. Call $q(u|s, a)$ the *transition rate* from state s to state u when using action a.

In the important case where the population consists of exactly (resp., at most) one individual at each point in time, so $\sum_{u \in \mathcal{S}} q(u|s, a) = 0$ (resp., ≤ 0) for all $a \in A_s$ and $s \in \mathcal{S}$, call the system *stochastic* (resp., *substochastic*). In either case $q(s|s, a) \leq 0$ for all $a \in A_s$ and $s \in \mathcal{S}$. The condition $\sum_{u \in \mathcal{S}} q(u|s, a) = 0$ has the interpretation that the expected rate of addition of individuals to the system generated by each individual is zero, so the system population size remains unchanged. Similarly, in the substochastic case the system population size remains unchanged or falls.

2 MAXIMUM T-PERIOD VALUE: BELLMAN'S EQUATION AND EXAMPLES [Be57]

A fundamental problem in the above setting is to find a policy that maximizes the T-period expected reward with terminal value vector $v = (v_s)$ in the various states. We begin with an informal analysis and make it rigorous later. Let V_s^t be the maximum expected reward that an individual (and his progeny) in state s at time t earns during the interval $[t, T]$. Then it is plausible that for small increments of time $h > 0$,

$$V_s^{t-h} = \max_{a \in A_s} [r(s,a)h + \sum_{u \in \mathcal{S}} q(u|s,a)hV_u^t + V_s^t + o(h)]$$

where $o(h)$ is a function for which $\lim_{h \downarrow 0} h^{-1}o(h) = 0$. Subtracting V_s^t from both sides, dividing by h, and letting $h \downarrow 0$ yields *Bellman's equation*

$$-\dot{V}_s^t = \max_{a \in A_s} [r(s,a) + \sum_{u \in \mathcal{S}} q(u|s,a)V_u^t] \tag{1}$$

for $t \geq 0$ and $s \in \mathcal{S}$ together with the terminal condition $V_s^T \equiv v_s$ for $s \in \mathcal{S}$. (Of course $\dot{V}_s^t \equiv \frac{d}{dt}V_s^t$.)

The expression in brackets on the right-hand side of (1) is the sum of two terms. The first is the reward rate $r(s,a)$ that an individual in state s earns while taking action a in that state at time t. The second is the extra reward rate during the interval $[t, T]$ that results from changes in the population size in each state at time t when an individual in state s takes action a in that state at time t and all individuals use an optimal policy thereafter. To see why this is so, observe that $q(u|s,a)V_u^t$ is the product of the expected rate $q(u|s,a)$ of addition of individuals to state u when an individual in state s takes action a in that state at time t and the maximum expected reward V_u^t that each such individual in state u and his progeny earn in the interval $[t, T]$. Thus, the second term $\sum_{u \in \mathcal{S}} q(u|s,a)V_u^t$ is the total extra expected reward rate during the interval $[t, T]$ that results from addition of individuals to the several states when an individual in state s takes action a in that state at time t. Finally, the right-hand side of (1) is the maximum expected reward rate that an individual and his progeny earn in state s at time t, and equals the left-hand side of (1), viz., the negative of the time derivative of the maximum value in state s at time t.

Observe that (1) is a system of S ordinary nonlinear differential equations. The sequel develops this equation rigorously, shows that it has a unique solution, and establishes the existence of a maximum T-period-value piecewise-constant (in time) policy.

Sometimes in the sequel, the symbol V_s^t will instead mean the maximum t-period value. (The reader will be warned when this is so.) In that event, the minus sign on the left-hand side of (1) disappears and the terminal condition becomes instead $V_s^0 \equiv v_s$ for $s \in \mathcal{S}$.

Example 1. Controlled Queues. Consider controlling a single-channel queue with independent Poisson arrivals and exponential service times. The state of the system is the number of cus-

tomers in the queue (including the one being served), and this number cannot exceed the queue capacity S. Let $\lambda_s \geq 0$ be the arrival rate of customers into the system when s customers are in the queue. Of course, $\lambda_S \equiv 0$. The actions are the indices $1, \ldots, K_s$ of the K_s available service rates $\mu_{s1}, \ldots, \mu_{sK_s} \geq 0$, say, when s customers are in the system. Of course $K_0 = 1$ and $\mu_{01} = 0$. Then for $0 \leq s, u \leq S$, the transition rates are

$$q(u|s,a) = \begin{cases} \mu_{sa} & , u = s-1 \\ -(\mu_{sa} + \lambda_s), & u = s \\ \lambda_s & , u = s+1 \end{cases}$$

and $q(u|s,a) = 0$ otherwise. Also $r(s,a) = r\mu_{sa} - c_{sa}$ where r is the price each customer pays on completing service and c_{sa} is the service cost rate when the action a is chosen with s customers in the queue. In this event Bellman's equation becomes

$$-\dot{V}_s^t = \max_{1 \leq a \leq K_s} [r\mu_{sa} - c_{sa} + \mu_{sa}(V^t_{(s-1)^+} - V^t_s) + \lambda_s(V^t_{(s+1)\wedge S} - V^t_s)], \ 0 \leq s \leq S.$$

Example 2. Supply Management. A supply manager seeks an ordering policy with minimum expected cost. The states are the possible inventory levels $0, \ldots, S$. An order for one or more items can be placed or canceled at any time with the time to deliver being exponentially distributed with mean μ^{-1}, $0 < \mu < \infty$. The actions in state s are the amounts a of stock on hand plus on order, so $s \leq a \leq S$. The times between demands are independent exponentially distributed random variables with common mean λ^{-1}, $0 < \lambda < \infty$. The demands are independent and identically distributed with d_k being the probability of a demand of size k. Unsatisfied demands are lost. The transition rates are

$$q(u|s,a) = \begin{cases} \lambda d_{s-u} & , 0 < u < s \\ \lambda \sum_{k \geq s} d_k, & u = 0 \\ \mu & , u = a > s \\ -(\lambda + \mu), & u = s \end{cases}$$

and $q(u|s,a) = 0$ otherwise. Also, the expected-cost-rate function is $c(s,a) = c(a-s)\mu + h(s) + \sum_{w=1}^{\infty} p(w)\lambda d_{s+w}$ where $c(\,\cdot\,)$ and $p(\,\cdot\,)$ are the ordering and shortage cost functions, $h(\,\cdot\,)$ is the storage-cost rate, and $c(0) = h(0) = p(0) = 0$.

Example 3. Project Scheduling. A project manager wants to allocate funds continuously among the activities of a project to maximize the probability of completing the project by time T. The project consists of a partially-ordered collection $\mathcal{A}$ of activities. An activity cannot be initiated until all activities preceding it in the partial order are completed. At any given point of time, $\mathcal{A}$ can be partitioned into two sets, viz., the set s of activities that have been completed

and the set s^c (the complement of s) of activities that have not been completed. From what we have said above, the set s is necessarily *decreasing*, i.e., if an activity is in s, then so is every activity that precedes it. Similarly, s^c is *increasing*, i.e., if an activity is in s^c, then so is every activity that follows it in the partial order. Denote by $\mathcal{S}$ the set of decreasing subsets of $\mathcal{A}$. For each $s \in \mathcal{S}$, denote by $\mathbb{A}_s$ the set of activities in s^c that are not preceded by any other activity in s^c. Evidently, $\mathbb{A}_s$ is the set of activities on which effort may be expended when $s \in \mathcal{S}$ is the set of completed activities. Funds may be spent continuously on the project at the rate of b million dollars per week. At each point in time at which $s \in \mathcal{S}$ is the set of completed activities, the funds can be allocated among the activities in $\mathbb{A}_s$ in multiples of one million dollars. An action a in state $s \in \mathcal{S}$ is an $|\mathbb{A}_s|$-tuple of nonnegative integers $a = (a_\alpha)$ indexed by elements α of $\mathbb{A}_s$ for which $\sum_{\alpha \in \mathbb{A}_s} a_\alpha = b$. The interpretation of a is that a_α million dollars per week are allocated to activity $\alpha \in \mathbb{A}_s$ for each such α. Denote by A_s the set of all such actions if $s \neq \mathcal{A}$ and the empty action $\emptyset$ if $s = \mathcal{A}$. The times to complete the activities are independent exponentially distributed random variables. The completion rate of activity $\alpha \in \mathcal{A}$ when $m \geq 0$ million dollars per month are allocated to that activity is $\lambda_\alpha \sqrt{m}$ with $\lambda_\alpha > 0$. Let V_s^t be the maximum probability of completing the project by time T when $s \in \mathcal{S}$ is the subset of activities completed by time $0 \leq t \leq T$. Then $v_{\mathcal{A}} = 1$ and $v_s = 0$ otherwise. Also, $r(s, a) = 0$ for all $a \in A_s$ and $s \in \mathcal{S}$. Moreover, in state $\mathcal{A}$, i.e., when the project is completed, $q(s|\mathcal{A}, \emptyset) = 0$ for all $s \in \mathcal{S}$. Finally, for each $s \in \mathcal{S} \setminus \{\mathcal{A}\}$ and $a = (a_\alpha) \in A_s$, the nonzero transition rates are contained among those for which $q(s \cup \{\alpha\}|s, a) = \lambda_\alpha \sqrt{a_\alpha}$ for some $\alpha \in \mathbb{A}_s$ or $q(s|s, a) = -\sum_{\alpha \in \mathbb{A}_s} \lambda_\alpha \sqrt{a_\alpha}$.

3 PIECEWISE-CONSTANT POLICIES, GENERATORS, TRANSITION MATRICES

A *decision* is a function δ that assigns to each state s in $\mathcal{S}$ an action $\delta^s \in A_s$. Let $\Delta \equiv \times_{s \in \mathcal{S}} A_s$ be the set of all decisions. A *policy* is a mapping $\pi = (\delta_t) : [0, \infty) \to \Delta$ $(t \to \delta_t)$ that is *piecewise constant*, i.e., there is a sequence of numbers $0 \equiv t_0 < t_1 < t_2 < \cdots$ with $t_N \to \infty$ as $N \to \infty$, such that $\delta_t = \delta_{t_N}$ for $t_N \leq t < t_{N+1}$ and each $0 \leq N$. Using π means that an individual in state s at time t chooses action δ_t^s at that time. Let Δ^∞ denote the set of all policies. Call a policy δ^∞ that uses δ at each point in time *stationary*.

For each $\delta \in \Delta$, let $r_\delta \equiv (r(s, \delta^s))$ be the *reward-rate vector* and $Q_\delta \equiv (q(u|s, \delta^s))$ be the *generator* matrix. Let P_δ^t be the $S \times S$ *transition matrix* whose su^{th} element is the expected number of individuals in state u at time t given that one individual was in state s at time 0 and δ^∞ is used. Assume that the P_δ^t satisfy $P_\delta^0 = I$,

$$P_\delta^{t+h} = P_\delta^t P_\delta^h = P_\delta^h P_\delta^t \text{ for } h, t \geq 0 \tag{1}$$

and

$$P_\delta^h = I + Q_\delta h + o(h) \text{ for } h > 0. \tag{2}$$

Substituting (2) into (1), subtracting P_δ^t from both sides, dividing by h, and letting $h \to 0$ establishes that ($\dot{P}_\delta^t \equiv \frac{d}{dt} P_\delta^t$)

$$P_\delta^t Q_\delta = \dot{P}_\delta^t = Q_\delta P_\delta^t \text{ for } t \geq 0. \tag{3}$$

These equations describe the growth of the population when the decision δ is used and may be interpreted as follows. Observe that Q_δ is the rate of addition to the population at any time per individual at that time. The left-hand side of (3) is the product of the population at time t and the rate of addition to the population per individual at that time. The right-hand side of (3) is the product of the rate of addition to the population at time zero and the population generated at time t by each individual added at time zero. Both products are simply different ways of expressing the rate of addition $\dot{P}_\delta^t$ to the population at time t. In any case, the unique solution of each of these two systems of S linear differential equations satisfying $P_\delta^0 = I$ is

$$P_\delta^t = e^{Q_\delta t} \text{ for } t \geq 0. \tag{4}$$

Nonnegative, Substochastic and Stochastic Transition Matrices

The nonnegativity of the off-diagonal elements of Q_δ implies that P_δ^t is nonnegative and has positive diagonal elements for $t \geq 0$. To see this, write $Q_\delta = (Q_\delta + \lambda I) - \lambda I$ where λ is a positive number chosen so that $Q_\delta + \lambda I \geq 0$. Now $(Q_\delta + \lambda I)t$ commutes with $\lambda I t$, so that $P_\delta^t = e^{Q_\delta t} = e^{(Q_\delta + \lambda I)t} e^{-\lambda I t}$. Also $e^{(Q_\delta + \lambda I)t} \geq 0$ because $(Q_\delta + \lambda I)t \geq 0$ and $e^{-\lambda I t} = e^{-\lambda t} I \geq 0$, so $P_\delta^t \geq 0$. Further, since $e^{Q_\delta t} = (e^{Q_\delta \frac{t}{n}})^n$ and $e^{Q_\delta \frac{t}{n}} = I + \frac{t}{n} Q_\delta + o(\frac{t}{n})$ is nonnegative and has positive diagonal elements for large enough n, it follows that $P_\delta^t = e^{Q_\delta t}$ also has positive diagonal elements.

If also $Q_\delta 1 = 0$ (resp., ≤ 0), then P_δ^t is stochastic (resp., substochastic). To see this, observe that

$$P_\delta^t 1 - 1 = (e^{Q_\delta t} - I)1 = \left(\int_0^t e^{Q_\delta u} du \right) Q_\delta 1,$$

so because the term in large parentheses is nonnegative, P_δ^t is stochastic (resp., substochastic) if $Q_\delta 1 = 0$ (resp., $Q_\delta 1 \leq 0$).

4 CHARACTERIZATION OF MAXIMUM T-PERIOD-VALUE POLICIES [Be57], [Mi68b]

Suppose that $\pi = (\delta_t)$ is any policy with $\delta_t = \gamma_N$ for $t_N \leq t < t_{N+1}$ for some $\{\gamma_N\}$ and numbers $0 \equiv t_0 < t_1 < t_2 < \cdots$ with $t_N \to \infty$ as $N \to \infty$. Then the transition function P_π^t determined by π is, for $t_N \leq t < t_{N+1}$, given by

$$P_\pi^t = P_{\gamma_0}^{t_1 - t_0} \cdots P_{\gamma_{N-1}}^{t_N - t_{N-1}} P_{\gamma_N}^{t - t_N}. \tag{1}$$

Thus, for $t_N < t < t_{N+1}$,

$$\dot{P}_\pi^t = P_\pi^t Q_{\delta_t}. \tag{2}$$

Interpret the su^{th} element of P_π^t as the expected number of individuals that an individual in state s at time 0 generates in state u at time t when using π. *Using* π during an interval means that π uses the decision δ_t at each time t in the interval.

Let $\pi = (\gamma_t)$ and $\pi^* = (\delta_t)$ be policies. Fix $T > 0$ and $v \in \Re^S$. Let $\pi^t\pi^*$ be the policy that uses π during $[0, t)$ and π^* during $[t, \infty)$, i.e., uses γ_u at $u \in [0, t)$ and uses δ_u at $u \in [t, \infty)$. Set $V_\pi^t = (V_{\pi s}^t)$ with $V_{\pi s}^t$ being the expected reward that π earns during the interval $[t, T]$ starting with a single individual in state s at time t.

The following Comparison Lemma gives an expression for the difference between the T-period values of two policies.

Lemma 1. Comparison. *If $\pi = (\gamma_t)$ and π^* are policies, then*

$$V_\pi^0 - V_{\pi^*}^0 = \int_0^T P_\pi^t G_{\gamma_t\pi^*}^t dt$$

where the comparison function is given by

$$G_{\gamma\pi^*}^t \equiv r_\gamma + Q_\gamma V_{\pi^*}^t + \dot{V}_{\pi^*}^t, \gamma \in \Delta \text{ and } 0 \le t \le T. \tag{3}$$

Proof. Evidently,

$$V_{\pi^t\pi^*}^0 = \int_0^t P_\pi^u r_{\gamma_u} du + P_\pi^t V_{\pi^*}^t. \tag{4}$$

Thus, for all continuity points t of π and π^*,

$$\frac{d}{dt} V_{\pi^t\pi^*}^0 = P_\pi^t[r_{\gamma_t} + Q_{\gamma_t} V_{\pi^*}^t + \dot{V}_{\pi^*}^t] = P_\pi^t G_{\gamma_t\pi^*}^t. \tag{5}$$

Hence since $V_\pi^0 - V_{\pi^*}^0 = \int_0^T (\frac{d}{dt} V_{\pi^t\pi^*}^0) dt$, the result follows from (4) and (5). ∎

The optimal-return operator plays a fundamental role in discrete-time-parameter problems. In continuous-time-parameter problems, the *optimal-return generator* $\mathcal{G}$ plays a similar role. Define $\mathcal{G} : \Re^S \to \Re^S$ by

$$\mathcal{G}V = \max_{\delta \in \Delta} [r_\delta + Q_\delta V] \text{ for } V \in \Re^S.$$

Observe that the optimal-return generator stands in the same relation to the generator matrices Q_δ that the optimal-return operator does to the transition matrices P_δ.

Theorem 2. Characterization of Maximum T-Period-Value Policies. *The policy $\pi^* = (\delta_t)$ has maximum T-period value with terminal-value vector v if and only if $V^t \equiv V_{\pi^*}^t$ is continuously differentiable in t on $[0, T]$, $V^T = v$ and*

$$-\dot{V}^t = \mathcal{G}V^t = r_{\delta_t} + Q_{\delta_t} V^t, \ 0 \le t \le T. \tag{6}$$

Proof. For the "if" part, observe that (6) implies that $G^t_{\gamma\pi^*} \le 0$ for all γ, so by the Comparison Lemma 1, $V^0_\pi \le V^0_{\pi^*}$ for all π. For the "only if" part, observe first that if $\pi = \pi^*$, (5) becomes $0 = P^t_{\pi^*} G^t_{\delta_t\pi^*}$. Also $P^t_{\pi^*}$ is nonsingular because it is a product of matrix exponentials, each of which is nonsingular. Thus $G^t_{\delta_t\pi^*} = 0$ for all continuity points of π^*. Now suppose $G^t_{\gamma\pi^*} \not\le 0$ for some continuity point t of π^* and some γ. For each state s with $G^t_{\gamma\pi^* s} \le 0$, redefine $\gamma^s = \delta^s_t$, whence $G^t_{\gamma\pi^* s} = G^t_{\delta_t\pi^* s} = 0$. Thus $G^u_{\gamma\pi^*} > 0$ for u in a sufficiently small neighborhood of t. Let π' be the policy that uses γ on that neighborhood and uses π^* elsewhere. Observe that $P^t_{\pi'}$ is nonnegative and has positive diagonal elements because it is a product of nonnegative matrix exponentials with those properties. Thus, it follows from the Comparison Lemma 1 that $V^0_{\pi'} > V^0_{\pi^*}$, which is a contradiction. Hence, $\max_{\gamma\in\Delta} G^t_{\gamma\pi^*} = 0 = G^t_{\delta_t\pi^*}$, or equivalently (6) holds, at continuity points of π^*. But V^t is continuous and so is $\mathcal{G}$, whence by the first equality in (6) at continuity points of $\dot V^\cdot$, $\dot V^t$ is continuous on $[0,T]$. ∎

Incidentally, $-\dot V^t = \mathcal{G}V^t$ is, of course, *Bellman's equation.*

5 EXISTENCE OF MAXIMUM T-PERIOD-VALUE POLICIES [Mi68b]

As a first step in showing that there is a maximum T-period-value piecewise-constant policy, we show how to find a stationary policy that has maximum t-period value for all small enough $t > 0$. To that end, let

$$\widehat V^t_\delta(v) = \left(\int_0^t e^{Q_\delta u}du\right) r_\delta + e^{Q_\delta t}v. \tag{1}$$

be the expected $t > 0$ period reward when δ^∞ is used and the terminal reward is v. For $t < 0$, $\widehat V^t_\delta(v)$ in (1) can be interpreted as the terminal reward for which the expected $-t > 0$ period reward when δ^∞ is used is v. To see this, notice that with this interpretation, we should have

$$v = \left(\int_0^{-t} e^{Q_\delta u}du\right) r_\delta + e^{-Q_\delta t}\widehat V^t_\delta(v). \tag{2}$$

Solving (2) for $\widehat V^t_\delta(v)$ gives

$$\widehat V^t_\delta(v) = -\left(\int_0^{-t} e^{Q_\delta(u+t)}du\right) r_\delta + e^{Q_\delta t}v = \left(\int_0^t e^{Q_\delta u}du\right) r_\delta + e^{Q_\delta t}v,$$

which shows that $\widehat V^t_\delta(v)$ satisfies (1) for $t < 0$.

Now consider the following two problems.

- **I. Maximum Value**. Choose $\delta \in \Delta$ to maximize $\widehat V^t_\delta(v)$ for all small enough $t > 0$.
- **II. Minimum Balloon Payment**. Choose $\delta \in \Delta$ to minimize $\widehat V^t_\delta(v)$ for all large enough $t < 0$.

The first problem has an obvious interpretation. The second problem can be interpreted as follows. Suppose that one must accumulate a total expected reward v_s during the next $-t$ periods starting with one individual in state s. Also suppose that there are no funds available initially and that if the system earns a total expected reward differing from v in the $-t$ periods, then each individual in state u after $-t$ periods must make a (possibly negative) balloon payment $\widehat{V}_\delta^t(v)_u$ at that time to make up the shortfall. The goal of problem II is to minimize the size of the balloon payments that each individual must make for small $-t$.

Example 4. Suppose $S = 1$, $\Delta = \{\gamma, \delta\}$, $r_\gamma = r_\delta = 0$, $v = 1$, $Q_\gamma = -1$ and $Q_\delta = -2$. Then γ^∞ solves problem I while δ^∞ solves problem II.

In order to solve problems I and II, substitute the Maclaurin-series expansion for $e^{Q_\delta u}$ into (1) and integrate term-by-term yielding

$$\pm\, \widehat{V}_\delta^{\pm t}(v) = \sum_{n=0}^{\infty} \frac{t^n}{n!} \psi_{\delta\pm}^n,\ t > 0 \tag{3}$$

where

$$\psi_{\delta\pm}^n = \begin{cases} \pm\, v & , n = 0 \\ (\pm\, Q_\delta)^{n-1}(r_\delta + Q_\delta v), & n = 1, 2, \ldots . \end{cases}$$

Let $\Psi_{\delta\pm}^n = (\psi_{\delta\pm}^0 \ \cdots \ \psi_{\delta\pm}^n)$ and $\Psi_{\delta\pm} = (\psi_{\delta\pm}^0 \ \psi_{\delta\pm}^1 \ \cdots\ )$. Let $\Delta_\pm$ be the set of $\delta \in \Delta$ that maximize (3) for all sufficiently small $t > 0$. From (3) we have $\Delta_\pm = \{\delta \in \Delta : \Psi_{\delta\pm} \succeq \Psi_{\gamma\pm}, \text{ all } \gamma \in \Delta\}$.

Theorem 3. Maximum-Value and Minimum-Balloon-Payment Stationary Policies for Small $|t|$. *There is a stationary policy maximizing (resp., minimizing) $\widehat{V}_\cdot^t(v)$ over the class of stationary policies for all sufficiently small $t > 0$ (resp., large $t < 0$), i.e., $\Delta_+ \neq \emptyset$ (resp., $\Delta_- \neq \emptyset$).*

Proof. Let $\Delta_\pm^n = \{\delta \in \Delta : \Psi_{\delta\pm}^n \succeq \Psi_{\gamma\pm}^n, \text{ all } \gamma \in \Delta\}$ for $n = 0, 1, \ldots$. Clearly $\Delta_\pm^0 = \Delta$, $\Delta_\pm^1 = \{\delta \in \Delta : r_\delta + Q_\delta v \geq r_\gamma + Q_\gamma v, \text{ all } \gamma \in \Delta\} \neq \emptyset$. Now suppose inductively that $\Delta_\pm^{n-1}$ is nonempty and is a product of action sets, one for each state, for some $n - 1 \geq 1$. Then $\Psi_{\delta\pm}^{n-1} = \Psi_\pm^{n-1}$ say, for $\delta \in \Delta_\pm^{n-1}$. Thus $\Delta_\pm^n = \{\delta \in \Delta_\pm^{n-1} : \psi_{\delta\pm}^n \geq \psi_{\gamma\pm}^n, \text{ all } \gamma \in \Delta_\pm^{n-1}\}$. Hence since $\psi_{\delta\pm}^n = \pm\, Q_\delta \psi_\pm^{n-1}$ for $\delta \in \Delta_\pm^{n-1}$, it follows that

$$\Delta_+^n = \{\delta \in \Delta_+^{n-1} : Q_\delta \psi_+^{n-1} \geq Q_\gamma \psi_+^{n-1}, \text{ all } \gamma \in \Delta_+^{n-1}\}$$

and

$$\Delta_-^n = \{\delta \in \Delta_-^{n-1} : Q_\delta \psi_-^{n-1} \leq Q_\gamma \psi_-^{n-1}, \text{ all } \gamma \in \Delta_-^{n-1}\},$$

so $\Delta_\pm^n$ is nonempty and is a product of action sets, one for each state. Thus $\Delta_\pm = \bigcap_{n=0}^\infty \Delta_\pm^n \neq \emptyset$. ∎

Theorem 4. Truncation. $\Delta_{\pm}^{S+1} = \Delta_{\pm}^{S+2} = \cdots = \Delta_{\pm}$.

Proof. If $\Psi_{\delta\pm}^{S+1} = \Psi_{\gamma\pm}^{S+1}$, then

$$(Q_\delta - Q_\gamma)\psi_{\delta\pm}^n = 0,\ n = 1, \ldots, S. \tag{4}$$

It suffices to show (4) holds for all $n \geq 1$ for this implies $\Psi_{\delta\pm} = \Psi_{\gamma\pm}$ from which the Theorem follows. That (4) holds for all $n \geq 1$ follows from Lemma 23 of §1.8 with $B = \pm Q_\delta$ and L the null space of $Q_\delta - Q_\gamma$. ∎

Theorem 5. Existence of Maximum T-Period-Value Policies. *For $T > 0$ and terminal value-vector v, there is a policy with maximum T-period value. Also $V = (V^t)$, where V^t is the maximum value on $[t, T]$, is the unique solution of Bellman's equation.*

Proof. Put $\Delta_\pm(v) \equiv \Delta_\pm$. Choose $\delta_1 \in \Delta_+(v)$. Let $\widehat{V}_\delta^t(v)$ be the t-period value when using δ and the terminal reward is v. Now show that δ_1 maximizes $\widehat{V}_\delta^t(v)$ over all piecewise-constant policies for all sufficiently small $t > 0$. By Theorem 2, this will be so if $\gamma = \delta_1$ maximizes $r_\gamma + Q_\gamma \widehat{V}_{\delta_1}^t(v)$ for all small enough $t > 0$. To see that this is so, notice on using $\widehat{V}_{\delta_1}^t(v) = \sum_{n=0}^\infty \frac{t^n}{n!}\psi_{\delta_1+}^n$ that

$$r_\gamma + Q_\gamma \widehat{V}_{\delta_1}^t(v) = [r_\gamma + Q_\gamma v] + tQ_\gamma\psi_{\delta_1+}^1 + \frac{t^2}{2!}Q_\gamma\psi_{\delta_1+}^2 + \cdots . \tag{5}$$

Hence, by construction, $\gamma = \delta_1$ lexicographically maximizes the coefficients of the above series and so indeed maximizes $r_\gamma + Q_\gamma \widehat{V}_{\delta_1}^t(v)$ for all small enough $t > 0$.

Let $t_1 \equiv \inf\{0 < t \leq T : \mathcal{G}\widehat{V}_{\delta_1}^t(v) > r_{\delta_1} + Q_{\delta_1}\widehat{V}_{\delta_1}^t(v)\} > 0$. Then $\pi^* = \delta_1$ for $T - t_1 \leq t \leq T$ is optimal. Now on replacing v by $\widehat{V}_{\delta_1}^{t_1}(v) \equiv v^1$, repeat the above construction obtaining $\delta_2 \in \Delta_+(v_1)$. Then define $\pi^* = \delta_2$ for $T - t_1 - t_2 \leq t < T_1 - t_1$ where $t_2 = \inf\{0 < t \leq T - t_1 : \mathcal{G}\widehat{V}_{\delta_2}^t(v_1) > r_{\delta_2} + Q_{\delta_2}\widehat{V}_{\delta_2}^t(v)\} > 0$. Continuing inductively, construct π^* and a sequence $t_1, t_2, \ldots$ of positive numbers. If $\sum_{i=1}^N t_i \geq T$ for some N, the claim is established. If not, $T^* \equiv T - \sum_{i=1}^\infty t_i \geq 0$. It turns out that this is impossible.

For let $v^* = \lim_{t\downarrow T^*} V_{\pi^*}^t$. From Theorem 3, there is a δ that minimizes $\widehat{V}_\cdot^{-t}(v^*)$ over the class of decisions for all small enough $t > 0$, i.e., $\delta \in \Delta_-(v^*)$. Now $\gamma = \delta$ maximizes $r_\gamma + Q_\gamma \widehat{V}_\delta^{-t}(v^*)$ for all small enough $t > 0$, say for $0 \leq t \leq \epsilon$, $\epsilon > 0$. To see this observe

$$r_\gamma + Q_\gamma \widehat{V}_\delta^{-t}(v^*) = [r_\gamma + Q_\gamma v^*] - tQ_\gamma\psi_{\delta-}^1 + \frac{t^2}{2}Q_\gamma\psi_{\delta-}^2 - \cdots . \tag{6}$$

It follows that on letting $\widetilde{V}^t \equiv \widehat{V}_\delta^{-t}(v^*)$,

$$-\dot{\widetilde{V}}^t = \mathcal{G}\widetilde{V}^t,\ \widetilde{V}^0 = v^*,\ 0 \leq t \leq \epsilon. \tag{7}$$

But from Theorem 4 and construction, $V^t \equiv V_{\pi^*}^{t+T^*}$ satisfies

$$-\dot{V}^t = \mathcal{G}V^t,\ V^0 = v^*,\ 0 \le t \le T - T^*. \tag{8}$$

Now the equations (7) and (8) are identical on $[0, \epsilon]$. Thus, since $\mathcal{G}$ is Lipschitz continuous, it follows from the theory of nonlinear differential equations in the Appendix that $\widetilde{V}^t = V^t$ for $0 \le t \le \epsilon$. Hence, redefine π^* to be δ on $[T^*, T^* + \epsilon]$ without altering its optimality, contradicting the assumption $t^N \downarrow 0$. ∎

Nonstationary Data. The above results concerning the existence of maximum T-period-value policies extend to situations in which the data are nonstationary. The simplest situation is that in which the action sets A_{st}, reward rates $r_t(s,a)$ and transition rates $q_t(u|s,a)$ at time t are piecewise constant in t on $[0,T]$, i.e., there exist constants $0 = T_0 < T_1 < T_2 < \cdots < T_N = T$ such that A_{st}, $r_t(s,a)$ and $q_t(u|s,a)$ are constant in t on the interval $[T_{i-1}, T_i)$ for each $i = 1, \ldots, N$. To see why there is a maximum T-period-value policy, suppose that one has found the maximum $(T-T_i)$-period value V^{T_i} and the corresponding policy π on the interval $[T_i, T]$ with terminal value v . Now apply Theorem 5 to find a maximum (T_i-T_{i-1})-period-value $V^{T_{i-1}}$ and the corresponding policy μ on the interval $[T_{i-1}, T_i)$ with terminal value V^{T_i}. Then the policy that uses μ on the interval $[T_{i-1}, T_i)$ followed by π on the interval $[T_i, T]$ has maximum $(T-T_{i-1})$-period value on the interval $[T_{i-1}, T]$ with terminal value v. Repeating this procedure for $i = N, \ldots, 1$ produces the desired maximum T-period-value policy and its value.

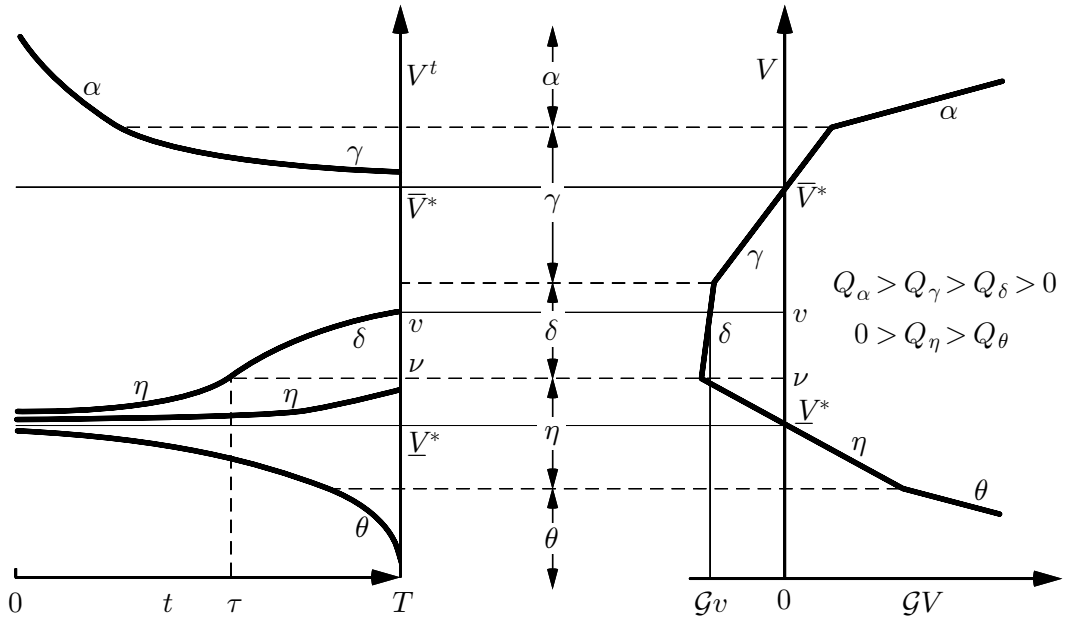
6 MAXIMUM T-PERIOD VALUE WITH A SINGLE STATE

It is instructive to apply the results for the maximum-T-period-value problem to the case where there is only a single state, i.e., $S = 1$. The general procedure is best described with an example.

Example 5. Suppose there is a single-state and five decisions $\alpha, \gamma, \delta, \eta, \theta$. Also suppose the graph of the piecewise-linear convex function $\mathcal{G}$ appears (landscape) in the right-hand side of Figure 1, with each linear segment labeled by the decision κ for which $\mathcal{G}V = r_\kappa + Q_\kappa V$. From the graph of $\mathcal{G}$ in the right-hand side of Figure 1, it is possible to sketch a rough graph of $V^\cdot$ in the left-hand side of Figure 1 that illustrates several features of the solution to $-\dot{V}^t = \mathcal{G}V^t$, $0 \le t \le T$, given any terminal value v. These features include:

- the monotonicity of V^t in t on $[0,T]$,
- the subinterval of $[0,T]$ during which it is optimal to use a decision,
- the monotonicity of $\dot{V}^t$ on each such subinterval,
- the values of V^t at the ends of each such subinterval and
- the order in which the decisions are used.

To see how, suppose one has at hand the value V^t at time t. Now find the value and sign of the derivative $\dot{V}^t$, and the decision κ to use at time t by moving horizontally from the value V^t on



$$-\dot{V}^t = \mathcal{G}V^t,\ 0 \le t \le T$$

Graphs of Four $V^{\cdot}$ Graph (Landscape) of $\mathcal{G}$

$$\mathcal{G}\underline{V}^* = 0 = \mathcal{G}\overline{V}^*$$

Figure 1. Graphs of $V^{\cdot}$ and $\mathcal{G}$ for a Single-State Problem

the vertical axis of the left-hand graph to the same value on the vertical axis of the right-hand graph and reading off $\dot{V}^t = -\mathcal{G}V^t = -r_\kappa - Q_\kappa V^t$. Using this information and starting with the given terminal value v, construct the graph of $V^{\cdot}$. This is done in the left-hand side of Figure 1 for four different terminal values.

To illustrate, consider the second largest v, say, of the four terminal values mentioned above. For that terminal value, $\dot{V}^T = -\mathcal{G}v = -r_\delta - Q_\delta v > 0$. Let ν be the value of V for which $r_\delta + Q_\delta \nu = r_\eta + Q_\eta \nu$, i.e., the value for which one is indifferent between using δ and η. Now as t decreases from T, V^t decreases and the derivative $\dot{V}^t$ increases as long as $V^t \ge \nu$. Call the time $t = \tau$ at which $V^t = \nu$ the *switching time*. Thus $V^{\cdot}$ is concave and increasing on the interval $[\tau, T]$. Now as t decreases from τ to 0, the derivative $\dot{V}^t = -r_\eta - Q_\eta V^t$ decreases, but remains positive, with V^t remaining above the value $\underline{V}^*$ at which the derivative of $V^{\cdot}$ vanishes, i.e., $\mathcal{G}\underline{V}^* = 0$. Thus $V^{\cdot}$ is convex and increasing on the interval $[0, \tau]$. And it is optimal to use η on $[0, \tau)$ and δ on $[\tau, T]$.

Actually, it is possible to give a closed form expression for the switching time τ. For at τ, we have

$$\mathcal{G}\nu = r_\delta + Q_\delta V^\tau = r_\delta + Q_\delta \left(\int_0^{T-\tau} e^{Q_\delta u} r_\delta du + e^{Q_\delta (T-\tau)} v \right) = e^{Q_\delta (T-\tau)} \mathcal{G}v.$$

Solving this equation for τ yields

$$\tau = T - Q_\delta^{-1}\ln\Big(\frac{\mathcal{G}\nu}{\mathcal{G}v}\Big)$$

provided that $\tau \geq 0$. In the contrary event, it is optimal to use δ on the entire interval $[0,T]$. Observe that $T-\tau$ is independent of T.

It is interesting to note that there are two zeros of $\mathcal{G}$, viz., $\underline{V}^*$ and $\overline{V}^*$. Both are equilibria in the sense that if v equals either of them, then $V^t = v$ for all $0 \leq t \leq T$. But the first equilibrium is a stable attractor and the second is an unstable repeller. The first is so because for all $v < \overline{V}^*$, $V^0 \to \underline{V}^* < \overline{V}^*$ as $T \to \infty$. The second is so from what was just shown and the fact that for all $v > \overline{V}^*$, $V^0 \to \infty$.

Theorem 6. Single State: Each Decision Used for an Interval. *If there is a single state, there is a maximum-T-period-value policy for which the set of times at which any given decision is used is a (possibly empty) interval and the maximum value V^t on $[t,T]$ is monotone in $0 \leq t \leq T$.*

Proof. First show that $\dot{V}^t$ has constant sign on $[0,T]$. To that end, suppose $\mathcal{G}v \geq 0$ (resp., ≤ 0). Then $-\dot{V}^t = \mathcal{G}V^t \geq 0$ (resp., ≤ 0) for all $0 \leq t \leq T$ because $\dot{V}^t$ is continuous in t and if there is a t for which $\dot{V}^t = 0$, then $V^\tau = V^t$ for all $0 \leq \tau \leq t$.

Next observe that $\mathcal{G}V$ is piecewise linear and convex in V. Hence, for each $\delta \in \Delta$, the set of values V for which $\mathcal{G}V = r_\delta + Q_\delta V$ is a subinterval of $\Re$. Now combine these facts with Theorem 2. ∎

One important implication of this result is that the number of switching times is bounded above by $|\Delta| - 1$.

7 EQUIVALENCE PRINCIPLE FOR INFINITE-HORIZON PROBLEMS [Ve69a]

This section studies the infinite-horizon version of the continuous-time-parameter problem. This problem can, for most purposes, be reduced to solving an "equivalent" discrete-time-parameter problem. To see this, it is useful to begin by discussing transient policies in continuous time.

Call a policy π *transient* if $\int_0^\infty P_\pi^t dt \ll \infty$. If $\pi = (\delta_t)$ is transient, its *value* is given by

$$V_\pi \equiv \int_0^\infty P_\pi^t r_{\delta_t} dt.$$

Now

$$I - e^{Q_\delta T} = -\int_0^T (\frac{d}{dt}e^{Q_\delta t})dt = (-Q_\delta)(\int_0^T e^{Q_\delta t}dt).$$

For δ^∞ to be transient, it is necessary and sufficient that $e^{Q_\delta T} \to 0$ as $T \to \infty$. The necessity is evident. For the sufficiency, observe that $I - e^{Q_\delta T}$ is nonsingular for large enough T, so the same is so of the two matrices in parentheses following the second equality in the last displayed equa-

tion. Thus, on premultiplying that equation by $-Q_\delta^{-1}$ and letting $T \to \infty$, it follows that $0 \leq \int_0^\infty e^{Q_\delta t} dt = -Q_\delta^{-1}$. Hence

$$V_\delta = \left(\int_0^\infty e^{Q_\delta t} dt \right) r_\delta = -Q_\delta^{-1} r_\delta.$$

Now on defining $P_\delta \equiv I + Q_\delta$, it follows that

$$V_\delta = (I - P_\delta)^{-1} r_\delta,$$

which is the value of δ^∞ for the equivalent discrete-time-parameter problem with P_δ and r_δ *provided* that $P_\delta \geq 0$. The nonnegativity of the P_δ can always be assured by taking the unit of time to be small enough, e.g., a minute rather than a day, in the continuous-time-parameter problem. A proportional change in the time unit necessitates a like change in the reward rate and the rate of addition of individuals to the population. In particular, on putting $\widehat{Q}_\delta \equiv aQ_\delta$, $\widehat{r}_\delta \equiv ar_\delta$ and $t \equiv a\tau$ where $a > 0$, and observing that the value V_δ of δ is *independent of the choice of the time unit*, it follows that

$$V_\delta = \left(\int_0^\infty e^{aQ_\delta \tau} d\tau \right) ar_\delta = -\widehat{Q}_\delta^{-1} \widehat{r}_\delta.$$

Thus if $a > 0$ is small enough, $\widehat{P}_\delta \equiv I + \widehat{Q}_\delta \geq 0$ and so $V_\delta = (I - \widehat{P}_\delta)^{-1} \widehat{r}_\delta$.

Equivalence of Discrete- and Continuous-Time-Parameter Problems. Suppose for this paragraph and the next that the time unit is chosen so $P_\delta = I + Q_\delta \geq 0$ for each δ and that each δ^∞ is transient in the continuous-time-parameter problem. As discussed above, the last is so if and only if each δ^∞ is transient in the *equivalent discrete-time-parameter problem* with P_δ and r_δ. Moreover, for each δ, $V_\delta = -Q_\delta^{-1} r_\delta = (I - P_\delta)^{-1} r_\delta$ is the value of δ^∞ in both problems. *Thus, in order to find a maximum or near-maximum value stationary policy for the continuous-time-parameter problem, it suffices to find such a policy for the equivalent discrete-time-parameter problem, e.g., by successive approximations, policy improvement or linear programming.*

This equivalence is also useful for theoretical studies. For example, V^* is the maximum value among the stationary policies in the continuous-time-parameter problem if and only if V^* is the maximum value among the stationary policies in the equivalent discrete-time-parameter problem. Thus, V^* is the unique fixed point of $\mathcal{R}$, or equivalently, the unique zero of $\mathcal{G}$ $(= \mathcal{R} - I)$, i.e., $0 = \mathcal{G}V^*$ or

$$0 = \max_{\delta \in \Delta} [r_\delta + Q_\delta V^*]. \tag{1}$$

This is consistent with the results for the t-period problem because if $V^t = V^*$ for all $t \geq 0$, then

$\dot{V}^t = 0$, whence (1) reduces to Bellman's equation for the t-period problem, viz., $0 = \dot{V}^t = \mathcal{G}V^t$ for all $t \geq 0$.

Maximum Present Value. Suppose instead that each stationary policy δ^∞ is *bounded*, i.e., $P_\delta^t = O(1)$. In this event, interest often centers on present-value optimality criteria. For that case, observe from what was shown above, for each δ the finite present value of future population sizes with the nominal interest rate $\rho > 0$ per unit time is $\int_0^\infty e^{-\rho t} P_\delta^t dt = \int_0^\infty e_\delta^{(Q_\delta - \rho I)t} dt = (\rho I - Q_\delta)^{-1}$ $= R_\delta^\rho$ where R_δ^ρ is, of course, the resolvent of Q_δ. Also, as long as the time unit is small enough (which here requires common rescaling of the r_δ, Q_δ *and* ρ) so that $P_\delta = I + Q_\delta \geq 0$ for all δ, then as shown on page 49, R_δ^ρ is also the present value of future population sizes in the equivalent discrete-time-parameter problem for all δ. *Moreover, in both cases the present values of* δ^∞ *coincide and equal* $V_\delta^\rho = R_\delta^\rho r_\delta$ *for each* δ. Thus both sets of maximum-present-value stationary policies coincide, and so do their present values. Also, the maximum-present-value V^* among the stationary policies for the continuous-time-parameter problem is the unique zero of the analog of (1) formed by replacing Q_δ by $Q_\delta - \rho I$, viz.,

$$0 = \max_{\delta \in \Delta} [r_\delta + Q_\delta V^* - \rho V^*]. \tag{2}$$

Thus, in order to find a maximum- or near maximum-present-value stationary policy for the continuous-time-parameter problem, it suffices to find such a policy for the equivalent discrete-time-parameter problem, e.g., by successive approximations, policy improvement or linear programming. Also, the sets of strong present-value optimal (resp., n-optimal) stationary policies in the continuous- and equivalent discrete-time-parameter problems coincide. Thus, one such policy can be found for the continuous-time-parameter problem by solving the equivalent discrete-time-parameter problem, e.g., by the strong policy-improvement method or linear programming.

8 MAXIMUM PRINCIPLE [PBGM62], [LM67]

It is often useful to consider models in which there are infinitely many states. This section does that for deterministic control problems whose transition laws are given by differential equations. The goals are to develop Bellman's equation and Pontryagin's celebrated *Maximum Principle* for these problems formally. By 'formal' is meant that differentiability of appropriate functions is assumed, but not shown. The Maximum Principle is a necessary condition for optimality that is also sufficient provided that the problem satisfies suitable convexity conditions, e.g., concavity of the objective function and linearity of the transition law.

The problem is that of finding a piecewise-continuous *control* $a^\cdot : [0, T] \to \Re^m$ that maximizes the T-period *value*

$$\int_0^T r(s^t, a^t) dt \tag{1}$$

subject to

$$\dot{s}^t = f(s^t, a^t),\ s^0 = s_0 \tag{2}$$

and

$$a^t \in A \tag{3}$$

for $0 \le t \le T$ where $s^{\cdot} : [0, T] \to \Re^n$ and s_0 is the given initial state. The interpretation is that $s^t \in \Re^n$ is the *state* of the system at time t, $a^t \in A \subseteq \Re^m$ is the *action* chosen at time t, $r(s^t, a^t)$ is the *reward rate* at time t, and (2) is the *transition law* governing the evolution of the system through the states.

In order to obtain the desired necessary condition for optimality, first formally generalize Bellman's equation

$$-\frac{\partial}{\partial t} V_s^t = \max_{a \in A} [r(s, a) + (Q_a V^t)_s],\ V_s^T = 0,$$

for $0 \le t \le T$ and all s, to the present case where V_s^t is the maximum value in $[t, T]$ starting in state s. The appropriate definition of $(Q_a V)_s$ in the present case is the differential operator

$$(Q_a V)_s = \lim_{h \downarrow 0} \frac{1}{h}(V_{s+hf(s,a)} - V_s) = f(s, a)\nabla_s V_s$$

where $\nabla_s V_s = (\frac{\partial}{\partial s_i} V_s)$ is the gradient of V_s because if $(s^t, a^t) = (s, a)$, then from (2),

$$s^{t+h} = s^t + hf(s^t, a^t) + o(h).$$

Thus *Bellman's equation* becomes the nonlinear partial differential equation

$$0 = \max_{a \in A} [r(s, a) + f(s, a)\nabla_s V_s^t + \frac{\partial}{\partial t} V_s^t],\ V_s^T = 0, \tag{4}$$

for $0 \le t \le T$ and all s. Now let $g(s, a, t)$ be the term in brackets in (4), $a^{\cdot}$ be an optimal control and $s^{\cdot}$ be the corresponding optimal trajectory satisfying the differential equations (2). Then from (4),

$$0 = \max_s g(s, a^t, t) = g(s^t, a^t, t) \tag{5}$$

because $g(s, a, t) \le 0$ for all states s, actions $a \in A$ and $0 \le t \le T$. Thus

$$\nabla_s g(s^t, a^t, t) = 0, \tag{6}$$

so on suppressing (s^t, a^t, t),

$$0 = \nabla_s r + \nabla_s f \nabla_s V + f\nabla_{ss}^2 V + \nabla_{st}^2 V$$

where $\nabla_{ss}^2 = \left(\frac{\partial^2}{\partial s_i \partial s_j}\right)$, $\nabla_{st}^2 = \left(\frac{\partial^2}{\partial s_i \partial t}\right)$ and $\nabla_s f = \left(\frac{\partial f_i}{\partial s_j}\right)$. Thus since $\dot{s} = f$,

$$\frac{d}{dt}\nabla_s V^t_{s^t} = \dot{s}^t \nabla^2_{ss} V^t_{s^t} + \nabla^2_{st} V^t_{s^t} = -\nabla_s r - \nabla_s f \nabla_s V^t_{s^t}. \tag{7}$$

Now set $v^t \equiv \nabla_s V^t_{s^t}$ and observe that $V^T_s = 0$ for all s, (7) becomes the *adjoint* linear differential equation

$$-\dot{v}^t = \nabla_s r(s^t, a^t) + \nabla_s f(s^t, a^t) v^t,\ v^T = 0 \tag{8}$$

for $0 \le t \le T$ and (4) implies the *maximum principle*

$$-\frac{\partial}{\partial t} V^t_{s^t} = \max_{a \in A} [r(s^t, a) + f(s^t, a) v^t] = r(s^t, a^t) + f(s^t, a^t) v^t \tag{9}$$

for $0 \le t \le T$. In order to check whether a control (a^t) satisfies (9), solve the ordinary differential equations (2) for (s^t) and then the linear differential equations (8) for (v^t). Finally, check whether a^t attains the maximum on the right-hand-side of (9).

In general the conditions (2), (8), and (9) are necessary, but not sufficient, for optimality of a control (a^t) because (6) is necessary, but not sufficient, for (5). Thus, like the analogous Karush-Kuhn-Tucker-Lagrange conditions for finite-dimensional mathematical programs, the maximum principle is necessary, but not sufficient, for optimality. However, again like analogous results for finite-dimensional concave programs, the maximum principle is sufficient for optimality under suitable convexity hypotheses, e.g., when $f(\,\cdot\,,\,\cdot\,)$ is linear, $r(s, a) = p(s) + q(a)$ is concave, A is convex, and mild regularity conditions are fulfilled. By contrast, a suitable form of Bellman's equation (4) is necessary and sufficient for optimality just as various versions of it introduced throughout this course are also necessary and sufficient for optimality.

Example 6. Linear Control Problem. For the special case in which r and f are linear, i.e.,

$$r(s, a) = sc + ad \text{ and } f(s, a) = sC + aD,$$

the adjoint equations (8) reduce to the linear differential equations with constant coefficients

$$-\dot{v}^t = c + Cv^t,\ v^T = 0$$

that are independent of the control and trajectory. Also the maximum principle simplifies to

$$\max_{a \in A} a(d + Dv^t) = a^t(d + Dv^t).$$

Observe that in this case, all that is required to find an optimal control is to solve the adjoint equations and then choose $a = a^t$ to achieve the maximum on the left-hand side of the above equation.

Example 7. Markov Population Decision Chain. Consider the S-state Markov branching decision chain discussed in the preceding section. In this event, s^t is the (row) vector of expected numbers of individuals in each state at time t. Then $A = \Delta$, $a = \delta$, $s^0 = s_0$,

$$r(s,a) = sr_\delta \text{ and } f(s,a) = sQ_\delta.$$

The adjoint equations simplify to the linear differential equations with variable coefficients

$$-\dot{v}^t = r_{\delta_t} + Q_{\delta_t} v^t,\ v^T = 0$$

that depend on the control, but not the trajectory. Thus, the maximum principle simplifies to

$$\max_{\delta\in\Delta} [s^t r_\delta + s^t Q_\delta v^t] = s^t r_{\delta_t} + s^t Q_{\delta_t} v^t = -s^t \dot{v}^t.$$

Now this is so for all initial populations $s^0 = s_0 \geq 0$, and hence all $s^t \geq 0$ in the corresponding convex cone of dimension S. Thus it follows by factoring out s^t in the above equation that

$$-\dot{v}^t = \mathcal{G} v^t = r_{\delta_t} + Q_{\delta_t} v^t,$$

which is precisely Bellman's equation (6) of §3.4 for the continuous-time-parameter problem.

9 MAXIMUM PRESENT-VALUE FOR CONTROLLED ONE-DIMENSIONAL DIFFUSIONS

[Ma67, 68], [Pu74]

The goal of this section is to introduce the theory of controlled one-dimensional diffusions in a manner analogous to the development of continuous-time-parameter Markov decision chains. As in the previous section, the generators here are differential operators. Otherwise, as will be seen in the sequel, the main features, formulas and results developed for Markov decision chains carry over in the same form to controlled one-dimensional diffusions, albeit with a few new complications. The development given below is analytic in the spirit of Feller, Volume I, and has the advantage of leading quickly to computational methods. A more fundamental treatment would, of course, require demonstrating that there exist stochastic processes with the properties described here.

Diffusions

Roughly speaking, a *one-dimensional diffusion* $\{S^t : t \geq 0\}$ with *state space* a compact interval of real numbers $\mathcal{S} = [s_0, s_1]$, $-\infty < s_0 < s_1 < \infty$, is a Markov process in which for all small $h > 0$, the conditional distribution of each increment $S^{t+h} - S^t$ given S^t is approximately normal with mean and variance proportional to h and with the proportionality constants depending on S^t. (If $\mathcal{S} = \Re$ and the proportionality constants are 0 and 1 respectively, the process is a *Wiener*

process or *Brownian motion.*) More precisely, assume that there exist piecewise-continuous *drift* and *diffusion* coefficients μ_s and σ_s^2, $s \in \mathcal{S}$, with $\sigma_s^2 > 0$ being bounded away from zero and bounded above such that for each $\epsilon > 0$,

$$(i) \qquad \int_{|u-s|>\epsilon} P^h(s, du) = o(h),$$

$$(ii) \qquad \int_{|u-s|\leq\epsilon} (u-s) P^h(s, du) = \mu_s h + o(h),$$

and

$$(iii) \qquad \int_{|u-s|\leq\epsilon} (u-s)^2 P^h(s, du) = \sigma_s^2 h + o(h)$$

at continuity points of μ and σ^2 where $P^t(s, B) = P(S^t \in B \mid S^0 = s)$. Observe that (i) asserts that the probability that the process deviates from any given initial state by more than ϵ in a small interval is small. Condition (ii) (resp., (iii)) asserts that the truncated mean drift (resp., square deviation) of the process from any initial state s in a small interval is essentially proportional to the length of the interval with the proportionality constant μ_s (resp., σ_s^2) depending on s.

Incidentally, the diffusion coefficient has an alternate interpretation as the infinitesimal variance. To justify this interpretation, it suffices to show that (ii) and (iii) are equivalent to (ii) and

$$(iii)' \qquad \int_{|u-s|\leq\epsilon} (u - s - \mu_s h)^2 P^h(s, du) = \sigma_s^2 h + o(h).$$

To see why this is so, observe that

$$\int_{|u-s|\leq\epsilon} (u - s - \mu_s h)^2 P^h(s, du) = \int_{|u-s|\leq\epsilon} (u-s)^2 P^h(s, du) - 2\mu_s h \int_{|u-s|\leq\epsilon} (u-s) P^h(s, du) + \mu_s^2 h^2 \int_{|u-s|\leq\epsilon} P^h(s, du).$$

Now the next to last term on the right-hand side of this equation is $o(h)$ by (ii) and the last term is $o(h)$ because the integral is bounded below by zero and above by one. Thus, it follows from the above equation that (ii) and (iii) are equivalent to (ii) and $(iii)'$ as claimed. While this formulation has intuitive appeal, it has the drawbacks that the conditions (ii) and $(iii)'$ are not independent and condition $(iii)'$ is less convenient to work with.

The next step is to derive the *generator* Q of the process defined by

$$(1) \qquad (QV)_s \equiv \lim_{h \downarrow 0} \frac{1}{h} \left[\int V_u P^h(s, du) - V_s \right]$$

for all bounded V with bounded piecewise-continuous $\ddot{V}$ and s an interior continuity point of $\ddot{V}$, μ and σ^2. It turns out that at such points

$$(QV)_s = \frac{1}{2}\sigma_s^2 \ddot{V}_s + \mu_s \dot{V}_s. \tag{2}$$

To see this, observe from (i)-(iii) that

$$\begin{aligned}(QV)_s &= \lim_{h\downarrow 0} \frac{1}{h}\int_{|u-s|\le\epsilon} (V_u - V_s)P^h(s,du)\\ &= \lim_{h\downarrow 0} \frac{1}{h}\int_{|u-s|\le\epsilon} [\dot{V}_s(u-s) + \frac{1}{2}\ddot{V}_s(u-s)^2]P^h(s,du)\\ &\qquad + \left[\lim_{h\downarrow 0} \frac{1}{h}\int_{|u-s|\le\epsilon} \frac{1}{2}(\ddot{V}_\theta - \ddot{V}_s)(u-s)^2 P^h(s,du)\right]\\ &= \frac{1}{2}\sigma_s^2\ddot{V}_s + \mu_s\dot{V}_s\end{aligned}$$

(with $|\theta - s| \le |u-s|$) because $\left|\ddot{V}_\theta - \ddot{V}_s\right|$, and hence the bracketed term above, can be made arbitrarily small by choosing $\epsilon > 0$ small enough.

The next step is to show that if the *reward rate* r_s earned per unit time in state s is piecewise continuous and bounded in s, then the expected *present value* V^ρ of rewards, viz.,

$$V^\rho = \int_0^\infty e^{-\rho t} P^t r dt,$$

for $\rho > 0$ where $(P^t r)_s \equiv \int_{\mathcal{S}} r_u P^t(s,du)$, satisfies

$$(\rho I - Q)V^\rho = r \tag{3}$$

at interior continuity points of r, μ and σ^2.[8] To see this, observe from (i) that $\lim_{h\downarrow 0} P^h r = r$ at such points. Thus

$$V^\rho = \int_0^h e^{-\rho t} P^t r dt + e^{-\rho h} P^h V^\rho = rh + (1-\rho h)P^h V^\rho + o(h).$$

Subtracting V^ρ, dividing by h and letting $h \downarrow 0$ establishes (3). Thus if, as can be shown, $\ddot{V}_\cdot^\rho$ is bounded and piecewise continuous, it follows by substituting (2) into (3) that V^ρ satisfies the second-order linear differential equation

$$\frac{1}{2}\sigma_s^2 \ddot{V}_s^\rho + \mu_s \dot{V}_s^\rho - \rho V_s^\rho = -r_s,\ s \in (s_0, s_1) \tag{4}$$

[8]Note that this equation is the same as the corresponding equation for Markov population decision chains except that here Q is a differential operator rather than a matrix.

at continuity points of r, μ and σ^2. Of course V^ρ is not the only solution to (4) because no boundary conditions have been specified.

The simplest type of boundary condition is *absorption* with a terminal value v_i at s_i for $i = 0, 1$. This implies that

$$V^\rho_{s_i} = v_i,\ i = 0, 1. \tag{5}$$

It turns out that (4) has a unique solution subject to the *two-point boundary conditions* (5). We show this under the hypotheses that r_s, μ_s and σ^2_s are piecewise constant. To see this, put

$$\overline{V} \equiv \begin{pmatrix} V^\rho \\ \dot{V}^\rho \end{pmatrix},\ \overline{Q}_s \equiv \begin{pmatrix} 0 & 1 \\ \dfrac{2\rho}{\sigma^2_s} & \dfrac{-2\mu_s}{\sigma^2_s} \end{pmatrix} \text{ and } \overline{r}_s \equiv \begin{pmatrix} 0 \\ \dfrac{-2r_s}{\sigma^2_s} \end{pmatrix},$$

and rewrite (4) as

$$\dot{\overline{V}}_s = \overline{r}_s + \overline{Q}_s \overline{V}_s,\ s \in (s_0, s_1). \tag{4}'$$

In this form, (4) is evidently the differential equation for the $s - s_0$ period value in a two-state continuous-time-parameter Markov population chain where $\overline{r}_s$ and $\overline{Q}_s$ are respectively the reward-rate vector and transition-rate matrix $s - s_0$ units of time from the end of the process. With this in mind, let $\overline{P}_{st}$ be the "reverse time" transition matrix from "time" s to "time" t. Since $\overline{Q}_s$ has positive 12^{th} element for $s \in \mathcal{S}$, so does $\overline{P}_{st}$ for all $t < s$ in $\mathcal{S}$. Then

$$\overline{V}_{s_1} = \overline{P}_{s_1 s_0} \overline{V}_{s_0} + \text{constant}.$$

Hence the first component of $\overline{V}_{s_1}$ is a strictly increasing affine function of $\overline{V}_{s_0}$. Thus since the first component of $\overline{V}_{s_0}$ is given as v_0, the second component of $\overline{V}_{s_0}$ can be chosen to assure the first component of $\overline{V}_{s_1}$ equals v_1. Therefore, (4) has a unique solution subject to the boundary conditions (5).

Observe that the above argument remains valid even when $\rho = 0$, whence $V \equiv V^0$ is the unique solution to (4), (5) with $\rho = 0$. In that event, V_s is the total expected reward earned starting from state s.

It is useful now to apply these results to solve some simple problems associated with the Wiener process with zero drift and unit diffusion coefficients.

Example 8. Probability of Reaching One Boundary Before the Other. What is the probability V_s of reaching s_1 before s_0 starting from $s \in [s_0, s_1]$? Then $r_s = 0$ on (s_0, s_1) and V satisfies

$$-\frac{1}{2}\ddot{V}_s = 0,\ V_{s_0} = 0,\ V_{s_1} = 1.$$

Thus $V_s = \alpha s + \beta$, so $0 = \alpha s_0 + \beta$ and $1 = \alpha s_1 + \beta$ whence

$$V_s = \frac{s - s_0}{s_1 - s_0}.$$

Example 9. Mean Time to Reach Boundary. What is the mean time V_s until absorption at the boundaries starting from $s \in [s_0, s_1]$? Then $r_s = 1$ on (s_0, s_1) and V satisfies

$$-\frac{1}{2}\ddot{V}_s = 1,\ V_{s_0} = V_{s_1} = 0.$$

Thus

$$\begin{aligned} -V_s &= s^2 + \alpha s + \beta \\ 0 &= s_0^2 + \alpha s_0 + \beta \\ 0 &= s_1^2 + \alpha s_1 + \beta \end{aligned}$$

so

$$V_s = (s - s_0)(s_1 - s).$$

Controlled Diffusions

A diffusion process is controlled by changing its drift and diffusion coefficients. Thus suppose A is the finite set of *actions* available in any state, and that μ_{sa} and $\sigma_{sa}^2 > 0$ are respectively the *drift* and *diffusion coefficients* and r_{sa} is the *reward rate* when one chooses action a in state s. Assume that μ_{sa}, $\sigma_{sa}^2 > 0$ and r_{sa} are each piecewise constant in s and σ_{sa}^2 is bounded away from zero and bounded above for each $a \in A$. Also assume that there is absorption with terminal rewards v_0 and v_1 at the two finite boundaries $s_0 < s_1$. When the interest rate is $\rho > 0$, earning the terminal reward v_i at the boundary s_i is equivalent to earning $r_{s_i a} \equiv \rho v_i$ per unit time there forever.

A *decision* is a piecewise constant function δ on the interval $[s_0, s_1]$ with δ^s being the action that δ chooses when the process is in state $s \in [s_0, s_1]$. When using δ, the *reward function* is $r_\delta = (r_{s\delta^s})$ for all $s \in [s_0, s_1]$. Also, in the interior of the state space, the *generator* Q_δ is defined by

$$(Q_\delta V)_s = \frac{1}{2}\sigma_{s\delta^s}^2 \ddot{V}_s + \mu_{s\delta^s}\dot{V}_s,\ s \in (s_0, s_1) \tag{6}$$

at continuity points s of $\ddot{V}_s$, δ^s, $\mu_{s\delta^s}$ and $\sigma_{s\delta^s}^2$. Since there is absorption at the boundaries, it follows from the definition (1) of the generator that Q_δ is defined at the boundaries by

$$(Q_\delta V)_s = 0,\ s = s_0, s_1. \tag{7}$$

A *stationary policy* δ^∞ entails using δ at each point in time. If the goal is to find a stationary policy that maximizes the expected present value of the rewards earned over an infinite time interval, Bellman's equation becomes

$$0 = \max_{\gamma\in\Delta}[r_\gamma + Q_\gamma V^\rho - \rho V^\rho] \tag{8}$$

where Δ is the set of decisions and V^ρ is the maximum expected present value of rewards. Note that with this formulation, the boundary conditions are absorbed into (8) since its restriction to the boundary states s_i are, in view of (7), equivalent to $\rho v_i - \rho V^\rho_{s_i} \equiv 0$ for $i = 0, 1$, or what is the same thing, $V^\rho_{s_i} \equiv v_i$, for $i = 0, 1$.[9] To see this, let R^ρ_δ be the nonnegative linear operator that assigns to r_δ the unique solution $V = V^\rho_\delta$ of the equation

$$(\rho I - Q_\delta)V = r_\delta. \tag{9}$$

Thus, $V^\rho_\delta = R^\rho_\delta r_\delta$ and R^ρ_δ is invertible with $(R^\rho_\delta)^{-1} = \rho I - Q_\delta$. This leads to the following version of the Comparison Lemma,

$$V^\rho_\gamma - V^\rho_\delta = R^\rho_\gamma[r_\gamma - (R^\rho_\gamma)^{-1}V^\rho_\delta] = R^\rho_\gamma G^\rho_{\gamma\delta}$$

where the *comparison function* is defined by

$$G^\rho_{\gamma\delta} \equiv r_\gamma + Q_\gamma V^\rho_\delta - \rho V^\rho_\delta.$$

Thus, $\gamma = \delta$ maximizes V^ρ_γ over Δ if $G^\rho_{\gamma\delta} \le 0$ for all $\gamma \in \Delta$ since (9) can be rewritten as $G^\rho_{\delta\delta} = 0$. Hence Bellman's equation (8) holds. On substituting the definition (6) and (7) of the differential operator Q_γ into (8), observe that (8) becomes

$$0 = \max_{a\in A}[r_{sa} + \frac{1}{2}\sigma^2_{sa}\ddot{V}^\rho_s + \mu_{sa}\dot{V}^\rho_s - \rho V^\rho_s],\ s \in (s_0, s_1) \tag{8$'$}$$

at interior continuity points of μ, σ and r, and

$$V^\rho_{s_i} \equiv v_i,\ i = 0, 1 \tag{8$''$}$$

at the boundaries.

Also, it can be shown that there is a stationary optimal policy δ^∞ with δ piecewise constant, even if the data A, r_{sa}, σ^2_{sa} and μ_{sa} are piecewise constant in s. The method of doing this entails first reducing the infinite-horizon controlled-diffusion problem to an equivalent finite-horizon ($T = s_1 - s_0$ period) two-state continuous-time-parameter Markov population decision chain as

[9]Here again the equation (8) is the same as that for the corresponding Markov population decision chain except that here the generators Q_γ are differential operators rather than matrices.

discussed above (see (4)′). Then show that a piecewise-constant (now in time) policy is optimal for the latter problem by using arguments like those given in §3.5—especially the last paragraph thereof.

Example 10. Maximizing the Probability of Accumulating Given Wealth. Consider an investor who seeks to maximize his probability of accumulating a million dollars before going bankrupt starting with s million dollars where $0 \le s \le 1$. Denote by V_s the desired maximum probability. Assume that the investor's wealth is a controlled diffusion with drift coefficient μ and diffusion coefficient $\sigma^2 > 1$ or 1 respectively according as the investor acts boldly or timidly.

Now V is the unique solution of Bellman's equation (with $\rho = 0$)

$$0 = (\frac{1}{2}\sigma^2 \ddot{V}_s + \mu \dot{V}_s) \vee (\frac{1}{2}\ddot{V}_s + \mu \dot{V}_s) = (\frac{1}{2}\sigma^2 \ddot{V}_s) \vee (\frac{1}{2}\ddot{V}_s) + \mu \dot{V}_s \tag{10}$$

for $s \in (0,1)$ subject to the boundary conditions $V_0 = 0$ and $V_1 = 1$. Observe from (10) that *bold* (resp., *timid*) *investing* is optimal if and only if $\ddot{V}_s \ge 0$ (resp., $\ddot{V}_s \le 0$) on $[0,1]$, i.e., the investor's maximum probability V_s is convex (resp., concave) on $[0,1]$. This is so if and only if $\mu \le 0$ (resp., $\mu \ge 0$). Here is the argument for bold investing; the other case is similar. Bold investing is optimal if and only if

$$0 = \frac{1}{2}\sigma^2 \ddot{V}_s + \mu \dot{V}_s,\ s \in (0,1), \tag{11}$$

$V_0 = 0$ and $V_1 = 1$. On rewriting (11) as

$$\frac{1}{2}\sigma^2 \frac{\ddot{V}_s}{\dot{V}_s} = -\mu$$

and then integrating gives

$$\frac{1}{2}\sigma^2 \ln \dot{V}_s = -s\mu + \alpha$$

for $s \in [0,1]$ and some α. Thus $\dot{V}$ is nondecreasing, i.e., $\ddot{V} \ge 0$, if and only if $\mu \le 0$ as claimed.

This result has the following intuitive explanation. If the investor earns the mean return, the investor would eventually accumulate a million dollars if the drift is positive and eventually go bankrupt if the drift is negative. Consequently, if the drift is positive, timid investing is the better option since it has higher probability of keeping the wealth close to the mean and this is all that is needed to reach a million dollars in this case. On the other hand, if the drift is negative, bold investing is the better option since it has higher probability of raising the wealth significantly above the mean which is needed to accumulate a million dollars in this case.

This page is intentionally left blank.

Appendix

Functions of Matrices

The purpose of this appendix is to summarize a few useful facts about functions of matrices, especially the Jordan form and its applications to matrix power series, matrix exponentials, etc.

1 MATRIX NORM

A norm $\|P\|$ of a complex matrix $P = (p_{ij})$ is a real valued function that satisfies 1°-4° below. A norm of a matrix is a measure of its distance from the null matrix. Throughout, all norms are presumed to satisfy 5° as well. One example is the Tchebychev norm $\|P\| = \max_i \sum_j |p_{ij}|$. Indeed, the term "norm" means the Tchebychev norm in the main text unless stated otherwise.

1° $\|P\| \geq 0$.

2° $\|P\| = 0$ if and only if $P = 0$.

3° $\|\lambda P\| = |\lambda| \|P\|$ for each complex number λ. (*homogeneity*)

4° $\|P + Q\| \leq \|P\| + \|Q\|$ if the matrix sum $P + Q$ is well defined. (*triangle inequality*)

5° $\|PQ\| \leq \|P\| \|Q\|$ if the matrix product PQ is well defined. (*Schwarz inequality*)

2 EIGENVALUES AND VECTORS

An *eigenvalue* of a square complex matrix P is a complex number λ such that $\lambda I - P$ is singular. The *spectrum* $\sigma(P)$ of P is the set of its eigenvalues. The number of eigenvalues, allowing for multiplicities, equals the order of P. The *spectral radius* σ_P of P is the radius of the smallest circle in the complex plane having center at the origin and containing the spectrum of P. For each eigenvalue λ of P, there is nonnull vector v for which $\lambda v = Pv$. Call v an *eigenvector* of P. As an application of these definitions, note that $\sigma_{\lambda P} = |\lambda|\sigma_P$ and $\sigma_P \leq \|P\|$. To see the last fact, let λ be an eigenvalue of P and v be an associated eigenvector. Then since $\lambda v = Pv$,

$$|\lambda|\|v\| = \|\lambda v\| = \|Pv\| \leq \|P\|\|v\|,$$

whence on dividing by $\|v\| \neq 0$, the result follows.

3 SIMILARITY

A square complex matrix P is *similar* to a square complex matrix Q of like order if there is a nonsingular matrix T such that $P = TQT^{-1}$. The similarity relation is evidently an *equivalence relation*, i.e., is transitive, reflexive and symmetric. More important is the fact that two similar matrices have the *same* spectrum. And if P is similar to Q, P^N is similar to Q^N. More generally, if $f(P) \equiv \sum_{N=0}^{\infty} a_N P^N$ is absolutely convergent, then so is $f(Q)$, and $f(P)$ is similar to $f(Q)$. Also $f(P)$ is absolutely convergent if that is so of $f(\|P\|)$. It turns out that among each equivalence class of similar matrices, there is a matrix having a particularly simple form and whose properties it is easier to study. That matrix is the *Jordan form*.

4 JORDAN FORM

Call a square complex matrix P *nilpotent* if $P^N = 0$ for some $N \geq 1$. A *Jordan matrix* of order S is an $S \times S$ matrix J of the form $J \equiv \lambda I + E$ where λ is a complex number and E is the matrix having ones just above the diagonal and zeroes elsewhere. For example, if $S = 4$,

$$E = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Observe that $J = \lambda I + E$ has the following properties:

1° λ is the unique eigenvalue of J;
2° $E^S = 0$ (so E is nilpotent);
3° $J^N = \sum_{k=0}^{S-1} \binom{N}{k} \lambda^{N-k} E^k$, $N \geq S$; and
4° $J^N \to 0$ if and only if $|\lambda| < 1$.

Property 3° follows from 2° and the binomial theorem

$$(P+Q)^N = \sum_{k=0}^{N} \binom{N}{k} P^{N-k} Q^k,$$

which is valid for square matrices P, Q that *commute*, i.e., $PQ = QP$, as do I and E. Property 4° follows from 3°. Notice that $\binom{N}{k}$ is a polynomial in N of degree k. Thus 3° implies that $J^N = O(N^{S-1}\lambda^N)$.

A square complex matrix is in *Jordan form* if it is block-diagonal and each diagonal block is a Jordan matrix. The spectrum of a matrix P in Jordan form is simply the union of the spectra of its diagonal blocks, say $J_1, \ldots, J_k$, where $J_i \equiv \lambda_i I + E$ for all i. Then

$$P = \begin{pmatrix} J_1 & & & 0 \\ & J_2 & & \\ & & \ddots & \\ 0 & & & J_k \end{pmatrix}$$

and $\{\lambda_1, \ldots, \lambda_k\}$ is the spectrum of P.

The following representation theorem for square complex matrices in terms of matrices in Jordan form is of great importance. It enables many questions about square complex matrices to be reduced to the same questions about matrices in Jordan form.

Theorem 1. Jordan Form. *Every square complex matrix is similar to a matrix in Jordan form.*

Corollary 2. Transient Matrices. *If P is a square complex matrix, then $P^N \to 0$ if and only if $\sigma_P < 1$.*

Proof. The result follows from 4° above for Jordan matrices and so for P by Theorem 1. ∎

Corollary 3. Nilpotent Matrices. *If P is an $S \times S$ complex matrix, the following are equivalent.*

1° *P is nilpotent.*

2° *$P^S = 0$.*

3° *$\sigma(P) = \{0\}$.*

Proof. The result follows easily from 3° above for Jordan matrices and so is true for P by Theorem 1. ∎

5 SPECTRAL MAPPING THEOREM

The next result asserts that the spectrum of a matrix power series is the set of complex numbers that one obtains by replacing the matrix in the power series by each of its eigenvalues. If f is a function and S is a subset of its domain, let $f(S)$ be the image of S under f, i.e., $\{f(\lambda) : \lambda \in S\}$.

Theorem 4. Spectral Mapping. *If P is a square complex matrix and $f(P) \equiv \sum_{N=0}^{\infty} a_N P^N$ is absolutely convergent, then $f(\lambda)$ is absolutely convergent for each $\lambda \in \sigma(P)$ and $\sigma(f(P)) = f(\sigma(P))$.*

Proof. By Theorem 1, it suffices to prove the result for the Jordan form of P, and hence for each of its Jordan blocks $J = \lambda I + E$, $\lambda \in \sigma(P)$. Now J^N is upper triangular with common diagonal element λ^N, so $f(J)$ is upper triangular with common diagonal element $f(\lambda)$. Thus if $f(J)$ is absolutely convergent, then $f(\lambda)$ is absolutely convergent and is the unique eigenvalue of $f(J)$, so $\sigma(f(J)) = f(\sigma(J))$. ∎

6 MATRIX DERIVATIVES AND INTEGRALS

Often interest centers on a *matrix function* $X(\,\cdot\,) = (x_{ij}(\,\cdot\,))$ of real-valued functions x_{ij} on an interval T of real numbers with $X(t)$ being $m \times n$ for each $t \in T$. The *matrix derivative* of X at t, denoted $\frac{d}{dt}X(t)$ (or $\dot{X}(t)$), is the matrix $(\frac{d}{dt}x_{ij}(t))$ (or $(\dot{x}_{ij}(t))$) of first-order derivatives of the x_{ij}. Similarly, the *matrix integral* of X, denoted $\int X(t)dt$, is the matrix $\left(\int x_{ij}(t)dt\right)$ of integrals of the x_{ij}. Matrix-valued functions inherit many properties of real-valued functions. It is easy to verify the following facts. If $X(t)$ and $Y(t)$ are matrix-valued functions, then

$$\frac{d}{dt}(X(t) + Y(t)) = \frac{d}{dt}X(t) + \frac{d}{dt}Y(t) \text{ and } \int (X(t) + Y(t))dt = \int X(t)dt + \int Y(t)dt$$

provided the matrices are of the same size. Also,

$$\frac{d}{dt}(X(t)Y(t)) = (\frac{d}{dt}X(t))Y(t) + X(t)\frac{d}{dt}Y(t)$$

provided the number of columns of $X(t)$ equals the number of rows of $Y(t)$. Finally,

$$\frac{d}{dt}\int_0^t X(u)du = X(t).$$

7 MATRIX EXPONENTIALS

There is a powerful and useful method of extending the definition of a function $f(x)$ of a variable x with a Maclaurin expansion to a function $f(Q)$ of a square matrix Q. The method entails substituting Q^N for x^N for each N in the Maclaurin expansion of $f(x)$, provided that the Maclaurin expansion of $f(Q)$ is absolutely convergent.

An important example of this method of defining a matrix function is the matrix exponential. The function e^x of the variable x has the Maclaurin expansion $e^{\mathrm{x}} = \sum_{N=0}^{\infty} \frac{1}{N!}x^N$. Hence one can define the *matrix exponential* e^Q of the square matrix Q by

$$e^Q \equiv \sum_{N=0}^{\infty} \frac{1}{N!}Q^N$$

because the Maclaurin series defining e^Q is absolutely convergent. Thus, from the Spectral Mapping Theorem, $\sigma(e^Q) = e^{\sigma(Q)}$. If also P is a square complex matrix of the same order as Q and commutes therewith, it follows easily from the Binomial Theorem in §A.4 above that

$$e^{P+Q} = e^P e^Q = e^Q e^P. \tag{1}$$

One consequence is that e^Q is nonsingular and its inverse is e^{-Q} because $I = e^0 = e^{Q-Q} = e^Q e^{-Q}$. Also, $\frac{d}{dt} e^{Qt} = Qe^{Qt} = e^{Qt}Q$.

8 MATRIX DIFFERENTIAL EQUATIONS

Matrix exponentials are important because they provide an elegant and useful tool for studying the solution of a system of linear differential equations.

Theorem 5. Solution of Matrix Differential Equation. *If Q is an $m \times m$ real matrix and q is a continuous function from the nonnegative real line to $\Re^{m+k}$, then the system of matrix linear differential equations*

$$\dot{X}(t) = q(t) + QX(t),\ X(0) = X_0 \tag{2}$$

has a unique continuously differentiable solution, viz.,

$$X(t) = \int_0^t e^{Q(t-u)} q(u)du + e^{Qt}X_0. \tag{3}$$

In particular, if $k = m$, $q = 0$ and $X_0 = I$, then $X(t) = e^{Qt}$.

Proof. To show that (3) satisfies (2), rewrite (3) as

$$X(t) = e^{Qt}\left(\int_0^t e^{-Qu} q(u)du + X_0\right).$$

Then differentiate both sides of this equation. Conversely, if X and $\widehat{X}$ are two solutions of (2), then $W \equiv X - \widehat{X}$ satisfies $\dot{W}(t) = QW(t)$, $W(0) = 0$. Thus

$$\frac{d}{dt} e^{-Qt} W(t) = -Qe^{-Qt}W(t) + e^{-Qt}\dot{W}(t) = 0$$

so $e^{-Qt}W(t)$ is constant in t. But that constant is 0 at $t = 0$, so $e^{-Qt}W(t) = 0$. Hence $W = 0$, and so $X = \widehat{X}$. ∎

The next result is a continuous-time analog of Corollary 2.

Lemma 6. Transient Matrices in Continuous Time. *If Q is a square complex matrix, the following are equivalent.*

1° $e^{Qt} \to 0$ *as* $t \to \infty$.

2° $\sigma_{e^Q} < 1$.

3° $\left|e^{\lambda}\right| < 1$ *for all* $\lambda \in \sigma(Q)$.

4° *The real parts of the eigenvalues of* Q *are negative.*

Proof. 1° $\Leftrightarrow$ 2°. Since $e^{QN} = (e^Q)^N$ for $N = 0, 1, \ldots$ by (1), the claim follows from Corollary 2 when t runs through the integers. It remains to show that $e^{Q[t]} \to 0$ implies $e^{Qt} \to 0$ where $[t]$ is the integer part of t. Evidently, $e^{Qt} = e^{Q[t]}e^{Q(t-[t])}$, so because $\left\|e^Q\right\| \leq \sum_{N=0}^{\infty} \frac{1}{N!}\|Q\|^N \leq e^{\|Q\|}$,

$$\left\|e^{Qt}\right\| = \left\|e^{Q[t]}\right\| \sup_{0 \leq u \leq 1} \left\|e^{Qu}\right\| \leq \left\|e^{Q[t]}\right\| e^{\|Q\|},$$

from which the claim follows.

2° $\Leftrightarrow$ 3°. Immediate from $\sigma(e^Q) = e^{\sigma(Q)}$.

3° $\Leftrightarrow$ 4°. Immediate from standard properties of complex exponentials. ∎

References

Books

Arrow, K. J. (1965). *Aspects of the Theory of Risk Bearing*. Academic Bookstore, Helsinki, Finland.

Aho, A., J. Hopcroft, and J. Ullman (1974). *The Design and Analysis of Computer Algorithms*. Addison-Wesley, Reading, Mass.

Bellman, R. E. (1957). *Dynamic Programming*. Princeton University Press, Princeton, NJ.

Bellman, R. E. and S. E. Dreyfus (1962). *Applied Dynamic Programming*. Princeton University Press, Princeton, NJ.

Bensoussan, A., E. G. Hurst, Jr., and B. Naslund (1974). *Management Applications of Modern Control Theory*. American Elsevier, New York.

Bertsekas, D. P. (2007). *Dynamic Programming and Optimal Control.* I and II. 3rd ed. Athena Scientific, Belmont, Massachusetts.

Bertsekas, D. P. and S. E. Shreve (1978). *Stochastic Optimal Control: The Discrete Time Case*. Academic Press, San Francisco.

Bertsekas, D. P. and J. N. Tsitsiklis (1996). *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

Chow, Y. S., H. Robbins and D. Siegmund (1971). *Great Expectations: The Theory of Optimal Stopping*. Houghton Mifflin, Boston.

Denardo, E. V. (1982). *Dynamic Programming: Models and Applications.* Prentice-Hall, Englewood Cliffs, NJ.

Derman, C. (1970). *Finite State Markovian Decision Processes*. Academic Press, New York.

Doob, J. L. (1953). *Stochastic Processes*. Wiley, New York.

Dubins, L. E. and L. J. Savage (1965). *How to Gamble if You Must*. McGraw-Hill, New York.

Dynkin, E. B. and A. A. Yushkevich (1969). *Markov Processes*. Translated by J. Wood. Plenum Press, New York.

Feller, W. (1957). *An Introduction to Probability Theory and Its Applications, Vol. 1, Second Ed.* Wiley, New York.

Fleming, W. H. and R. W. Rishel (1975). *Deterministic and Stochastic Control*. Springer-Verlag, New York.

Gittens, J. C. (1989). *Multiarmed Bandit Allocation Indices.* Wiley, Chichester.

Halmos, P. (1958). *Finite Dimensional Vector Spaces*. Van Nostrand, Princeton, NJ.

Hardy, G. H. (1949). *Divergent Series*. Clarendon Press, Oxford.

Holt, C. C., F. Modigliani, J. F. Muth and H. A. Simon (1960). *Planning Production, Inventories, and Work Force*. Prentice-Hall, Englewood Cliffs, NJ.

Hordijk, A. (1974). *Dynamic Programming and Markov Potential Theory*. Mathematical Centre Tract No. 51, Mathematical Center, Amsterdam.

Howard, R. A. (1960). *Dynamic Programming and Markov Processes*. The Technology Press, MIT, Cambridge, Mass.

Howard, R. A. (1971). *Dynamic Probabilistic Systems. Vol. II: Semi-Markov and Decision Processes.* Wiley, New York.

Karlin, S. (1959). *Mathematical Methods and Theory in Games, Programming, and Economics, Vol. I.* Addison-Wesley, Reading, Mass.

Kato, T. (1966). *Perturbation Theory for Linear Operators*. Springer-Verlag, New York.

Kemeny, J. G. and J. L. Snell (1960). *Finite Markov Chains*. Van Nostrand, Princeton, NJ.

Kushner, H. (1967). *Stochastic Stability and Control*. Academic Press, New York.

Kushner, H. (1971). *Introduction to Stochastic Control*. Holt, Rinehart, and Winston, New York.

Lee, E. B. and L. Markus (1967). *Foundations of Optimal Control Theory.* Wiley, New York.

Mandl, P. (1968). *Analytical Treatment of One-Dimensional Markov Processes*. Springer-Verlag, New York.

Marschak, J. and R. Radner (1972). *Economic Theory of Teams.* Cowles Foundation Monograph. New Haven and London, Yale University Press.

Martin, J. (1967). *Bayesian Decision Problems and Markov Chains*. Wiley, New York.

Pontryagin, L. S., V. G. Boltyanskiii, R. V. Gamkrelidze and E. F. Mishchenko (1962). *The Mathematical Theory of Optimal Processes.* Translated from Russian by K. N. Trirogoff and Edited by L. W. Neustadt. Interscience Publishers, Wiley, New York.

Puterman, M. L. (1994). Markov Decision Processes: Discrete Stochastic Dynamic Programming. Wiley, New York.

Ross, S. M. (1983). *Introduction to Stochastic Dynamic Programming.* Academic Press, New York.

Veinott, A. F., Jr. (1965). *Mathematical Studies in Management Science*. Macmillan, New York.

Whittle, P. (1983). *Optimization Over Time: Dynamic Programming and Stochastic Control. Vols. I & II.* Wiley, New York.

Surveys

Bellman, R. E. and R. Karush (1964). *Dynamic Programming: A Bibliography of Theory and Application.* RM-3951-PR, 140 pp. and RM-3591-1-PR, 142 pp., the RAND Corporation, Santa Monica, CA.

Breiman, L. (1964). *Stopping Rule Problems.* In E. Beckenbach (ed.). *Applied Combinatorial Mathematics.* Wiley, New York.

Dantzig, G. B. (1959). On the Status of Multistage Linear Programming Problems. *Man. Sci.* 6, 1, 53-72.

Dreyfus, S. E. (1969). An Appraisal of Some Shortest-Path Algorithms. *Opns. Res.* 17, 3, 395-412.

Fleming, W. (1966). Optimal Control of Diffusion Processes. In E. Caianello (ed.). *Functional Analysis and Optimization.* Academic Press, New York, 67-84.

Fulkerson, D. (1966). Flow Networks and Combinatorial Operations Research. *Amer. Math. Monthly* 73, 115-138.

Mandl, P. (1967). Analytical Methods in the Theory of Controlled Markov Processes. In J. Kozesnik (ed.). *Transactions of the Fourth Prague Conference on Information Theory, Statistical Decision Functions and Random Processes*, 1965. Academic Press, New York, 45-53.

Puterman, M. L. (1990). Markov Decision Processes. Chapter 8 in D. P. Heyman and M. J. Sobel (eds.). *Handbooks in Operations Research and Management Science* 2, Elsevier (North-Holland), 331-434.

Veinott, A. F., Jr. (1965). Commentary on Part Two: Stochastic Decision Models. In A. F. Veinott, Jr. (ed). *Mathematical Studies in Management Science.* Macmillan, New York, 313-321.

Veinott, A. F., Jr. (1974). Markov Decision Chains. In B. C. Eaves and G. B. Dantzig (eds.). *Studies in Optimization.* MAA Studies in Mathematics, Vol. 10, Mathematical Association of America, 124-159.

White, D. J. (1985). Real applications of Markov decision processes. *Interfaces* 15, 73-83.

White, D. J. (1988). Further real applications of Markov decision processes. *Interfaces* 18, 55-61.

White, D. J. (1993). A Survey of Applications of Markov Decision Processes. *Journal of the Operational Research Society* (UK) 44, 11, 1073-1096.

Articles

Balinski, M. (1961). *On Solving Discrete Stochastic Decision Problems. Study 2.* Mathematica, Princeton, NJ.

Beale, E. M. L. (1955). On Minimizing a Convex Function Subject to Linear Inequalities. *J. Royal Statist. Soc. Ser. B* 17, 2, 173-184.

Bellman, R. E. (1957). A Markovian Decision Process. *J. Math. Mech.* 6, 5, 679-684.

Bertsimas, D. and J. Niño-Mora (1996). Conservation Laws, Extended Polymatroids and Multiarmed Bandit Problems; A Polyhedral Approach to Indexable Systems. *Math. Operations Res.* 21, 2, 257-306.

Blackwell, D. (1962). Discrete Dynamic Programming. *Ann. Math. Stat.* 33, 2, 719-726.

Blackwell, D. (1967). Positive Dynamic Programming. In *Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I.* Univ. of Calif. Press, Berkeley, Calif., 415-418.

Brown, B. (1965). On the Iterative Method of Dynamic Programming on a Finite Space Discrete Time Markov Process. *Ann. Math Stat.* 36 4, 1279-1285.

Cantaluppi, L. J. (1984a). Optimality of Piecewise-Constant Policies in Semi-Markov Decision Chains. *SIAM J. Control Optim.* 22, 5, 723-739.

Cantaluppi, L. J. (1984b). Computation of Optimal Policies in Discounted Semi-Markov Decision Chains. *OR Spektrum* 6, 3, 147-160.

Chatwin, R. E. (1992). *Optimal Airline Overbooking.* Ph.D. Dissertation, Department of Operations Research, Stanford University, Stanford, CA, 111 pp.

Dantzig, G. B. (1955). Optimal Solutions of a Dynamic Leontief Model with Substitution. *Econometrica* 23, 3, 295-302.

Dantzig, G. B. (1955). Linear Programming Under Uncertainty. *Man. Sci.* 1, 3-4, 197-206.

de Ghellinck, G. (1960). Les Problèms de Décisions Séquentielles. *Cahiers du Centre d'Etudes de Recherche Opérationnelle* 2, 2, 161-179.

Denardo, E. V. (1967). Contraction Mappings in the Theory Underlying Dynamic Programming. *SIAM Review* 9, 2, 165-177.

Denardo, E. V. (1969). *Computing a Bias-Optimal Policy in Discrete Time Markov Decision.* Report No. 9 Dept. of Admin. Sci., Yale Univ., New Haven, 21 pp. (Revised version of *Computing (g,w)-Optimal Policies in Continuous and Discrete Markov Programs*, (1968).)

Denardo, E. V. (1970). On Linear Programming in a Markov Decision Problem. *Man. Sci.* 16, 5, 281-288.

Denardo, E. V. and B. L. Fox (1968). Multichain Markov Renewal Programs. *SIAM J. Appl. Math.* 16, 3, 468-487.

Denardo, E. V. and B. L. Miller (1968). An Optimality Condition for Discrete Dynamic Programming with No Discounting. *Ann. Math. Stat.* 39, 4, 1220-1227.

D'Epenoux, F. (1963). A Probabilistic Production and Inventory Problem. *Man. Sci.* 10, 1, 98-108. (Translation of an article published in *Revue Francaise de Recherche Opérationnelle* 14 (1960)).

Derman, C. (1962). On Sequential Decisions and Markov Chains. *Man. Sci.* 9, 1, 16-24.

Derman, C. (1963). Stable Sequential Rules and Markov Chains. *J. Math. Anal. and Appl.* 6, 2, 257-265.

Derman, C. (1964). On Sequential Control Processes. *Ann. Math. Stat.* 35, 1, 341-349.

Derman, C. and R. Strauch (1965). A Note on Memoryless Rules for Controlling Sequential Control Processes. *Ann. Math. Stat.* 37, 1, 276-278.

Dynkin, E. B. (1963). The Optimum Choice of the Instants for Stopping a Markov Process. *Soviet Math.* (English translation of *Doklady Akad. Nauk.*) 4, 627-629.

Eaton, J. and L. Zadeh (1962). Optimal Pursuit Strategies in Discrete State Probabilistic Systems. *Transactions of ASME Series D, Journal of Basic Engineering* 84, 23-29.

Federgruen, A. (1984). Successive Approximation Methods for Solving Nested Functional Equations in Markov Decision Problems. *Math. Opns. Res.* 9, 3, 319-344.

Fox, B. L. and D. M. Landi (1968). An Algorithm for Identifying the Ergodic Subchains and Transient States of a Stochastic Matrix. *Commun. Assoc. Comp. Mach.* 11, 9, 619-621.

Gittens, J. C. and D. M. Jones (1974). A Dynamic Allocation Index for the Sequential Design of Experiments. In J. Gani, K. Sarkadi and I. Vince (Eds.) *Progress in Statistics.* European Meeting of Statisticians, 1972, 1, North Holland, Amsterdam, 214-266.

Hordijk, A. and L. C. M. Kallenberg (1984). Constrained Undiscounted Stochastic Dynamic Programming. *Math. Opns. Res.* 9, 2, 276-289.

Howard, R. A. and J. E. Matheson (1972). Risk Sensitive Markov Decision Processes. *Man. Sci.* 18, 356-369.

Kantorovitch, L. (1939). The Method of Successive Approximations for Functional Equations. *Acta Mathematica* 71, 63-97.

Katehakis, M. N. and A. F. Veinott, Jr. (1987). The Multi-Armed Bandit Problem: Decomposition and Computation. *Math. Operations Res.* 12, 2, 262-268.

Lanery, E. (1967). Etude Asymptotique des Systemes Markoviens à Commande. *Révue d'Informatique et Recherche Opérationnelle* 1, 5, 3-56.

Lippman, S. A. (1968). On the Set of Optimal Policies in Discrete Dynamic Programming. *J. Math. Anal. Appl.* 440-445.

Lippman, S. A. (1969). Criterion Equivalence in Discrete Dynamic Programming. *Opns. Res.* 17, 5, 920-922.

MacQueen, J. and R. Miller, Jr. (1960). Optimal Persistence Policies. *Opns. Res.* 8, 3, 362-380.

Manne, A. S. (1960). Linear Programming and Sequential Decisions. *Man. Sci.* 6, 3, 259-267.

Martin-Lof, A. (1967). Optimal Control of a Continuous-Time Markov Chain with Periodic Transition Probabilities. *Opns. Res.* 15, 5, 872-881.

Miller, B. L. (1968a). Finite State Continuous-Time Markov Decision Processes with an Infinite Planning Horizon. *J. Math. Anal. Appl.* 22, 552-569.

Miller, B. L. (1968b). Finite State Continuous-Time Markov Decision Processes with a Finite Planning Horizon. *Siam J. Control* 6, 2, 266-280.

Miller, B. L. and A. F. Veinott, Jr. (1969). Discrete Dynamic Programming with a Small Interest Rate. *Ann. Math. Stat.* 40, 2, 366-370.

Ornstein, D. (1969). On the Existence of Stationary Optimal Strategies. *Proc. Amer. Math. Soc.* 20, 2, 563-569.

Radner, R. (1955). The Linear Team: An Example of Linear Programming Under Uncertainty. *Proc. Second Symp. Linear Prog.* National Bureau of Standards, Washington, 381-396.

Radner, R. (1961). Evaluation of Information in Organizations. In *Proc. 4th Berkeley Symp.*

Radner, R. (1962). Team Decision Problems. *Ann. Math. Stat.* 33, 2, 857-881.

Riis, J. (1965). Discounted Markov Programming in a Periodic Process. *Opns. Res.* 13, 6, 920-929.

Rosenblatt, D. (1957). On the Graphs and Asymptotic Forms of Finite Boolean Relation Matrices and Stochastic Matrices. *NRLQ* 4, 2, 151-167.

Ross, I. and F. Harary (1955). Identification of the Liason Persons of an Organization Using the Structure Matrix. *Man. Sci.* 1, 251-258.

Rothblum, U. G. (1975a). Multivariate Constant Risk Posture. *J. Econ. Theory* 10, 3, 309-332.

Rothblum, U. G. (1975b). Algebraic Eigenspaces of Nonnegative Matrices. *Lin. Alg. Appl.* 12, 281-292.

Rothblum, U. G. (1975c). Normalized Markov Decision Chains I: Sensitive Discount Optimality. *Opns. Res.* 16, 785-795.

Rothblum, U. G. (1978). Normalized Markov Decision Chains II: Optimality of Nonstationary Policies. *SIAM J. Control* 15, 221-232.

Rothblum, U. G. (1984). Multiplicative Markov Decision Chains. *Math. Opns. Res.* 9, 1, 6-24.

Rothblum, U. G. and A. F. Veinott, Jr. (1992). *Markov Branching Decision Chains: Immigration-Induced Optimality.* Technical Report No. 45, Department of Operations Research, Stanford University, 100 pp.

Rothblum, U. G. and P. Whittle (1982). Growth Optimality for Branching Markov Decision Chains. *Math. Opns. Res.* 7, 4, 582-601.

Rykov, V. V. (1966). Markov Decision Processes with Finite State and Decision Spaces. *Theory Prob. Appl.* 112, 2, 302-311.

Shapley, L. S. (1953). Stochastic Games. *Proc. Nat. Acad. Sci.* 39, 1095-1100.

Sladky, K. (1974). On the Set of Optimal Controls for Markov Chains with Rewards. *Kybernetika* 10, 350-367.

Strauch, R. E. (1966). Negative Dynamic Programming. *Ann. Math. Stat.* 37, 4, 871-890.

Strauch, R. E. and A. F. Veinott, Jr. (1966). *A Property of Sequential Control Processes.* RM-4772-PR, The RAND Corporation, Santa Monica, Calif., 8 pp.

Tardos, É. (1986). A Strongly Polynomial Algorithm to Solve Combinatorial Linear Programs. *Opns. Res.* 34, 2, 250-256.

Tarski, A. (1955). A Lattice Theoretical Fixpoint Theorem and Applications. *Pac. J. Math.* 5, 285-309.

Veinott, A. F., Jr. (1966). On Finding Optimal Policies in Discrete Dynamic Programming with No Discounting. *Ann. Math. Stat.* 37, 5, 1284-1294.

Veinott, A. F., Jr. (1968). Discrete Dynamic Programming with Sensitive Optimality Criteria. Preliminary Report. *Ann. Math. Stat.* 39, 1372.

Veinott, A. F., Jr. (1969a). Discrete Dynamic Programming with Sensitive Discount Optimality Criteria. *Ann. Math. Stat.* 40, 5, 1635-1660.

Veinott, A. F., Jr. (1969b). Minimum Concave Cost Solution of Leontief Substitution Models of Multi-facility Inventory Systems. *Opns. Res.* 17, 2, 262-291.

Warshall, S. (1962). A Theorem on Boolean Matrices. *J. Assoc. Comp. Mach.* 9, 11-12.

Wolfe, P. and G. B. Dantzig (1962). Linear Programming in a Markov Chain. *Opns. Res.* 10, 5, 702-710.

Zachrisson, L. (1964). *Markov Games.* In M. Dresher L. S. Shapley, and A. W. Tucker (eds.). *Advances in Game Theory.* Princeton University Press, Princeton, N. J., 211-253.

Index of Symbols

appearing on more than two successive pages.
Syntax: symbol, page number, meaning.

Index

Acknowledgement. Mai Vu suggested the creation of this index. It is a pleasure to acknowledge her encouragement and substantial collaboration in its preparation.

Homework 1 Due April 11

1. Exercising a Put Option. Consider the problem of deciding when to exercise a *put option* to *sell* a stock at the *strike price* $s^* > 0$ during any of the next N days. Suppose that if the price of the stock on one day is $s \geq 0$, then the price the next day is sR where the *return* R is a random variable. Assume that the returns on successive days are independent and have the same distribution as a nonnegative random variable R whose expected value is finite. The goal is to maximize the expected net revenue from an option to sell the stock any time in the next N days at the strike price when the price with N days to expiration is s. Assume that if one exercises the option to sell the stock on a day at the strike price, then one first buys the stock that day at the market price. Let $r \equiv R - 1$ be the rate of return.

(a) Dynamic-Programming Recursion. Write down a dynamic-programming recursion for the maximum expected net revenue V_s^N when the price with N days to expiration is s.

(b) Nonpositive Expected Rate of Return. Show that if $\mathrm{E}r \leq 0$, then it is optimal not to exercise the option before expiration.

(c) Positive Expected Rate of Return. Determine the optimal exercise policy with N days to expiration by induction on N where $\mathrm{E}r > 0$. [*Hint*: First show that $V_0^N = s^*$ for each $N = 0, 1, \ldots$. Next observe that if the stock price falls from a positive level to 0, which occurs with probability $1 - p = \Pr(R = 0)$, then the stock price can never become positive again. For this reason, it suffices to consider the case of positive stock prices $s > 0$. For that case, rewrite the dynamic programming recursion in (a) to reflect this fact and then make the change of variables $W_s^N \equiv \frac{1}{s}(V_s^N - (s^* - s))$.]

2. Requisition Processing. Requisitions arrive independently at a data-processing center in periods $1, \ldots, N$. At the end of each period, the center manager decides whether to process any requisitions. If the manager decides to process in a period, he processes all previously unprocessed requisitions arriving in that and all prior periods. The manager must process requisition in a period in which the number of unprocessed requisitions exceeds $M > 0$. The cost of processing a batch of requisitions of any size is $K > 0$ and the processing time is negligible. The waiting cost incurred in each period that a requisition awaits processing but is not processed is $w > 0$. All requisitions, if any, are processed in period N. The problem is to determine the periods in which to process that minimize the combined expected costs of processing and waiting over N periods. Suppose that p_i is the known probability that $i = 0, 1, \ldots, b$ requisitions arrive in a period where $\sum_{i=0}^{b} p_i = 1$. The expected number of requisitions arriving in a period is positive.

(a) Dynamic-Programming Recursion. Write down a dynamic-programming recursion for the minimum expected costs that permits an optimal N-period processing policy to be determined.

(b) Optimality of Processing-Point Policy. Show that one optimal processing policy in period n has the following form: if s is the number of requisitions accumulated before processing at the end of period n, process in period n if $s > s_n$ and wait (do not process) if $s \leq s_n$. Show how to determine the constants $s_1, \ldots, s_N$. [*Hint*: Prove these facts by induction on n by showing that the minimum expected cost in periods $n, \ldots, N$ is increasing in s.]

(c) Optimal Preplanned Processing Periods. Answer part (a) under the hypothesis that all the processing times must be chosen before any requisitions arrive, that no requisitions are on hand initially, and that $M = \infty$. [*Hint*: It suffices to consider only N states.]

(d) Comparison of Policies in Parts (b) and (c). Will the minimum expected cost in part (a) (assuming no requisitions are on hand initially and $M = \infty$) be larger or smaller than that in part (c), or can one say? What administrative and/or computational advantages does the policy in part (c) have over that in part (b)?

3. Matrix Products. Let $M_1, \ldots, M_n$ be matrices with M_i having r_{i-1} rows and r_i columns for $i = 1, 2, \ldots, n$ and some positive integers $r_0, \ldots, r_n$. The problem is to choose the order of multiplying the matrices that will minimize the number of multiplications needed to compute the product $M_1 M_2 \cdots M_n$. Assume that matrices are multiplied in the "usual" way.

(a) Dynamic-Programming Recursion. Give a dynamic-programming recursion for finding the optimal order in which to multiply the matrices that requires $O(n^3)$ operations.

(b) Example. Find the optimal order in which to multiply the matrices where $n = 4$ and $(r_0 \;\; r_1 \;\; r_2 \;\; r_3 \;\; r_4) = (10 \;\; 30 \;\; 70 \;\; 2 \;\; 100)$.

(c) Comparison of Minimum and Maximum Number of Multiplications. Also solve the problem of part (b) where the objective is instead to maximize the number of multiplications. What is the ratio of the maximum to minimum number of multiplications?

4. Bridge Clearance. Consider a network of S cities. Between each pair (s, t) of cities, there is a road having a bridge whose clearance height is $h(s, t) \geq 0$ with $h(s, t) = h(t, s)$. A trucking firm wants to determine a route from each city to city S that maximizes the minimum bridge height along the route subject to the restriction that at most N cities are visited enroute (excluding the initial city). Give a dynamic-programming recursion for solving the problem.

Answers to Homework 1 Due April 11

1. Exercising a Put Option. Suppose that the (nonnegative) market price of a stock on any day is a product of its price on the preceding day and its *return* that day. Assume that the returns on successive days are independent and have the same distribution as another nonnegative random variable R whose expectation is finite. Call $r \equiv R - 1$ the *rate of return*. Let V_s^N be the maximum expected revenue from a put option to sell a stock at the strike price $s^* > 0$ when the market price N days before expiration of the option is $s \geq 0$.

(a) Dynamic-Programming Recursion. The V_s^N satisfy the dynamic-programming recursion

$$V_s^N = \max(s^* - s,\, \mathrm{E}V_{sR}^{N-1}) \tag{1}$$

for $N = 1, 2, \ldots$ where $V_s^0 = (s^* - s)^+$ for all $s \geq 0$.

(b) Nonpositive Expected Rate of Return. Suppose $\mathrm{E}r \leq 0$, so $\mathrm{E}R \leq 1$. We claim that it is optimal not to exercise the option until the expiration day, and then only if the price that day is below the strike price. It suffices to show by induction that $V_s^N = \mathrm{E}V_{sR}^{N-1}$ for $N > 0$ and for all s. To see this, observe from (1) for $N > 1$ and by definition for $N = 1$ that $V_s^{N-1} \geq s^* - s$ for each $s \geq 0$. Thus, because $\mathrm{E}R \leq 1$, it follows that $\mathrm{E}V_{sR}^{N-1} \geq s^* - s\mathrm{E}R \geq s^* - s$ for each $s \geq 0$. Thus, $V_s^N = \mathrm{E}V_{sR}^{N-1}$ for $N > 0$ as claimed.

(c) Positive Expected Rate of Return. Suppose $\mathrm{E}r > 0$, so $\mathrm{E}R > 1$. For this part, it is helpful to note that once one reaches the market price $s = 0$, the subsequent price Rs is also zero, so the only market price accessible from 0 is 0. Hence, when the market price is zero, it is optimal to exercise the put option to sell the stock at the strike price s^*, so $V_0^N = s^*$ for $N = 0, 1, \ldots$. It is convenient to restrict attention to positive market prices in the sequel. To that end, let $p = \Pr(R > 0)$ and, for any random variable W, let $\mathrm{E}_+W \equiv \mathrm{E}(W|R > 0)$. Then $\mathrm{E}V_{sR}^{N-1} = (1-p)s^* + p\mathrm{E}_+V_{sR}^{N-1}$, so the restriction of (1) to positive market prices can be rewritten as

$$V_s^N = \max(s^* - s,\, (1-p)s^* + p\mathrm{E}_+V_{sR}^{N-1}) \tag{2}$$

for $s > 0$ and $N = 1, 2, \ldots$ where $V_s^0 = (s^* - s)^+$ for all $s > 0$.

Subtract $s^* - s$ from both sides of (2) and make the change of variables $U_s^N = V_s^N - (s^* - s)$ for $N = 0, 1, \ldots$. Then (2) reduces to the equivalent system

$$U_s^N = \max(0, -s\mathrm{E}r + p\mathrm{E}_+U_{sR}^{N-1})$$

for $s > 0$ and $N = 1, 2, \ldots$ where $U_s^0 = (s - s^*)^+$ for all $s > 0$. Now make the second change of variables $W_s^N = \frac{1}{s}U_s^N$ for $s > 0$. Then the above equation becomes

$$W_s^N = \max(0, -\mathrm{E}r + p\mathrm{E}_+RW_{sR}^{N-1}) \tag{2}'$$

for $s > 0$ and $N = 1, 2, \ldots$ where $W_s^0 = (1 - \frac{s^*}{s})^+$ for $s > 0$. We claim that for each N, W_s^N is increasing and continuous in $s > 0$ with $\lim_{s\uparrow\infty} W_s^N = 1$ and $\lim_{s\downarrow 0} W_s^N = 0$. Certainly that is so for $N = 0$. Suppose it is so for $N - 1$ and consider N. Then $p\mathrm{E}_+ RW_{sR}^{N-1}$ is increasing and continuous in $s > 0$, and converges to $\mathrm{E}R$ as $s \uparrow \infty$ and to 0 as $s \downarrow 0$, whence the claim follows. Consequently, there is a largest $s = s_N$ such that $-\mathrm{E}r + p\mathrm{E}_+ RW_{sR}^{N-1} \le 0$. Thus, the maximum on the right side of $(2)'$ is $-\mathrm{E}r + p\mathrm{E}_+ RW_{sR}^{N-1}$ if $s > s_N$ and 0 if $s \le s_N$. Since (2) and $(2)'$ are equivalent, it follows that this rule is optimal with (2) as well, i.e., it is optimal to wait if $s > s_N$ and to exercise the option if $s \le s_N$. In short, *if the expected rate of return is positive and if N days remain until expiration of the option, then there is an exercise price s_N such that it is optimal to wait if the price is above s_N and to exercise the option if the price is below s_N.*

2. Requisition Processing. The state of the system is the number of requisitions on hand at the end of a period before processing and the actions are to *process* or to *wait*.

(a) Dynamic-Programming Recursion. Let C_s^n be the minimum expected cost in periods $n, \ldots, N$ when there are s requisitions on hand at the end of period n before processing in that period. Then $C_s^N = 0$ or K according as $s = 0$ or $s > 0$. Also, for each $1 \le n < N$,

$$C_s^n = \min\left[K + \textstyle\sum_t p_t C_t^{n+1},\ ws + \sum_t p_t C_{t+s}^{n+1}\right] \text{ for } 0 \le s \le M$$

and

$$C_s^n = K + \textstyle\sum_t p_t C_t^{n+1} \text{ for } M < s$$

where the first and second terms π_n and ω_s^n in brackets on the right-hand-side of the first equation correspond respectively to processing and to waiting.

(b) Optimality of Processing-Point Policy. We show first that C_s^n is increasing in s by induction on n. This is surely true for $n = N$. Thus suppose it is so for $n+1, \ldots, N$ and consider n. Then since C_{t+s}^{n+1} is increasing in s for each t and ws is increasing in s, ω_s^n is increasing in s. Thus, since π_n is independent of s, $C_s^n = \pi_n \wedge \omega_s^n \le \pi_n$ is increasing in s.

Now let $s_N = 0$ and for each $1 \le n < N$, let s_n be the largest integer s for which $\omega_s^n \le \pi_n$. Since ω_s^n is increasing in s, it follows that $\omega_s^n \le \pi_n$ if and only if $s \le s_n$. Thus one optimal policy in period n is to wait if $s \le s_n$ and to process if $s_n < s$.

(c) Optimal Preplanned Processing Periods. Let C_n be the minimum expected cost in periods $1, \ldots, n$ when the processing times must be chosen before any requisitions arrive, there are no requisitions on hand initially, $M = \infty$, and one processes in period n. Let $r = \sum_t t p_t$ be the expected number of requisitions arriving in a period. Without loss of generality, assume that $r > 0$ because in the contrary event there are never any requisitions to process. The expected

processing and waiting costs c_i incurred when all requisitions arriving in i consecutive periods are processed in the last of those periods is

$$c_i = K + wr\sum_{j=1}^{i}(j-1) = K + wr\frac{i(i-1)}{2}.$$

Then the C_n may be found from the (deterministic) forward equations ($C_0 \equiv 0$)

$$C_n = \min_{0\le m<n} [C_m + c_{n-m}],\ n = 1, \ldots, N.$$

Let m_n be the least m minimizing the right-hand-side of this equation. Then m_n is the next to last optimal period in which to process assuming that one processes in period n. Thus since one processes last in period N, one processes next to last in period m_N, 2nd from last in period m_{m_N}, etc.

(d) Comparison of Policies in Parts (b) and (c). Since the optimal solution in part (b) allows the decision maker to base his decisions on the information accumulated whereas that in part (c) does not, the minimum expected cost in part (b) will not exceed that in part (c). On the other hand, the solution in part (c) has the advantages that it requires less computational effort, it utilizes deterministic rather than random processing times (which makes scheduling easier), and it does not require one to keep records on the number of requisitions remaining unprocessed at the end of each period.

We now give a more precise comparison of the computational effort required to find the optimal policies in parts (b) and (c) under the assumption that $M = O(N)$, as seems reasonable to make the comparison fair. The numbers of comparisons, additions and multiplications required to find the policies in parts (b) and (c) without further exploiting the special structure of the problem is summarized in the table below.

	Comparisons	Additions	Multiplications
(b)	$O(N^2)$	$O(N^2b)$	$O(N^2b)$
(c)	$O(N^2)$	$O(N^2)$	$O(N)$

Comparison of Computational Effort to Find Solutions in (b) and (c)

Observe that finding the solution in (b) requires the same order of comparisons, $O(b)$ times as many additions and $O(Nb)$ times as many multiplications when compared to the number of operations required for the solution in (c). The factor $O(b)$ reduction in additions arises because the solution in (c) depends only on the expected number r of requisitions arriving in a period (which we take to be given) whereas that in (b) depends on the number $b+1$ of elements in the support of the distribution of the number of requisitions arriving in a period.

3. Matrix Products. Every intermediate matrix product must be of the form $M_{i+1} \cdots M_k$, and that matrix product must be a product of $M_{i+1} \cdots M_j$ and $M_{j+1} \cdots M_k$ for some $i < j < k$. Let m_{ik} be the minimum number of multiplications needed to compute $M_{i+1} \cdots M_k$. Thus, the minimum number of multiplications needed to compute $M_{i+1} \cdots M_j$ and $M_{j+1} \cdots M_k$ is $m_{ij} + m_{jk}$. Also, the number of multiplications needed to compute the product of those two matrices is $r_i r_j r_k$. Since j should be chosen optimally, one has the branching recursion ($m_{i,i+1} = 0$ for $0 \le i < n$)

$$m_{ik} = \min_{i<j<k} \left[m_{ij} + m_{jk} + r_i r_j r_k \right] \text{ for } 0 \le i < k \le n.$$

(b) Example. With $(r_0 \;\; r_1 \;\; r_2 \;\; r_3 \;\; r_4) = (10 \;\; 30 \;\; 70 \;\; 2 \;\; 100)$.

$$\begin{aligned}
m_{02} &= r_0 r_1 r_2 = 21000 \\
m_{13} &= r_1 r_2 r_3 = \;\; 4200 \\
m_{24} &= r_2 r_3 r_4 = 14000 \\
m_{03} &= \min \left[m_{01} + m_{13} + r_0 r_1 r_3,\; m_{02} + m_{23} + r_0 r_2 r_3 \right] \\
&= \min \left[4200 + 600, 21000 + 14000 \right] = 4800. \\
m_{14} &= \min \left[m_{12} + m_{24} + r_1 r_2 r_4,\; m_{13} + m_{34} + r_1 r_3 r_4 \right] \\
&= \min \left[14000 + 210000,\; 4200 + 6000 \right] = 10200. \\
m_{04} &= \min \left[m_{01} + m_{14} + r_0 r_1 r_4,\; m_{02} + m_{24} + r_0 r_2 r_4,\; m_{03} + m_{34} + r_0 r_3 r_4 \right] \\
&= \min \left[10200 + 30000,\; 21000 + 14000 + 70000,\; 4800 + 2000 \right] = 6800.
\end{aligned}$$

The minimum-multiplication order of matrix-multiplication is thus $(M_1(M_2M_3))M_4$.

(c) Comparison of Minimum and Maximum Number of Multiplications. If M_{ik} is the maximum number of operations required to compute the matrix product $M_{i+1} \cdots M_k$, then $M_{ik} = m_{ik}$ for $k - i \le 2$. Also $M_{03} = 22400$, $M_{14} = 224000$, and

$$M_{04} = \max \Big[224000 + 30000,\; 21000 + 14000 + 70000,\; 22400 + 2000 \Big] = 254000.$$

The maximum-multiplication order of matrix-multiplication is thus $M_1(M_2(M_3M_4))$. Also the ratio of the maximum to the minimum number of multiplications is

$$\frac{M_{04}}{m_{04}} = 37.35!$$

4. Bridge Clearance. Let V_s^n be the maximum of the minimum bridge heights along routes from s to S among those that visit at most n cities. Set $V_S^n \equiv \infty$ for $n = 1, 2, \ldots$. Then

$$V_s^1 = h(s, S)$$

and

$$V_s^n = \max_{t \ne s} h(s,t) \wedge V_t^{n-1}$$

for $s = 1, \ldots, S-1$ and $n = 2, \ldots, N$.

Revised Homework 2 Due April 18

(Do any three problems—problems 2-4 if you took MS&E 251.)

1. Dynamic Portfolio Selection with Constant Multiplicative Risk Posture. A financial planner seeks a strategy for dynamically revising a portfolio of M securities that she manages for a client to maximize the expected utility of distributions to him over n years. At the beginning of each year $N < n$, the planner observes the market value (in dollars) $s > 0$ of her client's portfolio, distributes a portion $p_0 s$ to him in cash, and invests fractions $p_1, \ldots, p_M$ of the remaining funds $(1 - p_0)s$ in securities $1, \ldots, M$. The *distribution fraction* p_0 and *portfolio* $p \equiv (p_1, \ldots, p_M)$ satisfy $0 < p < 0 \leq 1$, $p \geq 0$ and $\sum_{i=1}^{M} p_i = 1$. At the beginning of the last year n, the planner distributes the entire market value of the portfolio in cash. There are no transaction costs for buying or selling securities. Each dollar she invests in security i in year N is worth $R_i^N > 0$ dollars at the end of the year. The joint distribution of $R^N \equiv (R_1^N, \ldots, R_M^N)$ in each year N is known, the random vectors $R^1, \ldots, R^n$ are independent, and R_i^N and $\ln R_i^N$ have finite expectations for all i, N. Her client's *utility* of distributing $w^1, \ldots, w^n \geq 0$ to him in years $1, \ldots, n$ is $\sum_{N=1}^{n} \ln w^N$.

(a) Dynamic-Programming Recursion. Give a dynamic-programming recursion satisfied by the maximum expected utility $V^N(s)$ of distributions to the client in years $N, \ldots, n$ when the market value of the client's portfolio at the beginning of year N is s.

(b) Maximum Expected Utility has Constant Multiplicative Risk Posture. Show by induction on N that $V^N(s) = a^N \ln s + b^N$ for some constants a^N and b^N, and show how to evaluate the constants.

(c) Optimal Distribution Fraction and Portfolio. Conclude that the optimal policy in year $N < n$ is to choose the distribution fraction $p_0 = 1/(n-N+1)$ (the reciprocal of the number of years remaining) and portfolio p that maximizes $\mathrm{E} \ln(pR^N)$ (invest myopically). Note that both p_0 and p are independent of the value of the portfolio at the beginning of year N.

2. Airline Overbooking. An airline seeks a reservation policy for a flight with S seats that maximizes its expected profit from the flight. Reservation requests arrive hourly according to a Bernoulli process with p being the probability of a reservation request during an hour. A passenger with a reservation pays the fare $f > 0$ at flight time. If $b \geq 0$ passengers with reservations are denied boarding at flight time, they do not pay the fare and the airline pays them a penalty $c(b)$ (divided among them) where c is increasing on the nonnegative integers and $c(0) = 0$. Consider the N^{th} hour ($N > 0$) before flight time. At the beginning of the hour, the airline reviews the number of reservations on hand and decides whether to *book* (accept) or decline any reservation request during the hour. Each passenger with a reservation after any reservation request is booked or declined, but before cancellations, independently cancels the reservation with proba-

bility q during the hour. For this reason, the airline is considering the possibility of overbooking the flight to compensate for cancellations. Let V_r^N be the maximum expected future profit when r seats have been booked *before* reservation requests *and* cancellations during the hour. Let v_r^N be the maximum expected future profit when r seats have been booked *after* booking or declining any reservation request, but *before* cancellations, during the hour

(a) Dynamic-Programming Recursion. Give a dynamic-programming recursion from which an optimal reservation policy can be determined.

(b) Optimality of Booking-Limit Policies. Assume, as can be shown, that if g is quasiconcave on the integers, then $\mathrm{E}g(B_r)$ is quasiconcave in r where B_r is a sum of r independent identically distributed Bernoulli random variables. Use this result to show the following facts. First, v_r^N is quasiconcave in r, i.e., there is a nonnegative integer *booking limit* b^N such that v_r^N is increasing in r on $[0, b^N]$ and decreasing in r on $[b^N, \infty)$. Second, V_r^N is quasiconcave in r and achieves its maximum at $r = b^N$. Third, a booking-limit policy is optimal, i.e., accept a reservation request when N hours remain before flight time if and only if $r < b^N$.

(c) Solving the Problem with MATLAB. Assume that $c(b) = fb$. Graph V_r^N for $0 \le N \le 30$ and $0 \le r \le 20$ by using MATLAB[1] with the following parameter values: $S = 10$, $f = \$300 + F$, $p = .2$ and $.3$, and $q = .05$ and $.10$, where F is the sum of the digits representing the positions in the alphabet of the first letters of your first and last names. For example, Arthur Veinott would have $F = 1 + 22 = 23$ because A is the first and V is the 22^{nd} letter of the alphabet. In each case, estimate the booking limit ten hours before flight time from your graphs. Discuss whether your graphs confirm the claim in (b) that V_r^N is quasiconcave in r. What conjectures do the graphs that you found suggest about the optimal reservation policy and/or maximum expected reward and their variation with the various data elements? You will lose points on your conjectures only if your graphs are inconsistent with or do not support your conjectures or if you don't make enough interesting conjectures. The idea here is to brainstorm intelligently.

3. Optimal Capacitated Supply Policy. A manager seeks a supply policy for a single product that minimizes the expected costs of ordering, storage and shortage over n periods. The demands $D_1, \ldots, D_n$ for the product in periods $1, \ldots, n$ are independent real-valued random variables. At the beginning of period N, the supply manager observes the possibly negative *initial stock* $x \in \Re$ on hand. If $x < 0$, $-x$ units are backlogged. The supply manager then orders a nonnegative

[1]To develop the graphs required for this part and the next one, you need to run the *MATLAB* m-file *airlin27.m.* This file calls the files *finhor3r.m* and *booklim3.m.* To do this, download these files from the MS&E 351 course directory. *MATLAB* is available on the UNIX clusters in Terman and Gates. Printers are available in both places for a fee. You can also run *MATLAB* remotely. Of course, if you own *MATLAB,* you can run it on your personal computer as well.

quantity at unit cost $c \geq 0$ with immediate delivery, bringing the *starting stock* after receipt of the order to y, $x \leq y \leq x + u$ where $u > 0$ is the given capacity. Unsatisfied demands are backlogged so the stock on hand at the end of period N is $y - D_N$. There is a convex storage and shortage cost function $g(z)$ where z is the (possibly negative) stock on hand at the end of a period. Assume that $G_N(y) \equiv \mathrm{E}g(y - D_N)$ is finite for all y, $G_N(y) \to \infty$ as $y \to \infty$ and $cy + G_N(y) \to \infty$ as $y \to -\infty$. Let $C_N(x)$ be the minimum expected cost in periods $N, \ldots, n$ starting with the initial stock x in period N. Assume that $C_N(x)$ is finite for all x. The terminal cost $C_{n+1}(\,\cdot\,)$ at the end of period n is bounded below. Assume, as can be shown, that all convex functions arising in this problem are continuous and that convex functions are closed under addition and positive scalar multiples.

(a) Dynamic-Programming Recursion. Give a dynamic-programming recursion for calculating $C_N(x)$.

(b) Convexity of $C_N(x)$. Show that if the terminal cost is convex, then $C_N(x)$ is convex and bounded below in x. [*Hint*: Use part (d) below.]

(c) Optimality of Constrained Base-Stock Policies. Show that one optimal policy $y_N(x)$ in period N has the form $y_N(x) = (y_N^* \vee x) \wedge (x + u)$ for some real number y_N^*.

(d) Projections of Convex Functions. Suppose that $f(x, y)$ is a real-valued convex function on a nonempty closed convex set $W \subseteq \Re^k$ and let $F(x) \equiv \inf_{(x,y)\in W} f(x, y)$. Show that if F is finite on $X \equiv \{x : (x, y) \in W\}$, then F is convex thereon.

4. Sequencing: Optimality of Index Policies. Consider a sequence $1, \ldots, n$ of n tasks. When a task is *executed*, it *succeeds* or *fails*. Given a permutation of the tasks, execute them in turn until some task succeeds or until all tasks fail, whichever occurs first. The goal is to choose the order of executing the tasks that minimizes the resulting expected cost. Let $\sigma = (\sigma_1, \ldots, \sigma_n)$ be a permutation of the tasks. Assume that the cost c_{σ_i} of executing task σ_i depends on σ_i, but not the other tasks. Let $p_{\sigma_i|\sigma_{i-}}$ be the conditional probability that task σ_i fails given that the tasks $\sigma_{i-} \equiv (\sigma_1, \ldots, \sigma_{i-1})$ fail. Assume that $p_{\sigma_i|\sigma_{i-}}$ has the property that $1 - p_{\sigma_i|\sigma_{i-}} = f_{\sigma_i} \cdot g_{\sigma_{i-}}$ for some positive functions f and g. Let $p_{\sigma_{i-}}$ be the (unconditional) probability that the tasks σ_{i-} fail. Assume that $p_{\sigma_{i-}}$ is *symmetric*, i.e., is independent of the order of executing tasks $\sigma_1, \ldots, \sigma_{i-1}$.

(a) Optimality of Index Policies by Interchanging Tasks. Show that there is an *index* function $I(\,\cdot\,)$ such that it is optimal to choose any permutation $\sigma = (\sigma_1, \ldots, \sigma_n)$ for which $I(\sigma_i)$ is increasing in i. [*Hint*: For an arbitrary permutation $\sigma = (\sigma_1, \ldots, \sigma_n)$ of the tasks, consider when it is optimal not to interchange the order of executing tasks σ_i and σ_{i+1}.]

Show how to apply the result of problem (a) to solve each of the following two problems.

(b) Testing a Machine to Find a Fault. Suppose that a machine does not operate properly and n independent tests $1, \ldots, n$ can be performed to find the cause. Test i costs c_i and has probability p_i of failing to find the cause given that the tests preceding it fail to find the cause. Once the cause is found, no further tests are executed. All tests may fail to find the cause. What is the minimum-expected-cost sequence to carry out the tests?

(c) Search for a Plane. Suppose a plane is down in one of n locations $1, \ldots, n$ with known positive probabilities $q_1, \ldots, q_n$, $\sum_i q_i = 1$. Denote by t_i the time to search location i to find the plane or determine that it is not there. What search sequence minimizes the time to find the plane?

Answers to Homework 2 Due April 18

1. Dynamic Portfolio Selection with Constant Multiplicative Risk Posture.

(a) Dynamic-Programming Recursion. The dynamic-programming recursion is

$$V^N(s) = \max_{\substack{0<p_0<1\\ p\in P}} \{\ln(sp_0) + \mathrm{E}V^{N+1}(s(1-p_0)pR^N)\},\ N = 1,\dots,n-1,\ s>0 \tag{1}$$

where P is the set of row M-vectors $\{p : \sum_{i=1}^M p_i = 1,\ p \geq 0\}$, $R^N = (R_i^N)$ is a column M-vector and $V^n(s) = \ln s$ for $s > 0$.

(b) Maximum Expected Utility has Constant Multiplicative Risk Posture.[1] It suffices to show by induction on N that

$$V^N(s) = a^N \ln s + b^N \tag{2}$$

where $a^n \equiv 1$ and $b^n \equiv 0$, and for $N = 1,\dots,n-1$,

$$a^N \equiv n-N+1 = 1 + a^{N+1} \tag{3}$$

and

$$b^N \equiv \max_{0<p_0<1} \{\ln p_0 + a^{N+1}\ln(1-p_0)\} + a^{N+1}\max_{p\in P} \mathrm{E}\ln(pR^N) + b^{N+1}. \tag{4}$$

To see this, observe that (2) certainly holds for $N = n$. Suppose (2) holds for $N+1 \leq n$. Then on substituting (2) for $N+1$ into the right-hand side of (1) and using (3) and (4), it follows that (2) holds for N.

(c) Optimal Distribution Fraction and Portfolio. It follows from (4) that the optimal distribution fraction $p_0 = p_0^N$ in year N maximizes the first term on the right-hand side of (4). Maximizing that term and using (3) yields $p_0^N = 1/a^N = 1/(n-N+1)$. Similarly, it follows from (4) that the optimal portfolio $p = p^N$ in each year $1 \leq N < n$ maximizes the second term on the right-hand side of (4), viz., maximizes $\mathrm{E}\ln(pR^N)$ subject to $p \in P$. Note that the optimal distribution fraction and portfolio in each year are both independent of the level of wealth at the beginning of the year. Also, the optimal portfolio in each year is *myopic*, i.e., maximizes the expected instantaneous rate of return for that year alone without regard for subsequent investment opportunities.

[1]Incidentally, note that the client's utility function $u(w) = \sum_{N=1}^n \ln w^N$ for distributions $w = (w^N) \gg 0$ in the n years exhibits constant multiplicative risk posture. Indeed, up to positive affine transformations, it is the only such utility function that is also symmetric and strictly risk averse.

2. Airline Overbooking.

(a) Dynamic-Programming Recursion. Let V_r^N be the maximum expected profit that can be earned from a flight when $r \geq 0$ seats have been booked N hours before flight time. Then the profit at flight time ($N = 0$) is

$$V_r^0 = v_r^0 \equiv fr \wedge [fS - c((r-S)^+)].$$

When $N = 1, 2, \ldots$ hours remain before flight time, the airline can either *book*, i.e., accept, or *decline* a new reservation request during the ensuing hour. Reflecting these two possibilities, the maximum expected profit satisfies the recursion

$$V_r^N = \max\{v_r^N, (1-p)v_r^N + pv_{r+1}^N\} \tag{1}$$

where $v_r^N \equiv \mathrm{E}V_{C_r}^{N-1}$ and C_r is the number of the r reservations that do not cancel during the hour. Then C_r has a binomial distribution with parameters r and $1-q$. Interpret v_r^N as the maximum expected profit when there are r reservations after booking or declining new requests, but before cancellations, with N hours remaining until flight time.

(b) Optimality of Booking-Limit Policies. We prove by induction on N that V_r^N and v_r^N are quasiconcave in r, and a booking-limit policy is optimal with N hours to flight time and *booking limit* b^N that is a maximizer of v_r^N. This is true for $N = 0$ with booking limit $b^0 = S$ because $V_r^0 = v_r^0 = fr$ is increasing in $r \leq S$ and $V_r^0 = v_r^0 = fS - c((r-S)^+)$ is decreasing in $r \geq S$. Suppose that $N - 1 \geq 0$ and V_r^{N-1} is quasiconcave in r. Then $v_r^N = \mathrm{E}V_{C_r}^{N-1}$ is quasiconcave in r by the first assertion in the statement of (b). It is clear from (1) that $r = b^N$ also maximizes V_r^N and $V_{b^N}^N = v_{b^N}^N$. Thus since v_r^N is increasing in $r \leq b^N$, the same is so of $V_r^N = (1-p)v_r^N + pv_{(r+1)\wedge b^N}^N$. Similarly, since v_r^N is decreasing in $r \geq b^N$, the same is so of $V_r^N = v_r^N$. Thus V_r^N is quasiconcave in r. Also, a booking-limit policy is optimal with N hours to flight time and booking limit b^N.

(c) Solving the Problem with MATLAB. The graphs of the V_r^N and the corresponding booking limits b^N for $N = 0, \ldots, 30$ and $r = 0, \ldots, 20$ with $S = 10$ seats and $f = \$300$ per seat are attached for $(p, q) = (.2, .05)$, $(.2, .10)$, $(.3, .05)$, $(.3, .10)$. The booking limits 10 hours before flight time are respectively 15, 20, 15, 20. The graphs do confirm the conjecture in (b) that V_r^N is quasiconcave in r for each N. The graphs are interpolated linearly between successive integer values of r to define V_r^N for fractional values of r. The graphs suggest the following additional conjectures:

- **Monotonicity of Booking Limits.** b^N is increasing in $(N, -p, q)$ and $b^0 = S$.
- **Equilibrium Reservation Level.** There is a (possibly fractional) equilibrium reservation level $e \leq S$ such that $V_e^N = V_e$ is independent of N. The equilibrium appears to be roughly

the reservation level e where the expected number of new reservation requests equals the expected number of cancellations, viz., $p = (e + p)q$ or $e = p(1 - q)/q \leq p/q$. This formula is probably valid only for the case where e is significantly smaller than S. Also, e is increasing in p and decreasing in q. Finally, if the number of reservations is less than e, the expected revenue from the flight never exceeds $V_e = V_e^0 = ef$, regardless of the number of hours before flight time.

• **Monotonicity/Concavity/Quasiconcavity of Maximum Value.** V_r^N is concave in r, increasing in N for $0 \leq r \leq e$, decreasing in N for $e \leq r \leq S$, quasiconcave in N for $S \leq r$, increasing in p, and decreasing in q for $0 \leq r \leq S$. In fact, if e is sufficiently below S, it appears that V_r^N is essentially linear in r for r sufficiently below S. This together with the fact $V_e^N = ef$ from the preceding paragraph implies that $V_r^N \approx [(1 - q)^N(r - e) + e]f$ for r sufficiently below S.

• **Monotonicity of Differences of Maximum N-Period Values.** For each $0 \leq N$ and $0 < k$, let $\Delta_k V_r^N \equiv V_r^{N+k} - V_r^N$. Let u_k^N be the largest number r such that $\Delta_k V_r^N = 0$. Then $b^N \leq u_k^N \leq b^{N+k}$. Also, $\Delta_k V_r^N$ is positive for $r < e$, negative for $e < r < u_k^N$ and positive for $u_k^N < r$.

• **Limiting Maximum Value Independent of Initial Reservations.** Then $\lim_{N\to\infty} V_r^N$ exists and equals V_e for all r.

3. Optimal Capacitated Supply Policy.

(a) Dynamic-Programming Recursion. The desired dynamic-programming recursion is

$$C_N(x) = \min_{x \leq y \leq x+u} [c \cdot (y - x) + G_N(y) + \mathrm{E}C_{N+1}(y - D_N)]$$

for $x \in \Re$ and $N = 1, 2, \ldots, n$ where $C_{n+1}(x)$ is given.

(b) Convexity of $C_N(x)$. We claim that $C_N(x)$ is convex and bounded below in x by induction on N. The claim is so by hypothesis for $N = n + 1$. Suppose it holds for $N + 1 \leq n + 1$ and consider N. Then each term in the minimand on the right side of the recursion in (a) is convex and bounded below on the convex feasible set $x \leq y \leq x + u$. Thus, by (d), $C_N(x)$ is convex and bounded below.

(c) Optimality of Constrained Base-Stock Policies. We claim that one optimal policy $y_N(x)$ in period N has the form $y_N(x) = (y_N^* \vee x) \wedge (x + u)$ for some real number y_N^*. To see this, let $J_N(y) \equiv cy + G_N(y) + \mathrm{E}C_{N+1}(y - D_N)$ and $y = y_N^*$ be a minimizer of $J_N(y)$. The existence of such a minimizer is assured because $\mathrm{E}C_{N+1}(y - D_N)$ is bounded below in y, $cy + G_N(y) \to \infty$ as $y \to -\infty$, $G_N(y) \to \infty$ as $y \to \infty$, and $J_N(y)$ is convex on the real line, and so continuous. Also,

$y_N(x)$ minimizes $J_N(y)$ subject to $x \le y \le x+u$. To see this, it suffices to consider the following cases:

- $x+u < y_N^*$. Since $J_N(y)$ is decreasing in $y \le y_N^*$, $y_N(x) = x+u$.
- $x \le y_N^* \le x+u$. Since $y = y_N^*$ minimizes $J_N(y)$, $y_N(x) = y_N^*$.
- $y_N^* \le x$. Since $J_N(y)$ is increasing in $y \ge y_N^*$, $y_N(x) = x$

(d) Projections of Convex Functions. Let $f(x,y)$ be a real-valued convex function on a non-empty closed convex set $W \subseteq \Re^k$ and let $F(x) \equiv \inf_{(x,y)\in W} f(x,y)$. We show that if F is finite on the set $X \equiv \{x : (x,y) \in W\}$, then F is convex thereon. To see this, let $\epsilon > 0$. Suppose $p, p' \ge 0$; $p+p' = 1$; and $x, x' \in X$. Choose y, y' such that (x,y), $(x',y') \in W$, $f(x,y) \le F(x) + \epsilon$ and $f(x',y') \le F(x') + \epsilon$. Then

$$F(px + p'x') \le f(p \cdot (x,y) + p' \cdot (x',y')) \le pf(x,y) + p'f(x',y') \le pF(x) + p'F(x') + \epsilon.$$

Since this is so for all $\epsilon > 0$, it is so for $\epsilon = 0$, whence F is convex.

4. Sequencing: Optimality of Index Policies. Let σ be a sequence of tasks and $C(\sigma)$ be its associated expected cost. Without loss of generality, assume that $\sigma = (1, \ldots, n)$, i.e., perform the tasks in numerical order. Let $1- = 0$ and $i- = (1, \ldots, i-1)$ for $1 < i < n$.

(a) Optimality of Index Policies by Interchanging Tasks. Fix $1 \le i < n$ and let σ' be the sequence formed from σ by interchanging tasks i and $i+1$. Let C_{i-} be the expected cost of tasks $i-$. Let C^j be the expected cost of tasks $j, \ldots, n$ for $1 \le j \le n$. Then ($C_0 \equiv C^{n+1} \equiv 0$)

$$C(\sigma) = C_{i-} + p_{i-}(c_i + p_{i|i-}c_{i+1}) + C^{i+2}$$

and

$$C(\sigma') = C_{i-} + p_{i-}(c_{i+1} + p_{i+1|i-}c_i) + C^{i+2}.$$

Consequently, on comparing the above formulas, it follows that $C(\sigma) \le C(\sigma')$ if and only if

$$c_i + p_{i|i-}c_{i+1} \le c_{i+1} + p_{i+1|i-}c_i,$$

or equivalently,

$$\frac{c_i}{1 - p_{i|i-}} \le \frac{c_{i+1}}{1 - p_{i+1|i-}}.$$

Thus since

$$(*) \qquad 1 - p_{j|i-} = f_j \cdot g_{i-}$$

for each $j \ge i$, the above inequality simplifies to

$$\frac{c_i}{f_i} \le \frac{c_{i+1}}{f_{i+1}}.$$

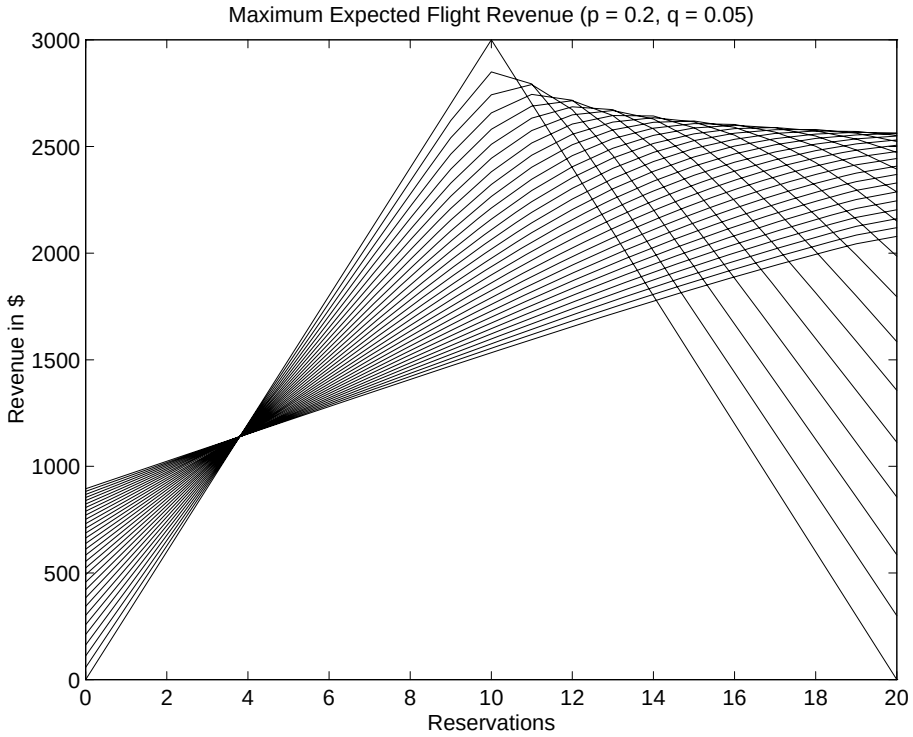
Consider the index rule in which the tests, after possibly renumbering, are carried out in an order that assures that the index function $I_i \equiv \frac{c_i}{f_i}$ is increasing in i. We claim that this index rule is optimal. For in any other sequence σ, there must be a first test with higher index than its successor. Let σ' be the sequence in which one interchanges these two tests. The above argument shows that $C(\sigma') \leq C(\sigma)$. Also, σ' is *lexicographically smaller* than σ, i.e., the first nonzero element of $\sigma - \sigma'$ is positive. Repeating this process up to $\binom{n}{2}$ times produces the index rule because each sequence is lexicographically smaller than its predecessor and lexicographic order is a total (or linear) order of the sequences. Thus, the index rule is optimal.

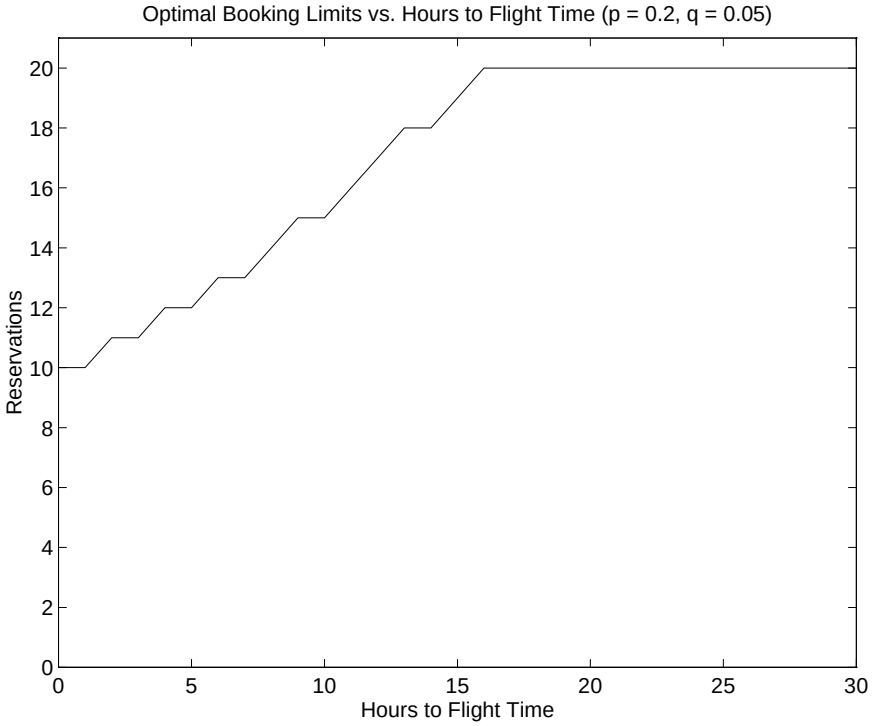
(b) Testing a Machine to Find a Fault. Let σ be a permutation of the tests, $g_{i-} = 1$, $f_j = 1 - p_j$ and $p_{i-} \equiv p_1 \cdots p_{i-1}$ for all $i, j \geq 1$. Note that $(*)$ holds and that g_{i-} and p_{i-} are symmetric in $i-$. Thus, the hypotheses of part (a) are satisfied. Hence from part (a), one optimal sequence σ in which to perform the tests is so that the *cost-success ratio* $\frac{c_i}{1-p_i}$ is increasing in i.

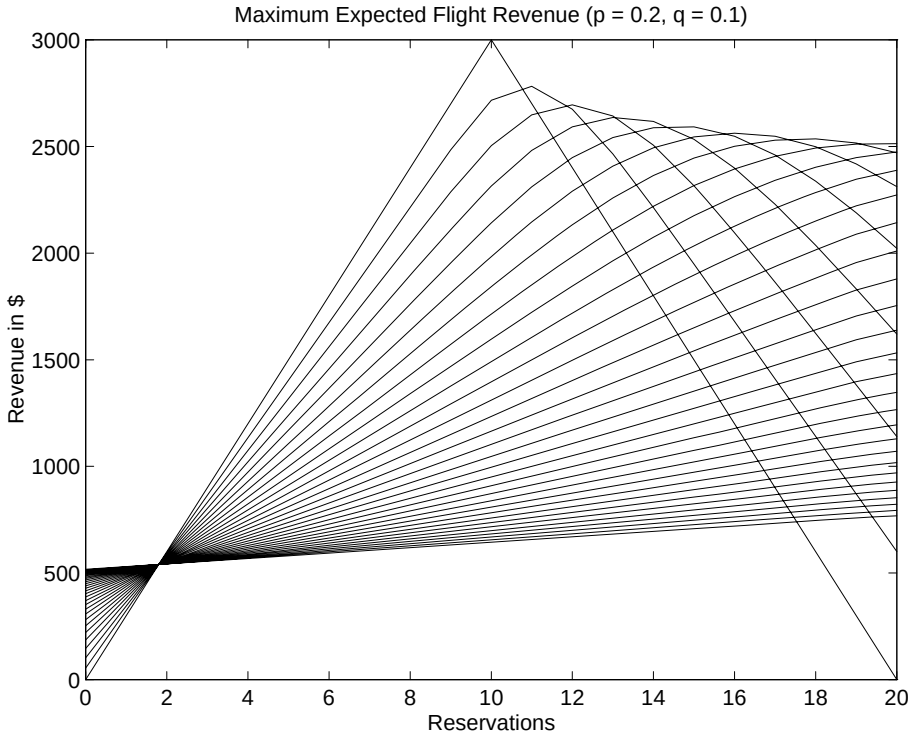
(c) Search for a Plane. Let σ be a permutation of the locations, $p_{i-} \equiv 1 - q_1 - \cdots - q_{i-1}$, $g_{i-} = \frac{1}{p_{i-}}$, $f_j = q_j$ and $c_j = t_j$ for all $i, j \geq 1$. Note that $(*)$ holds and that g_{i-} and p_{i-} are symmetric in $i-$. Thus, the hypotheses of part (a) are satisfied. Hence from part (a), one optimal sequence σ in which to search the locations is so that the *time-success ratio* $\frac{t_i}{q_i}$ is increasing in i.

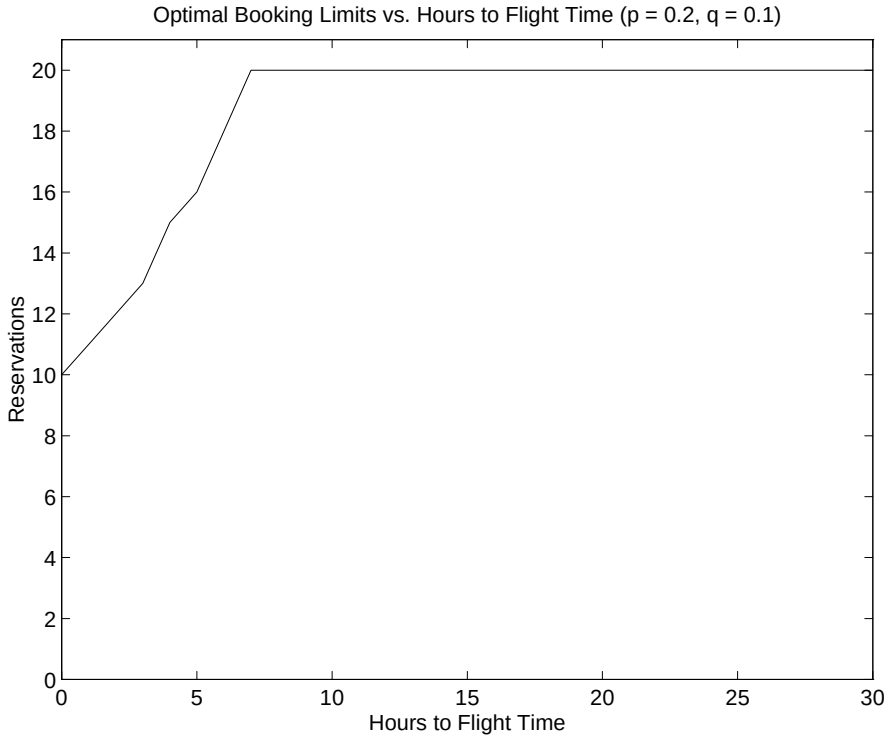
Remark. The function g can be dispensed with if one replaces the condition $(*)$ by the equivalent condition that

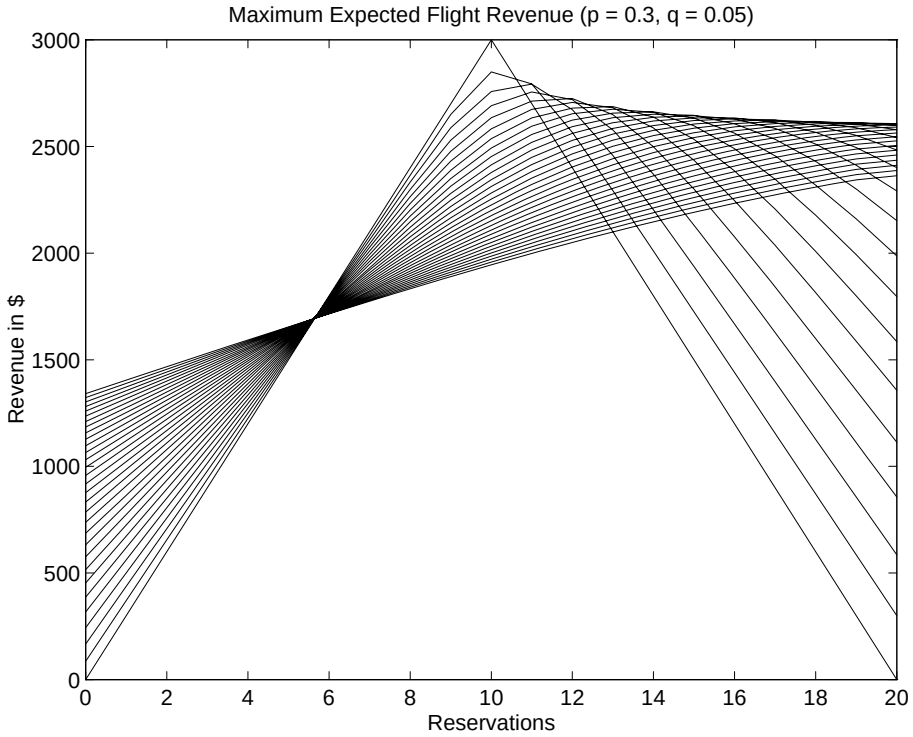
$$\frac{1 - p_{k|i-}}{1 - p_{j|i-}} = \frac{f_k}{f_j} \text{ for all } k, j \geq i.$$

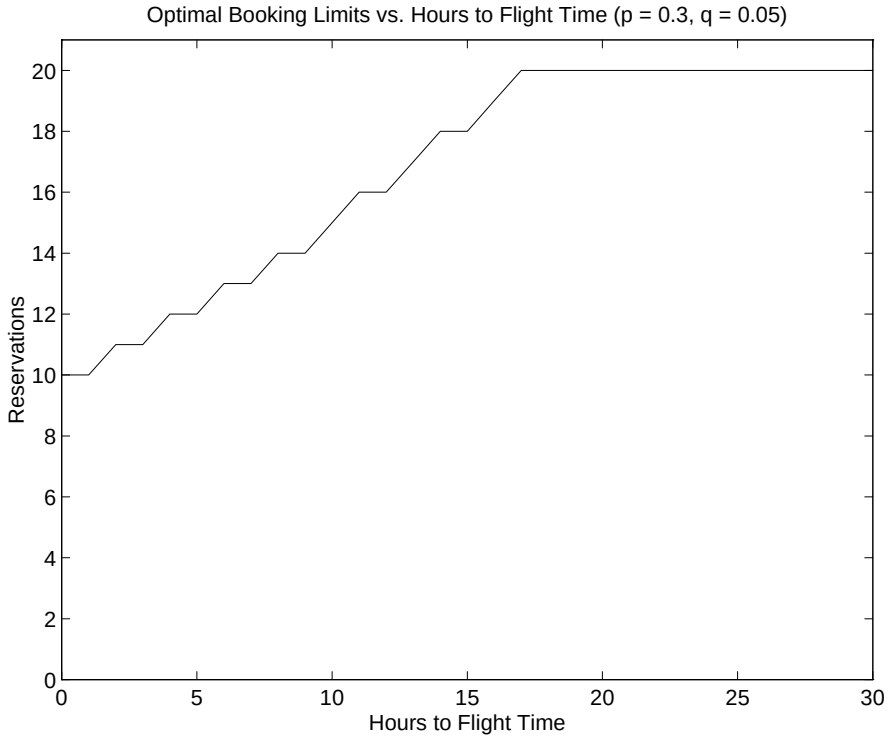

Maximum Expected Flight Revenue (p = 0.2, q = 0.05)
Revenue in $
Reservations
0
500
1000
1500
2000
2500
3000
0
2
4
6
8
10
12
14
16
18
20

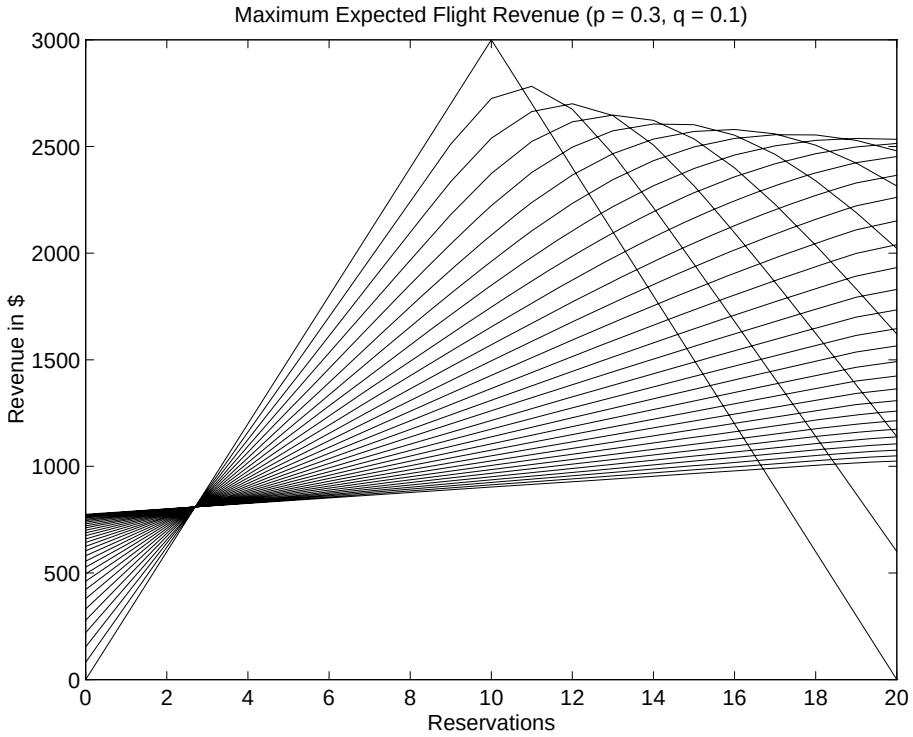

Optimal Booking Limits vs. Hours to Flight Time (p = 0.2, q = 0.05)
Reservations
Hours to Flight Time
0
2
4
6
8
10
12
14
16
18
20
0
5
10
15
20
25
30

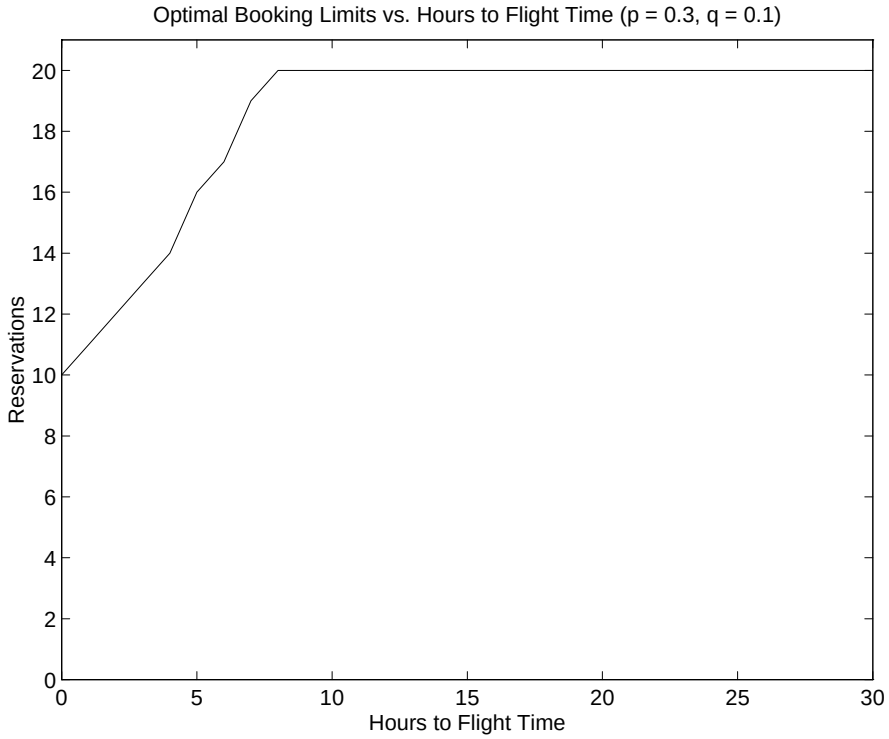

Maximum Expected Flight Revenue (p = 0.2, q = 0.1)
Revenue in $
Reservations
0
500
1000
1500
2000
2500
3000
0
2
4
6
8
10
12
14
16
18
20


Optimal Booking Limits vs. Hours to Flight Time (p = 0.2, q = 0.1)
Reservations
Hours to Flight Time
0
2
4
6
8
10
12
14
16
18
20
0
5
10
15
20
25
30


Maximum Expected Flight Revenue (p = 0.3, q = 0.05)
Revenue in $
Reservations
0
500
1000
1500
2000
2500
3000
0
2
4
6
8
10
12
14
16
18
20


Optimal Booking Limits vs. Hours to Flight Time (p = 0.3, q = 0.05)
Reservations
Hours to Flight Time
0
2
4
6
8
10
12
14
16
18
20
0
5
10
15
20
25
30


Maximum Expected Flight Revenue (p = 0.3, q = 0.1)
Revenue in $
Reservations
0
500
1000
1500
2000
2500
3000
0
2
4
6
8
10
12
14
16
18
20


Optimal Booking Limits vs. Hours to Flight Time (p = 0.3, q = 0.1)
Reservations
Hours to Flight Time
0
2
4
6
8
10
12
14
16
18
20
0
5
10
15
20
25
30

Homework 3 Due April 25

1. Multifacility Linear-Cost Production Planning. Suppose K facilities, labeled $1, \ldots, K$, each produce a single product. Production of each unit at facility k in period $n = 1, \ldots, N$ costs c_n^k and consumes $e_{mn}^{jk} \geq 0$ units at facility j in period m for each j, k and $1 \leq m < n$. The cost c_n^k includes the costs of in-transit storage and shipping e_{mn}^{jk} units of each product j in each period $m < n$, but excludes the cost of supplying those units of product j in period m. The unit cost of storage at facility k at the end of period n is h_n^k. There are known nonnegative demands for the product produced at each facility in each period. Let C_n^k be the minimum total cost of satisfying a unit of demand at facility k in period n. This unit cost includes all costs of production, storage and in-transit storage at all facilities involved in supplying that unit of demand at facility k in period n.

(a) Dynamic-Programming Recursion. Give a dynamic-programming recursion for solving the problem. [*Hint*: Express C_n^k in terms of the C_m^j for $m < n$. Use the fact that the demand at facility k in period n can be satisfied either by producing at k in n or producing at k in a prior period and storing from the previous period $n - 1$.]

(b) Running Time. How many multiplications, additions and comparisons are required to solve the problem? Your answer should be of the form $p(N, K) + O(q(N, K))$ for some polynomials p and q with p being of higher order than q.

(c) Running Time: Special Cases. How long would it take a computer to execute the operations needed to solve the problem when $K = 1000$ and $N = 50$? $K = 10000$ and $N = 500$? (Assume that the computer will do 10^8 multiplications, 10^9 additions and 10^9 comparisons respectively per second. Use $p(N, K)$ defined in (b) for these evaluations.)

(d) Nonlinear Costs. Suppose the production and storage costs are nonlinear in the amounts produced and stored. Can you modify your recursion in part (a) to solve this case without expanding the number of "states" and "actions"? If so, show how to do this; if not, explain why.

2. Discovering System Transience. The purpose of this problem is to give two methods of discovering whether or not a system, i.e., a finite discrete-time-parameter Markov population decision chain, is transient. Improvements in one of the methods are given for substochastic systems.

(a) Characterization of Transient Matrices. Suppose P is a square real nonnegative matrix. Show that $P^N \to 0$ if and only if the linear inequalities $(I - P)v = r$, $v \geq 0$ have a solution for every $r \geq 0$ (resp., some $r \gg 0$). [*Hint*: For the "if" part, write the first equation in the form $v = r + Pv$ and iterate.]

(b) Discovering System Transience by Policy Improvement. Show how the policy-improvement method can be used to check whether or not a system is transient. [*Hint*: Let $r_\delta = 1$ for all δ and apply the policy-improvement method. Use the result of (a) to show how to detect when a decision δ is found for which δ is not transient.]

(c) Discovering System Transience in Substochastic Systems by a Combinatorial Method. Consider a substochastic system with S states and A state-action pairs. Let $q(s, a)$ be the proba-

bility that an individual who is in state s and chooses action a exits the system in one step. Thus, $q(s,a) \equiv 1 - \sum_{t \in \mathcal{S}} p(t|s,a)$ for each $a \in A_s$ and $s \in \mathcal{S}$. Let $q_\pi^N \equiv (1 - P_\pi^N)1$ be the vector of probabilities that an individual exits the system in $N \geq 0$ steps or less starting from each state while using π. For $N \geq 0$, let T^N be the set of states $s \in \mathcal{S}$ from which it is possible to exit the system in N steps or less with every policy, i.e., $\min_\pi q_{\pi s}^N > 0$. Show that the following are equivalent: 1° the system is transient; 2° $T^S = \mathcal{S}$; 3° $\min_\pi q_\pi^S \gg 0$. Give a low order polynomial in S and A that is an upper bound on the number of comparisons needed to find T^S. [*Hint*: For $N \geq 1$, let U^N be the set of states $s \in \mathcal{S}$ from which it is possible to exit the system in N steps or less with every policy, but not less than N steps for some policy. Then $T^N = T^{N-1} \cup U^N$ for $N \geq 1$. Show that there is an integer $0 \leq M \leq S$ for which $T^M = T^{M+1}$. Show that $1^\circ \Rightarrow 2^\circ \Rightarrow 3^\circ \Rightarrow 1^\circ$. Show that $1^\circ \Rightarrow 2^\circ$ by contraposition.]

3. Successive Approximations and Newton's Method Find Nearly Optimal Policies in Linear Time for Discounted Markov Decision Chains. Consider a finite discrete-time-parameter S-state discounted Markov decision chain with discount factor $\beta = \frac{1}{1+\rho}$ and interest rate $100\rho\% > 0$, so $P_\delta 1 = \beta 1$ for all δ. Since adding a constant to the rewards earned in all states does not affect the set of maximum-value policies, there is no loss of generality in assuming that $r_\delta \geq 0$ for all δ. Let n be the larger of the number of state-action pairs and the number of nonzero data elements $r(s,a)$, $p(t|s,a)$. Let $\epsilon > 0$ be given and V^* be the maximum value over all policies. Call a policy π ϵ-*optimal* if $\|V^* - V_\pi\| \leq \epsilon \|V^*\|$.

(a) Computing $\mathcal{R}V$ in Linear Time. Show that the number of operations required to compute $\mathcal{R}V$ for any fixed $V \in \Re^S$ is at most n. (An operation is an addition, a comparison, and a multiplication).

(b) Successive Approximations Finds an ϵ-Optimal Policy in Linear Time. Show that $V^N \equiv \mathcal{R}^N 0$ is increasing in N. Also show, for fixed ϵ and β, how successive approximations can be used to find a stationary ϵ-optimal policy with at most $O(n)$ operations. [*Hint*: Let δ^∞ be a stationary maximum-value policy and γ be such that $V^N = r_\gamma + P_\gamma V^{N-1}$. Show that $V_\delta^N \leq V^N \leq V_\gamma \leq V_\delta^N + P_\delta^N V_\delta$. Also show how to choose N depending on ϵ and β, but not on n, such that γ^∞ is ϵ-optimal.]

(c) Newton's Method Finds an ϵ-Optimal Policy in Linear Time. Show that the values $\overline{V}^N \equiv V_{\delta_N}$ of the successive decisions $\delta_N, N \geq 1,$ selected by Newton's method are increasing in N, and that $\overline{V}^N \geq V^N$ for all N. Thus conclude that Newton's method requires no more iterations to find a stationary ϵ-optimal policy than does successive approximations. Show that if Gaussian elimination is used to compute the value of a decision δ_N in Newton's method, then the number of operations needed to find that value is $O(S^3)$. Thus conclude that the number of operations required to find a stationary ϵ-optimal policy with Newton's method (when computing values by Gaussian elimination) is $O(n + S^3)$ for fixed ϵ and β. Show that this running time can be reduced to $O(n)$ if the values of successive decisions selected by Newton's method are instead estimated by successive approximations.

Answers to Homework 3 Due April 25

1. Multifacility Linear-Cost Production Planning.

(a) Dynamic-Programming Recursion. The forward dynamic-programming recursion for finding the minimum unit costs is ($C_1^k = c_1^k$)

$$C_n^k = \min\,[c_n^k + \sum_{m<n} \sum_j e_{mn}^{jk}\, C_m^j, h_{n-1}^k + C_{n-1}^k]$$

for $1 < n \leq N$ and $1 \leq k \leq K$. The right-hand side of the above equation is the minimum of two terms, the first representing production of product k in period n and the second production of that product before period n and storing the product from period $n-1$ to period n.

(b) Running Time. We summarize below the number of additions, multiplications and comparisons required to compute the C_n^k.

Additions
$$\sum_k \sum_n \sum_{m<n} \sum_j 1 = \frac{1}{2}\, N^2K^2 + O(NK^2).$$

Multiplications
$$\sum_k \sum_n \sum_{m<n} \sum_j 1 = \frac{1}{2}\, N^2K^2 + O(NK^2).$$

Comparisons
$$\sum_k \sum_n 1 = NK + O(K).$$

(c) Running Time: Special Cases. The approximate running times for special cases are summarized in the table below.

K	1000	10,000
N	50	500
Approx. Run Time	14 sec	38 hrs

Approximate Running Times

(d) Nonlinear Costs. It is not possible to modify the recursion in (a) to handle nonlinear costs without dramatically enlarging the state and actions spaces. The reason is that the recursion in (a) is based on the fact that the marginal costs are independent of the volume of production. That would no longer be true with nonlinear cost functions.

2. Discovering System Transience.

(a) Characterization of Transient Matrices. The "only if" part follows from the fact that $P^N \to 0$ implies $I - P$ is nonsingular and $(I-P)^{-1} = \sum_{N=0}^{\infty} P^N \geq 0$ by Lemma 3. Thus, $v = (I-P)^{-1}r \geq 0$ solves $(I-P)v = r$, $v \geq 0$ for all $r \geq 0$. For the "if" part, suppose $r \gg 0$ and $v \geq 0$ solves $(I-P)v = r$, $v \geq 0$. Then

$$v = r + Pv = \cdots = \sum_{i=0}^{N-1} P^i r + P^N v \geq \sum_{i=0}^{N} P^i r$$

for $N = 0, 1, \ldots$. Thus since $r \gg 0$, $\sum_{i=0}^{\infty} P^i$ converges, whence $P^N \to 0$.

(b) Using Policy Improvement to Discover System Transience. Let $r_\delta = 1$ for all $\delta \in \Delta$. Suppose $\delta \in \Delta$ and $(I - P_\delta)v = 1$ has no nonnegative solution. Then from (a), δ^∞ is not transient. If the equations have a nonnegative solution, then that solution is $V_\delta \geq 1$. Now choose an improvement γ of δ, if one exists, replace δ by γ, and repeat the above construction. After finitely many steps, we will either find a δ^∞ that is not transient, in which case the system is not transient, or a δ that has no improvement. In the latter event, $V_\delta \geq 0$ is a fixed point of $\mathcal{R}$, whence $V_\delta = \max_{\gamma \in \Delta}(1 + P_\gamma V_\delta) \gg P_\gamma V_\delta$ for all γ. Thus, for each γ, $\overline{r}_\gamma \equiv (I - P_\gamma)V_\delta \gg 0$, so from (a), γ is transient. Hence the system is transient.

(c) Discovering System Transience in Substochastic Systems by a Combinatorial Method.

For $N \geq 0$, let T^N be the set of states $s \in \mathcal{S}$ from which it is possible to exit the system in N steps or less with every policy. For $N \geq 1$, let U^N be the set of states $s \in \mathcal{S}$ from which it is possible to exit the system in N steps or less with every policy, but not in less than N steps with some policy. Then $T^N = T^{N-1} \cup U^N$ for $N \geq 1$.

$1^\circ \Rightarrow 2^\circ$. The proof is by contraposition. If 2° is false, then $T^S \subset \mathcal{S}$. Let $\overline{T} = \mathcal{S} \setminus T^S$. By definition of the T^N, $\emptyset = T^0 \subseteq T^1 \subseteq T^2 \subseteq \cdots$, so there is a smallest integer $0 \leq M \leq S$ such that $U^{M+1} = \emptyset$, whence $T^M = T^{M+1}$. We claim that $U^N = \emptyset$ for all $N > M$. This is certainly so for $N = M + 1$. Suppose it is so for $N > M$ and consider $N + 1$. Then, for each state $s \in U^{N+1}$, each policy that exits the system in $N + 1$ steps, but not less, first takes an action that sends the system only to states in U^N, which is impossible because $U^N = \emptyset$. Hence $T^M = T^S$. Then there is a decision δ such that $P_{\delta \overline{T}\,\overline{T}}$ is stochastic, whence δ^∞ is not transient.

$2^\circ \Rightarrow 3^\circ$. Immediate from the definitions involved.

$3^\circ \Rightarrow 1^\circ$. By hypothesis, $\min_\pi q_\pi^S \gg 0$ and so $\max_\pi P_\pi^S 1 \ll 1$. Thus there exists $0 < \lambda < 1$ such that $\max_\pi P_\pi^S 1 \leq \lambda 1$. Iterating this inequality shows that $\max_\pi P_\pi^{iS} 1 \leq \lambda^i 1$ for $i = 1, 2, \ldots$. Since $\max_\pi P_\pi^N 1$ is decreasing in $N \geq 0$, $\sum_{N=0}^{\infty} P_\pi^N 1 \leq S \sum_{i=0}^{\infty} P_\pi^{iS} 1 \leq S(1 - \lambda^S)^{-1}$, whence π is transient.

We show that the total number of comparisons to find T^M is at most AS. For $N \geq 0$, let A_s^N be the set of actions $a \in A_s$ in state $s \in \mathcal{S}$ that assure it is not possible to exit the system in N steps or less starting from s and using a initially. For $s \in \mathcal{S}$, $A_s^0 = \emptyset$. For $N \geq 1$ and $s \in T^{N-1}$, $A_s^N = \emptyset$. For $N = 1$ (resp., $N > 1$) and $s \in \mathcal{S} \setminus T^{N-1}$, A_s^N is the set of $a \in A_s^{N-1}$ such that $q(s, a) = 0$ (resp., $p(t|s, a) = 0$ for $t \in U^{N-1}$); given $A_{(\cdot)}^{N-1}, U^{N-1}$, T^{N-1} one can find A_s^N for all $s \in \mathcal{S} \setminus T^{N-1}$ with A (resp., $A \cdot |U^{N-1}|$) comparisons. For $N \geq 1$, U^N is the set of states $s \in \mathcal{S} \setminus T^{N-1}$ for which $A_s^N = \emptyset$, and the U^N are disjoint. If $T^M = \mathcal{S}$ (resp., $\subset \mathcal{S}$), it suffices to find A_s^N, and so U^N and $T^N = T^{N-1} \cup U^N$, for $s \in \mathcal{S} \setminus T^{N-1}$ and $N \leq M$ (resp., $N \leq M+1$); the number of comparisons to do so is at most $A(1 + \sum_2^M |U^{N-1}|) \leq A \cdot |T^M| = AS$ (resp., $A(1 + \sum_2^{M+1} |U^{N-1}|) \leq A(|\mathcal{S} \setminus T^M| + |T^M|) = AS$).

3. Successive Approximations and Newton's Method Find an ϵ-Optimal Policy in Linear Time for Discounted Markov Decision Chains.

(a) Computing $\mathcal{R}V$ in Linear Time. Recall that

$$(\mathcal{R}V)_s = \max_{a \in A_s}\{r(s,a) + \sum_{t \in \mathcal{S}} p(t|s,a)V_t\}$$

for each $V \in \Re^S$ and $s \in \mathcal{S}$. Observe that each $r(s,a)$ and $p(t|s,a)$ appears exactly once in the formula above for each s and therefore the number of operations needed to compute $(\mathcal{R}V)_s$ does not exceed the number n_s number of nonzero elements among the $r(s,a)$ and $p(t|s,a)$ for all $a \in A_s$ and $t \in \mathcal{S}$, for each fixed $s \in \mathcal{S}$. Thus, the total number of operations required to compute $\mathcal{R}V$ is bounded above by $\sum_s n_s = n$. Let $V^0 \equiv \overline{V}^0 \equiv 0$.

(b) Successive Approximations Finds an ϵ-Optimal Policy in Linear Time. Notice that $V^1 = \mathcal{R}V^0 = \mathcal{R}0 \geq 0$. Thus, by the monotonicity of $\mathcal{R}$, and so of $\mathcal{R}^N$, it follows that $V^{N+1} = \mathcal{R}^N V^1 \geq \mathcal{R}^N 0 = V^N$. Also, $V^N = r_\gamma + P_\gamma V^{N-1} \leq r_\gamma + P_\gamma V^N$, so $(I - P_\gamma)V^N \leq r_\gamma$. Premultiplying the last inequality by $(I - P_\gamma)^{-1}$, which exists and is nonnegative because γ^∞ is transient, gives $V^N \leq (I - P_\gamma)^{-1} r_\gamma = V_\gamma$. Since the system is strictly substochastic, it is transient and there is a stationary maximum-value policy δ^∞. Hence,

$$V_\delta^N \leq V^N \leq V_\gamma \leq V_\delta = V_\delta^N + P_\delta^N V_\delta = V^*.$$

Therefore,

$$0 \leq V^* - V_\gamma \leq V^* - V^N \leq P_\delta^N V_\delta = P_\delta^N V^*$$

and so

$$\|V^* - V_\gamma\| \leq \|V^* - V^N\| \leq \|P_\delta^N\| \|V^*\| = \beta^N \|V^*\|.$$

Thus, taking N iterations with $N = \left\lceil \frac{\ln \epsilon}{\ln \beta} \right\rceil = \left\lceil \frac{\ln (1/\epsilon)}{\ln (1+\rho)} \right\rceil$ assures that γ^∞ is ϵ-optimal and that V_γ is within $100\epsilon\%$ of V^*. Since N is independent of n and each iteration requires at most n operations, successive approximations finds a stationary ϵ-optimal policy with $O(n)$ operations for fixed ϵ and β.

(c) Newton's Method Finds an ϵ-Optimal Policy in Linear Time. Since Newton's method is an instance of the Policy-Improvement Method and the values of successive policies increase with that method, it follows that $\overline{V}^{N+1} \geq \overline{V}^N$ for $N \geq 0$. Thus, if $\overline{V}^N \geq V^N$, then

$$\overline{V}^{N+1} \geq r_{\delta_{N+1}} + P_{\delta_{N+1}} \overline{V}^N = \max_\delta\{r_\delta + P_\delta \overline{V}^N\} \geq \max_\delta\{r_\delta + P_\delta V^N\} = V^{N+1}.$$

This shows that the values at iteration N with Newton's method are at least as large as those with successive approximations. Thus Newton's method requires no more iterations than successive approximations, so the number N of iterations in (b) will find a stationary ϵ-optimal policy. Each

iteration N of Newton's method requires finding the "best" improvement δ_N and then its value V_{δ_N}. The first task can be done in $O(n)$ time as shown in (a) and the second can be done in $O(S^3)$ time if Gaussian elimination is used to solve the required system of S linear equations. Since the number of iterations is independent of both n and S, this implementation of Newton's method requires $O(n + S^3)$ operations to find an ϵ-optimal policy. However, if instead of solving a system of linear equations to compute V_{δ_N}, we use successive approximations to estimate this value as in (b), then both steps of Newton's method can be implemented in $O(n)$ time. Then this implementation of Newton's method requires $O(n)$ operations to find an ϵ-optimal policy.

Homework 4 Due May 2

(Do Problems 1, 2 and either 3 or 4; if you took MS&E 251, do Problems 2, 3 and 4)

1. Component Replacement. A manager of a fleet of M new machines desires to develop a maintenance policy that maximizes the expected net profit from the machines as long as they are "working". Working machines are rented out each morning and are returned by 5 p.m. that day with the daily rental revenue per machine being e. Each machine consists of four components, two of type α and two of type β. Each component may be "working" or "failed". Working components fail independently of one another while in service, though customers receive no refunds. The probability that working components of types α and β fail during a day are .2 and .1 respectively. A machine is "working" if it has at least one working component of each type; otherwise the machine is "failed". When a machine is returned at the end of a day, the manager observes the numbers of working components of each type therein. If the machine has failed, the manager discards it. If the machine is working, the manager decides which failed components, if any, to replace during the evening in order to prolong the machine's working life. Components of types α and β cost respectively 1 and 2. The labor cost to replace one failed component of either type in a working machine during an evening is 5 and that to replace two failed components, one of each type, is 6. Initially, all components of all machines are working.

(a) Solution by Newton's Method. Find the component replacement policy that maximizes expected profit for $e = 4$ using the MATLAB file *newtonr*1.*m* (Newton's method).

(b) Formulation as Linear Program. Formulate the problem of finding a component replacement policy for the machines that maximizes the expected net profit as a linear program. [*Hint*: The problem has nine variables and four equality constraints.]

(c) Solution by Linear Programmming. Find the optimal policy and the maximum expected profit therefrom by solving the linear program in (b) by using the MATLAB file *tranlpsc.m*. (Type "help linprog" at the MATLAB prompt to see explanations for the linprog solver.) Do this for each of $e = .5$, 2, 4 and 6, and where M is the sum of the two integers representing the positions in the alphabet of the first letter in your first and last names. For example, Walt Disney would have to solve the problem for the case $M = 23 + 4 = 27$ since W is the twenty-third and D is the fourth letter of the alphabet. Please be sure to state the value of M that you use. Discuss qualitatively the effect that you observe in the above computations of increasing the daily revenue e per machine on the optimal replacement policy.

(d) Extending the Service Life. The manager has been asked by the Vice-President for Marketing to extend the average working life of a machine to at least 17 days. For the case $e = 4$, what is the maximum-expected profit and corresponding optimal policy among those that satisfy this constraint? Use the MATLAB file *tranlpsc.m* to solve the problem.

2. Optimal Stopping Policy. Consider a finite S-state Markov branching decision chain. Call a policy π *stopping* if $P_\pi^N \to 0$. Call $V_\pi \equiv \limsup_{N\to\infty} V_\pi^N$ the *value* of π and call the supremum V^* of V_π over all stopping policies π the *maximum stopping value*. (The limit superior and supremum are taken coordinate-wise.) If there are no stopping policies, let $V^* \equiv -\infty$. Let $w \gg 0$ be a row S-vector and consider the primal linear program of finding a vector (x_δ) that

$$\text{maximizes} \qquad \sum_{\delta\in\Delta} x_\delta r_\delta$$

subject to

$$\sum_{\delta\in\Delta} x_\delta (I - P_\delta) = w$$
$$x_\delta \geq 0 \text{ all } \delta \in \Delta$$

(a) Existence of Stopping Policies. Show that the following statements are equivalent: 1° *There is a stopping policy.* 2° *There is a stationary stopping policy.* 3° *The primal linear program has a feasible solution.* [*Hint*: Show that 1° implies 2° by establishing that there is a transient policy $\pi = (\gamma_1, \gamma_2, \dots)$ that is *periodic*, i.e., for some $M \geq 1$, $\gamma_N = \gamma_{N+M}$ for $N = 1, 2, \dots$. Then, for this part only, replace each r_δ by -1. Show that $\mathcal{R}^N 0 \downarrow V \geq V_\pi$, that V is a finite negative fixed point of $\mathcal{R}$, and each δ^∞ for which $\mathcal{R}V = -1 + P_\delta V$ is a stopping policy.]

(b) Existence of Optimal Stopping Policies. Show that the following statements are equivalent: 1° V^* *is finite.* 2° *There is a stationary maximum-value stopping policy and* V^* *is the common least fixed point and least excessive point of* $\mathcal{R}$. (*An excessive point of* $\mathcal{R}$ *is a vector* $V \in \Re^S$ *with* $V \geq \mathcal{R}V$.) 3° *The primal linear program has an optimal solution.* 4° *There is a stopping policy and an excessive point of* $\mathcal{R}$. [*Hint*: Show that $4° \Rightarrow 3° \Rightarrow 2° \Rightarrow 1° \Rightarrow 4°$. Show that 3° implies 2° by establishing that one optimal solution of the linear program is basic and that for the corresponding decision δ, V_δ is optimal for the dual program, and so is a fixed point of $\mathcal{R}$. Then show that $V_\delta = V^*$. Establish that 1° implies 4° by using part (a) to show that since V^* is finite, there is a stationary stopping policy δ^∞, whence the primal is feasible. Then show $\mathcal{R}^N V_\delta$ is nondecreasing in N and converges to a fixed point of $\mathcal{R}$, so the dual is feasible.]

(c) Maximum Value Need Not Equal Maximum Stopping Value. Suppose $\mathcal{S} = \{1\}$, $\Delta = \{\gamma, \delta\}$, $r_\gamma = 0$, $P_\gamma = 1$, $r_\delta = -1$ and $P_\delta = 0$. Show that $V^* = -1$, but that $V = 0$ is the maximum value over all policies and is the greatest nonpositive fixed point of $\mathcal{R}$.

3. Simple Stopping Problems. Let P be the stochastic transition matrix of an S-state Markov chain. In each period N, say, the chain is observed in state $s \in \mathcal{S}$, there are two actions available. One is to *stop* and receive a *terminal reward* r_s. The other is to *continue* by paying an *entrance fee* c_s that permits the state of the chain to be observed in period $N+1$. Put $r = (r_s)$ and $c = (c_s)$.

(a) Existence of Stationary Optimal Stopping Policy. Show that there is a stationary maximum-value stopping policy if and only if $V \geq r \vee (-c + PV)$ for some $V \in \Re^S$.

(b) Solution in S Iterations by Linear Programming. Show that if the simplex method is used to solve the corresponding primal linear program starting with the decision that stops in every state, then the simplex method will terminate in S iterations or less. [*Hint*: Show that since the value of the stopping policy encountered at each iteration of the simplex method exceeds that of its predecessor, once the simplex method introduces the continuation action in a state, that action will not be changed back to the stopping action in that state at a subsequent iteration.]

(c) Entrance-Fee Problem. An *entrance-fee problem* is one in which there are no terminal rewards. Show how to reduce the above problem to an equivalent entrance-fee problem. [*Hint*: Subtract r from the equation $V = r \vee (-c + PV)$, thereby forming $\widehat{V} = 0 \vee (-\widehat{c} + P\widehat{V})$ where $\widehat{V} \equiv V - r$ and $\widehat{c} \equiv c + r - Pr$.]

(d) Optimality of Myopic Policies for Entrance-Fee Problem. Assume $\mathcal{S} = \{1, \ldots, S\}$. Consider an entrance-fee problem in which (i) P is upper triangular with diagonal elements less than 1, except that the last one equals 1, and (ii) there is an s^*, $1 \leq s^* \leq S$, such that $c_s < 0$ for $s < s^*$ and $c_s \geq 0$ for $s \geq s^*$. Show that it is optimal to continue if $s < s^*$ and to stop if $s \geq s^*$. Call this policy *myopic* because it maximizes the one-period reward.

4. House Buying. You have just moved to a city and are considering buying a house. The purchase prices of houses that interest you are independent identically distributed random variables with $p_1, \ldots, p_S$ $(\sum_i p_i = 1)$ being the positive probabilities that the purchase prices are respectively $1, \ldots, S$. Each time you look at a house to determine its price, you incur a cost $c > 0$. Determine explicitly a stationary policy that minimizes the expected cost of looking and buying among those policies in which you eventually buy with probability one. [*Hint*: Assume that you always have the option of buying the cheapest house that you have observed todate. Then apply the result of problem 3(d) to find an optimal policy. Also show that one optimal policy has the property that if you do not buy a house in the period in which you observe its price, then you never buy it.]

Answers to Homework 4 Due May 2

1. Component Replacement. The states of a machine at the end of a day are the ordered pairs (i, j) where i and j are respectively the number of working components of types α and β, so $\mathcal{S} = \{(1,1), (1,2), (2,1), (2,2)\}$. The actions available in state (i, j) are the ordered pairs $(k, l) \geq (i, j)$ of numbers of working components of the two types after replacement. Let $p(m, n \mid k, l)$ be the conditional probability that a machine in state (k, l) after replacements in an evening is in state $(m, n) \leq (k, l)$ at 5:00 p.m. the following day. Denote by $r(i, j, k, l)$ the net profit earned during a day by a machine in state (i, j) at 5:00 p.m. on the previous day when action (k, l) is chosen for the machine that evening. The $p(t \mid a)$ are given by the formula

$$p(m, n|k, l) = \binom{k}{m} .8^m .2^{k-m} \binom{l}{n} .9^n .1^{l-n}$$

and are tabulated below. The $r(s, a)$ are given by $r(s, a) = e - c(s, a)$ where

$$c(i, j, k, l) = (k - i) + 2(l - j) + c(k + l - i - j)$$

and $c(u)$ is the labor cost of replacing u components in the machine, $c(0) = 0$, $c(1) = 5$, and $c(2) = 6$. The $c(s, a)$ are as tabulated below. Also note that the nine state-action pairs are as given in a table below. In this instance, there are four actions in state 11, two in each of states 12 and 21, and one in state 22.

$p(t \mid a)$

	a			
t	11	12	21	22
11	.7200	.1440	.2880	.0576
12		.6480	-	.2592
21			.5760	.1152
22				.5184

$c(s, a)$

	a			
s	11	12	21	22
11	0	7	6	9
12		0	-	6
21			0	7
22				0

State-Action Pairs

s	a
11	11
11	12
11	21
11	22
12	12
12	22
21	21
21	22
22	22

Let $p(t|s, a) = p(t|a)$. Let w_t be the number of machines initially in state t. Assume that $w_t = 0$ for $t \neq (2, 2)$.

(a) Solution by Newton's Method. Below is the MATLAB output resulting from running a file *compdata.m* to create the input data and then the file *newtonr1.m* in the course directory to solve the component replacement problem. Then the MATLAB output is:

```
>> compdata

P =
    0.7200         0         0         0
    0.1440    0.6480         0         0
    0.2880         0    0.5760         0
    0.0576    0.2592    0.1152    0.5184
    0.1440    0.6480         0         0
    0.0576    0.2592    0.1152    0.5184
    0.2880         0    0.5760         0
    0.0576    0.2592    0.1152    0.5184
    0.0576    0.2592    0.1152    0.5184

r =
     4
    -3
    -2
    -5
     4
    -2
     4
    -3
     4

>> newtonr1
stop =
yes
ans =
Halted after 2 iterations

V =
   17.6774
   20.6774
   21.4412
   26.6774

LastDecision =
   4  2  1  1
```

The maximum expected net revenues earned and optimal actions in states 11, 12, 21 and 22 are respectively the elements of V above and the actions 22, 22, 21 and 22 indicated in the vector *LastDecision* (note that action 4 in state 11 is 22, action 2 in state 12 is 22, action 1 in state 21 is 21, action 1 in state 22 is 22).

(b) Formulation as Linear Program. The problem is to choose the expected numbers $x_{sa} \geq 0$ of machine-days that machines end a day in state s and take replacement action a for each s, a, to maximize the expected net revenue

$$\sum_{s,a} r(s,a) x_{sa} \tag{1}$$

subject to

$$\sum_{a} x_{ta} \; - \; \sum_{s,a} p(t \mid s,a) x_{sa} = w_t, \; t \in \mathcal{S}. \tag{2}$$

The matrix of nonzero coefficients of (2) is tabulated below (with $w_{22} = 1$) together with $r(s,a)$ for each of the four cases $e = .5$, 2, 4 and 6.

Daily Reward Vectors for Several e's

s	11				12		21		22
e a	11	12	21	22	12	22	21	22	22
0.5	0.5	−6.5	−5.5	−8.5	0.5	−5.5	0.5	−6.5	0.5
2	2	−5	−4	−7	2	−4	2	−5	2
4	4	−3	−2	−5	4	−2	4	−3	4
6	6	−1	0	−3	6	0	6	−1	6

Coefficient Matrix of Equations (2)

s	11				12		21		22	
t a	11	12	21	22	12	22	21	22	22	w_t
11	.28	.856	.712	.9424	−.144	−.0576	−.288	−.0576	−.0576	
12		−.648		−.2592	.352	.7408		−.2592	−.2592	
21			−.576	−.1152		−.1152	.424	.8848	−.1152	
22				−.5184		−.5184		−.5184	.4816	1

(c) Solution by Linear Programming. The optimal action to take and the maximum unit value V_s^* of each machine in each state s are tabulated below for each of the required values of e. Notice that the optimal actions in each state are the same for all nonnegative initial populations, and the maximum values are linear therein. In particular, if M is the initial number of machines in state 22, then the maximum value therein is $V_{22}^* M$.

Observe that as the daily rental price e increases, it is optimal to replace more components to prolong the lives of the machines. Moreover, if it is optimal to replace any components, then all failed components should be replaced because of the scale economies in labor cost of replacement. Finally, when a machine has only one failed component and $e = 4$, then it is optimal to replace component α, but not β, since the former is both cheaper and more likely to fail than the latter.

Optimal Actions and Maximum Unit Values V_s^* in Each State s for several e's

	Optimal Action in s				Maximum Unit Value V_s^* in s			
e s	11	12	21	22	11	12	21	22
0.5	11	12	21	22	1.79	2.15	2.39	2.98
2	11	12	21	22	7.14	8.60	9.57	11.92
4	22	22	21	22	17.68	20.68	21.44	26.68
6	22	22	22	22	53.90	56.90	55.90	62.90

Below is the MATLAB output resulting from running the file *compdata.m* to create the input data and then the file *tranlpsc.m* in the course directory to solve the component replacement problem by linear programming. The MATLAB output for the case $e = 4$ is:

```
>> compdata

P =
    0.7200         0         0         0
    0.1440    0.6480         0         0
    0.2880         0    0.5760         0
    0.0576    0.2592    0.1152    0.5184
    0.1440    0.6480         0         0
    0.0576    0.2592    0.1152    0.5184
    0.2880         0    0.5760         0
    0.0576    0.2592    0.1152    0.5184
    0.0576    0.2592    0.1152    0.5184

r =
     4
    -3
    -2
    -5
     4
    -2
     4
    -3
     4

>> w = [0 0 0 1]

w = 0     0     0     1

>> tranlpsc

x = 0  -0.0000  0.0000  1.5696  0  2.9948  3.1392  0  6.9895

V =
   17.6774
   20.6774
   21.4412
   26.6774

d = 4     2     1     1
```

The maximum expected net revenues earned and optimal actions in states 11, 12, 21 and 22 are respectively the elements of V above and the actions 22, 22, 21 and 22 indicated in the vector d (note that action 4 in state 11 is 22, action 2 in state 12 is 22, action 1 in state 21 is 21, action 1 in state 22 is 22). These results agree with those using Newton's method.

(d) Extending the Service Life. The service life constraint

$$\sum_{s,a} x_{sa} \geq 17 \tag{3}$$

must be appended to the constraints (2).

Below is the MATLAB output resulting from running the file *compdata.m* with $e = 4$ and then *tranlpsc.m*, to solve the component replacement problem with service constraint by linear programming. The MATLAB output is:

```
>> compdata

P =
    0.7200         0         0         0
    0.1440    0.6480         0         0
    0.2880         0    0.5760         0
    0.0576    0.2592    0.1152    0.5184
    0.1440    0.6480         0         0
    0.0576    0.2592    0.1152    0.5184
    0.2880         0    0.5760         0
    0.0576    0.2592    0.1152    0.5184
    0.0576    0.2592    0.1152    0.5184

r =
     4
    -3
    -2
    -5
     4
    -2
     4
    -3
     4

>> w = [ 0 0 0 1]

w = 0     0     0     1

>> tranlpsc

x =  0   0.0000   0   1.3973   0   3.9360   1.8148   0.9799   8.8720
pr = 0   0.0000   0   1.0000   0   1.0000   0.6494   0.3506   1.0000

v = 24.9490
```

Here the elements of *pr* are the optimal conditional probabilities of taking each action starting in each state. In particular, it is optimal to take actions 22 in states 11, 12 and 22. In state 21, it is optimal to choose action 21 with probability .6494 and action 22 with probability .3506. One consequence of adding the service constraint is to reduce the maximum expected reward from $V_{22} = 26.6774$ to $v = 24.9490$.

2. Optimal Stopping Policy.

(a) Existence of Stopping Policies. $1^\circ \Rightarrow 2^\circ$. Let π be a stopping policy. Then there is an M such that $\|P_\pi^M\| < \frac{1}{2}$. Let $\pi = (\gamma_1, \gamma_2, \dots)$, $\pi^M = (\gamma_1, \dots, \gamma_M)$ and $\mu = (\pi^M, \pi^M, \dots)$, so μ is periodic. Notice that μ is also transient because

$$\|\sum_{N=0}^{\infty} P_\mu^N\| = \|\sum_{k=0}^{\infty}\sum_{n=0}^{M-1} P_\mu^{Mk+n}\| \le \sum_{k=0}^{\infty} \|(P_\pi^M)^k\| \sum_{n=0}^{M-1} \|P_\pi^n\| \le 2 \sum_{n=0}^{M-1} \|P_\pi^n\| < \infty.$$

Let $r_\delta = -1$ for all δ, $V^0 = 0$ and $V^{N+1} = \mathcal{R}V^N$ for $N \ge 0$. We claim V^N is nonincreasing in N. This is certainly so for $N = 1$ since $V^1 = -1 \le 0 = V^0$. Thus, for $N > 1$, it follows from the monotonicity of $\mathcal{R}$, and so of $\mathcal{R}^N$, that $V^{N+1} \le V^N$. Since $r_\delta = -1 \ll 0$ for all δ, then $V^N = \mathcal{R}^N 0 \ge V_\mu^N \ge V_\mu \gg -\infty$, so the sequence $\{V^N\}$ is bounded below and has a limit V. Also, by

the continuity of $\mathcal{R}$, $\mathcal{R}V = \mathcal{R}(\lim_{N\to\infty}\mathcal{R}^N 0) = \lim_{N\to\infty}\mathcal{R}^{N+1}0 = V$. Thus, there is a decision δ such that $V = -1 + P_\delta V \le -1 \ll 0$, so from Problem 2 of Homework 3, δ^∞ is transient and so stopping.

$2^\circ \Rightarrow 3^\circ$. Let δ be a stationary stopping policy. Then $P_\delta^N \to 0$ and, by Lemma 3, $(I - P_\delta)^{-1}$ exists and is nonnegative. Thus $x_\gamma = 0$ for $\gamma \neq \delta$ and $x_\delta = w(I - P_\delta)^{-1} \ge 0$ is a basic feasible solution of the primal.

$3^\circ \Rightarrow 1^\circ$. If there is a feasible solution to the primal, there is a basic feasible solution. Hence by Theorem 15, there is a transient policy δ^∞, and δ^∞ is stopping.

(b) Existence of Optimal Stopping Policies. $4^\circ \Rightarrow 3^\circ$. Since there is a stopping policy, by (a), the primal is feasible. If there exists V such that $V \ge \mathcal{R}V$, then the dual is also feasible. This implies that both the primal and dual have optimal solutions.

$3^\circ \Rightarrow 2^\circ$. If there is an optimal solution of the primal, then by Theorem 15 there is a basic optimal solution with row basis $I - P_\delta$ for some decision δ for which δ^∞ is transient, and so stopping. Thus the basic optimal solution is such that $x_\delta(I - P_\delta) = w$ and $x_\gamma = 0$ for $\gamma \neq \delta$ and there exist optimal dual prices V. Since V is feasible for the dual, $V \ge \mathcal{R}V$. And by complementary slackness, $V = r_\delta + P_\delta V$, so $V = V_\delta$ is finite and a fixed point of $\mathcal{R}$. Then $V_\delta = \mathcal{R}V_\delta = \cdots = \mathcal{R}^N V_\delta \ge V_\pi^N + P_\pi^N V_\delta$. If π is stopping, $P_\pi^N \to 0$, so on taking limit superiors on both sides, $V_\delta \ge V_\pi$ and $V^* = V_\delta$. That V_δ is the least excessive and fixed point of $\mathcal{R}$ follows from the fact that if V is excessive, then $V \ge r_\delta + P_\delta V$, i.e., $(I - P_\delta)V \ge r_\delta$. Premultiplying the last inequality by $(I - P_\delta)^{-1}$, which exists and is nonnegative, implies that $V \ge (I - P_\delta)^{-1} r_\delta = V_\delta$ as claimed.

$2^\circ \Rightarrow 1^\circ$. Since there exists a stationary maximum-value stopping policy δ^∞, $V^* = V_\delta$ and, since δ is transient, V^* is finite.

$1^\circ \Rightarrow 4^\circ$. Since V^* is finite, there is a stopping policy and so, by (a), a stationary stopping policy δ^∞, say. Then $\mathcal{R}V_\delta \ge r_\delta + P_\delta V_\delta = V_\delta$. Hence, $\mathcal{R}^{N+1}V_\delta \ge \mathcal{R}^N V_\delta$ by the monotonicity of $\mathcal{R}$, and so $\mathcal{R}^N$. Let $\pi = (\gamma_i)$ be a policy such that $\mathcal{R}^N V_\delta = V_\pi^N(V_\delta) = V_{\pi^N\delta^\infty}$ where $\pi^N \equiv (\gamma_1, \ldots, \gamma_N)$. Since δ^∞ is stopping, so is $\pi^N\delta^\infty$; and it's value is $\mathcal{R}^N V_\delta$, whence $\mathcal{R}^N V_\delta \le V^*$. Hence since $\mathcal{R}^N V_\delta$ is bounded above and nondecreasing in $N \ge 0$, $V \equiv \lim_{N\to\infty}\mathcal{R}^N V_\delta$ exists and, by the continuity of $\mathcal{R}$, $V = \mathcal{R}V$. Thus V is excessive.

(c) Maximum Value Need Not Equal Maximum Stopping Value. Observe that since the rewards are nonpositive, V_μ is nonpositive for every policy μ. Also, notice that $V_\gamma = 0$ is the maximum value. Let π be a stopping policy. Then π uses δ in some period, so $V_\pi = -1$ is the maximum stopping value. Now $\mathcal{R}V = V \vee -1$ and 0 (resp., -1) is the greatest nonpositive (resp., least) fixed point of $\mathcal{R}$.

3. Simple Stopping Problems. The irredundant form of the linear program in Problem 2 is that of choosing S-element row vectors x and y that maximize

$$xr - yc$$

subject to

$$x + y(I - P) = w, \;\; x, y \geq 0$$

where w is a positive S-element row vector.

(a) Existence of Stationary Optimal Stopping Policy. For the "if" part, observe that V is excessive since $V \geq r \vee (-c + PV) = \mathcal{R}V$. Let δ^∞ be the stopping policy that stops in every state, so $P_\delta = 0$. Then there is a stationary maximum-value stopping policy by Problem 2(b). For the "only if" part, observe that if there is a stationary maximum-value stopping policy, then V^* is finite. Hence, by Problem 2(b), there is an excessive point of $\mathcal{R}$.

(b) Solution in S Iterations by Simplex Method. Recall that the simplex method is the special case of the policy improvement method in which each iteration improves the action in exactly one state. Now suppose the simplex method begins with the decision δ that stops in every state. Each subsequent iteration will either produce a stationary stopping policy (c.f., Theorem 15) whose value is greater than that of its predecessor (and strictly so in the state whose action is changed), or will establish that the primal objective function is unbounded above, in which case no stationary maximum-value stopping policy exists. Consider the former case. Observe that it is not possible to improve the action in a state s from "stop" to "continue" and then back to "stop" again at subsequent iterations. This is because the former change increases the value in s from r_s to a strictly higher value while the latter returns the value in s to r_s, which contradicts the fact that the value in a state rises with each iteration. Thus the number of states in which one continues at an iteration properly contains that at the previous iteration. Hence, by the result of (a) and of Problem 2(b), the simplex method terminates in at most S iterations with a stationary maximum-value stopping policy or by finding that the primal objective function is unbounded above, in which case no such policy exists.

(c) Entrance-Fee Problem. Observe that $V = r \vee (-c + PV)$ if and only if $\widehat{V} = 0 \vee (-\widehat{c} + P\widehat{V})$, and the latter is an entrance-fee problem. Thus there is a solution of the former system if and only if that is so of the latter, so there is a stationary maximum-value stopping policy in one problem if and only if that is so in the other. In either case, V^* is the least solution of the former system if and only if $\widehat{V}^* = V^* - r$ is the least solution of the latter. Also, $V_s^* = r_s$ if and only if $\widehat{V}_s^* = 0$, so the optimal states in which to stop are the same in both problems.

(d) Optimality of Myopic Policies for Entrance-Fee Problems. It is certainly optimal to continue in states $s < s^*$ with negative entrance fees $c_s < 0$ because one earns an immediate re-

ward $-c_s > 0$ from so doing compared with a zero reward from stopping therein. On the other hand, if $s \geq s^*$ it is optimal to stop and earn a zero reward. This is because the entrance fees are nonnegative in all states in $[s^*, S]$, and once one enters this set of states it is never possible to return to any state $t < s^*$ with a negative entrance fee because the elements of P below the diagonal are zero. Thus continuation in any state in $[s^*, S]$ cannot earn positive expected rewards. A more formal proof may be given by showing by induction on s, starting with $s = S$, that the least fixed point $V^* = (V_s^*)$ of the optimal return operator is nonegative and is given by

$$V_s^* = \begin{cases} 0 & , s \geq s^* \\ -c_s + \sum_{t<s^*} p_{st} V_t^* , & s < s^*. \end{cases}$$

4. House Buying. This is a special case of the simple stopping problem in Problem 3 in which the states are the lowest prices $1, \ldots, S$ seen so far. If s is the lowest price seen so far, then one can either stop and buy the house whose price was s with revenue $r_s = -s$ or continue looking at a cost $c_s = c$. Then $(Pr)_s = -\sum_1^S (t \wedge s) p_t$. Observe first that $V = (V_s) = 0$ satisfies $V_s \geq -s \vee (-c + PV)$, so by the result of Problem 3(a), a stationary optimal stopping policy exists. Also, by the result of Problem 3(c), the problem is equivalent to the entrance-fee problem in which $\widehat{c} = (\widehat{c}_s) = c + r - Pr$ so

$$\widehat{c}_s = c - s + \sum_{t=1}^{S} (t \wedge s) p_t = c - \sum_{t=1}^{S} (t - s)^- p_t.$$

Also, the conditions for the optimality of myopic policies in Problem 3(d) are fulfilled with the order of the states reversed. For the elements of $P = (p_{st})$ above the diagonal are zero and the diagonal elements are less than one except for the first which equals one since $p_{tt} \leq 1 - p_1 < 1$ for $t > 1$ and $p_{11} = 1$. Also, $\widehat{c}_1 = c > 0$, so there is a maximum $s = s^*$ such that $\widehat{c}_s > 0$. Moreover, $\widehat{c}_s$ is nonincreasing in s so $\widehat{c}_s > 0$ for $s \leq s^*$ and $\widehat{c}_s \leq 0$ for $s > s^*$. Thus, if s is the lowest house price seen to date and $s \leq s^*$, it is optimal to buy the house whose price was s; otherwise it is optimal to continue looking.

Observe that in implementing this policy, one buys the first time that the lowest price s to date is s^* or less. But this happens only if the lowest price to date is in fact that for the last house observed. Thus, the policy of buying the first house whose price is s^* or less is optimal for the problem in which the option of buying a house whose price was observed in a previous period is not available.

Homework 5 Due May 9

(Do Problem 3 and Problem 1 or 2—Problem 2 if you took MS&E 251)

1. Bayesian Statistical Quality Control and Repair. A firm manufactures a product under a continuing contract with the government that allows the government to cancel at any time without penalty. During any given day, the process for producing the product is either *in control* or *out of control*. Whether the process is out of control is not observed directly, but can only be inferred from the results of production. Each day the firm produces one item, inspects it and classifies it as good or defective. The government accepts a good item and pays the firm $r > 0$ for it. The firm discards a defective item. When the process is out of control, each item produced is defective. Thus if a good item is produced, the process was in control during its production. When the process is in control, the (known) probability that it produces a good item is p, $0 < p < 1$. The (known) probability that the process is in control at the time of production of an item given that it is in control at the time of production of its predecessor is q, $0 < q < 1$. Once the process is out of control, it remains that way until it is repaired. Independently of the production process, there is a probability β, $0 < \beta < 1$, that the firm will retain the contract for another day. The government informs the firm at the beginning of a day whether or not the contract is to be continued that day. If the decision is to cancel, the firm receives no further revenue from the government and incurs no further costs. If the decision is to continue, there are two possibilities that day: immediately repair at a cost $K > 0$ or don't repair. Repair is done quickly and brings the process into control with probability q (regardless of whether or not it was in control at the time of repair). Repair is the only permissible option when S consecutive defectives have been produced since the later of the times of the last repair and last good item. This Bayesian sequential statistical decision problem is to choose a repair policy that maximizes the expected value of profits before the government cancels the contract.

(a) Posterior Probability the Next Item is Good. Let p_s be the conditional probability that the item $s+1$ is good given that items $1, \ldots, s$ were defective and either item 0 was good or the process was repaired just before producing item 1. Give a formula for p_s in terms of p, q and s.

(b) Maximum Expected Value. Write down explicitly an equation (in terms of β, K, r and the p_s) whose solution is the maximum expected profit before the government cancels the contract. Define the states and actions. Is this system transient? Why? Does there exist a stationary policy achieving the maximum expected profit? Explain briefly. Discuss whether or not the policy-improvement method could be used to solve this problem.

(c) Optimality of Control-Limit Policy. Show for the problem in part (b) that a *control-limit* policy is optimal, i.e., there is a number s^* such that if the number of consecutive defective items since the later of the last good item or repair is s, it is optimal to repair if $s \geq s^*$ and not to re-

pair otherwise. [*Hint*: First show that p_s is decreasing in s. Next show by induction on N that the maximum expected profit over N days is decreasing in the state where the terminal value function is zero. Then use successive approximations to show that the maximum expected profit over the infinite horizon is decreasing in the state. Finally, establish the optimality of a control-limit policy.]

Remark. The approach to establishing the form of the optimal policy in the infinite-horizon problem given in part (c), viz., establishing a desired property of the N-period value function by induction and then retaining the property as $N \to \infty$, is the most commonly used way to establish the form of the optimal policy in infinite-horizon problems.

2. Minimum Expected Present Value of Sojourn Times. Consider a finite transient substochastic Markov decision chain with state space $\mathcal{S}$, one-period rewards r_δ and transition matrices P_δ for each decision $\delta \in \Delta$. Append a stopped state τ to $\mathcal{S}$. The vector of probabilities that the system enters τ from each state in $\mathcal{S}$ in one step using decision δ is $(I - P_\delta)1$. Let $T_{\pi s}$ be the sojourn time of the system in $\mathcal{S}$ before reaching τ starting from state s and using policy π. Let $T^\rho_{\pi s} = \sum_{j=1}^{T_{\pi s}} \beta^j$ be the present value of the sojourn time of the system in $\mathcal{S}$ starting from state $s \in \mathcal{S}$ and using policy π where $0 < \beta = \frac{1}{1+\rho} < 1$ is the discount factor. Let $T_\pi = (T_{\pi s})$, $\mathrm{E}T_\pi = (\mathrm{E}T_{\pi s})$, $\mathrm{E}T^2_\pi = (\mathrm{E}T^2_{\pi s})$, $\mathrm{Var}\, T_\pi = (\mathrm{Var}\, T_{\pi s})$ and $\mathrm{E}T^\rho_\pi = (\mathrm{E}T^\rho_{\pi s})$ be column vectors.

(a) Factorial Moments of Sojourn Times. Let $T = T_\delta$, $P = P_\delta$ and $D = D_\delta$. Show that $D = (I - P)^{-1}$ and $\mathrm{E}\binom{T+k-1}{k} = D^k 1$, $k = 1, 2, \ldots$ where $\mathrm{E}\binom{T+j}{k} = \left(\mathrm{E}\binom{T_s+j}{k}\right)$ is a column vector. [*Hint*: Let $I^N_s = 1$ if the system starts in state s, uses δ^∞, and does not reach state τ before or in period N, and let $I^N_s = 0$ otherwise. Then $T_{\delta s} = \sum_{N=0}^\infty I^N_s$. Now compute the desired expectations by induction on k using the familiar binomial identity $\binom{T+j+1}{k} = \binom{T+j}{k-1} + \binom{T+j}{k}$.]

(b) Minimum Limiting Expected Present Value of Sojourn Times of Individuals and Populations, and Moment-Optimality. Suppose one exogenous individual arrives in each state s in each period. Let $\underline{T}^\rho_{\delta s}$ be the present value of sojourn times in $\mathcal{S}$ of all exogenous individuals who first enter the system in state s and their progeny, and let $\underline{T}^\rho_\delta = (\underline{T}^\rho_{\delta s})$. Show that the following problems are equivalent:

1° Minimum Limiting Expected Present Value of Population Sojourn Time. Find δ for which $\limsup_{\rho \downarrow 0} (\mathrm{E}\underline{T}^\rho_\delta - \mathrm{E}\underline{T}^\rho_\gamma) \le 0$ for all $\gamma \in \Delta$.

2° Minimum Limiting Expected Discounted Individual Sojourn Time of Order One. Find δ for which $\limsup_{\rho \downarrow 0} \rho^{-1}(\mathrm{E}T^\rho_\delta - \mathrm{E}T^\rho_\gamma) \le 0$ for all $\gamma \in \Delta$.

3° Maximum Expected Population Reward with Negative Unit One-Period Reward. Find δ for which $\liminf_{\rho \downarrow 0} \rho^{-1}(V^\rho_\delta - V^\rho_\gamma) \ge 0$ for all $\gamma \in \Delta$ where $r_\gamma = -1$ for all $\gamma \in \Delta$.

4° Policies with No 1-Improvement and Negative Unit One-Period Reward. Find δ that maximizes (v^0_δ, v^1_δ) lexicographically where $r_\gamma = -1$ for all $\gamma \in \Delta$.

5° Moment Optimality. Find δ that minimizes $(\mathrm{E}T_\delta, -\mathrm{Var}\, T_\delta)$ lexicographically.

How would you have to modify the above results if the rate of exogenous arrivals into each state were constant, but state dependent? Random (and independent) with constant rate, but state dependent? [*Hint*: Establish the following equivalences: $1^\circ \Leftrightarrow 2^\circ \Leftrightarrow 3^\circ \Leftrightarrow 4^\circ \Leftrightarrow 5^\circ$. To establish the first equivalence, show that ρ^{-1} is the present value of the number of exogenous individuals entering the system in each given state. Use the result of (a) to establish the last equivalence.]

(c) Risk Posture. For fixed $\rho > 0$, does an individual who wishes to minimize $\mathrm{E}T_\delta^\rho$ exhibit risk aversion, risk preference or neither toward his sojourn time T_δ in $\mathcal{S}$? Why? Explain briefly why your conclusion is consistent with the equivalence of 2° and 5° in (b).

(d) Limiting Maximum Expected Sojourn Times. The problems in parts (b) and (c) seem most appropiate in situations where the sojourn times are painful, so we want to get them over quickly. But often sojourn times are pleasant, e.g., a firm's time to bankruptcy, a student's stay at Stanford, or even life itself, so we want to prolong them. In this circumstance, how would you modify the results of parts (b) and (c).

3. Optimal Control of Tandem Queues. Consider a job waiting to be processed by a factory at the beginning of some hour. The job must first be processed at station one and then at station two. The job may be processed at each station by a skilled or unskilled worker. Unskilled workers are paid the hourly rate h_i at station i while working on the job and skilled workers are paid twice that much. If the job is processed by a worker at station one during an hour and does not require rework, the job is sent on to station two for service at the beginning of the next hour and earns the shop the progress payment r_1. If the job is processed by a worker at station two during an hour and does not require rework, the job is *shipped* and earns the shop an additional payment r_2. If the job is processed at either station in an hour by an unskilled worker, the job requires rework at that station during the subsequent hour with probability p, $\frac{1}{2} < p < 1$, whether or not the job was previously reworked. By contrast if the job is processed by a skilled worker at either station in an hour, the job requires rework at that station during the following hour with probability $2p - 1$. Thus, the probability a skilled worker who processes the job at a station during an hour successfully completes the job there is $2(1 - p)$, which is double the corresponding probability for an unskilled worker. The problem is to decide whether to use a skilled or unskilled worker at each station to process the job there. In short, the issue is whether it is better to use a slow worker with low pay or a fast worker with high pay at each station.

(a) Strong-Maximum-Present-Value Policies by Strong Policy Improvement Method. Assume $p = .7$, $h_i = 3 + 3i$ and $r_i = 17i$ for $i = 1, 2$. Determine which stationary policies have maximum value. Use the strong policy improvement method as implemented on page 59 of *Lectures in Dynamic Programming and Stochastic Control* to find a strong-maximum-present-value policy starting with the policy in which an unskilled worker is used at each station.

(b) Nonstationary Arrivals. Suppose the expected number of jobs arriving for service at the factory in hour $N+1$ is $w^N = 3000 + N + 8\sin(\frac{N\pi}{12})$, $N = 0, 1, \ldots$. Also assume that the numbers of unskilled and skilled workers available at each station each hour is large enough so that all jobs requiring service during the hour can be processed without delay. What stationary policies have strong maximum present value? [*Hint*: Show that the present value of profits is the product of the present values of the arrival stream (w^N) and the profits earned by a single job.]

Answers to Homework 5 Due May 9

1. Bayesian Statistical Quality Control and Repair.

(a) Posterior Probability the Next Item is Good. Let G be the event that item $s{+}1$ is good; L be the event that items $1,\dots,s$ are defective and either item 0 is good or the process is repaired before producing item 1; and I be the number of the items $1,\dots,s$ that are produced while the process is in control. Then $L = \cup_{k=0}^{s}\{L, I=k\}$. Let $p_s = \Pr(G \,|\, L)$. Then

$$p_s = \Pr(G \,|\, L) = \frac{\Pr(G,L)}{\Pr(L)} = \frac{\Pr(G,L,I=s)}{\sum\limits_{k=0}^{s}\Pr(L,I=k)} = \frac{qp[q(1-p)]^s}{(1-q)\sum\limits_{k=0}^{s-1}[q(1-p)]^k + [q(1-p)]^s}.$$

(b) Maximum Expected Value. The states are the numbers $0,1,\dots,S$ of consecutive defective items produced since the later of the time of the last repair or good item. In each state other than S, there are two actions, *repair* or *don't repair*. In state S the only action is to repair. Let V_s be the maximum expected net income given that the system is in state s. Observe that $p_0 = qp$. Then for $0 \le s < S$,

$$(*) \qquad V_s = \beta\max\,(rp_0 - K + p_0V_0 + (1-p_0)V_1, rp_s + p_sV_0 + (1-p_s)V_{s+1}),$$

and $V_S = \beta(rp_0 - K + p_0V_0 + (1-p_0)V_1)$. Since the probability of retaining the contract on a day is β, $0 \le \beta < 1$, the system is strictly substochastic and so transient. Thus there is a stationary maximum-value policy. Hence policy-improvement method could be used to solve this problem.

(c) Optimality of Control-Limit Policy. To see that a control limit policy is optimal, i.e., there exists s^* such that one repairs if $s \ge s^*$ and doesn't repair otherwise, first show that p_s is decreasing in s as follows. Put $\alpha = q(1-p)$. Then

$$p_s = \frac{qp\alpha^s}{(1-q)\sum\limits_{k=0}^{s-1}\alpha^k + \alpha^s} = \frac{qp}{(1-q)\sum\limits_{k=0}^{s-1}\alpha^{k-s} + 1} = \frac{qp}{(1-q)\sum\limits_{k=1}^{s}\alpha^{-k} + 1}$$

is decreasing in s because $\sum_{k=1}^{s}\alpha^{-k}$ is increasing in s and $q < 1$.

Next show that the maximum value V_s is decreasing in $s = 0,\dots,S$. To this end, put $V^0 \equiv 0$, so V_s^0 is decreasing in $s = 0,\dots,S$. Now suppose that V_s^N is decreasing in $s = 0,\dots,S$ and consider $N+1$. Then for $s = 0,\dots,S{-}1$,

$$(**) \qquad V_s^{N+1} = \beta\max\,(rp_0 - K + p_0V_0^N + (1-p_0)V_1^N, rp_s + p_sV_0^N + (1-p_s)V_{s+1}^N).$$

Since $r > 0$ and p_s is decreasing in s, rp_s is decreasing in $s = 0,\dots,S$. Also, since $p_s > 0$, it follows from the induction hypothesis that

$$p_sV_0^N + (1-p_s)V_{s+1}^N \ge p_{s+1}V_0^N + (1-p_{s+1})V_{s+1}^N \ge p_{s+1}V_0^N + (1-p_{s+1})V_{s+2}^N,$$

i.e., $p_s V_0^N + (1-p_s)V_{s+1}^N$ is decreasing in $s = 0, \ldots, S-1$.[1] Thus, the second term in parenthesis on the right side of $(**)$ is decreasing in $s = 0, \ldots, S-1$. Since the first term in parenthesis on the right side of $(**)$ is a constant, the pointwise maximum of a constant and a decreasing function is decreasing, V_s^{N+1} is decreasing in $s = 0, \ldots, S-1$. Also, $V_{S-1}^{N+1} \geq \beta(rp_0 - K + p_0 V_0^N + (1-p_0)V_1^N)$ $= V_S^{N+1}$, so V_s^{N+1} is decreasing in $s = 0, \ldots, S$.

Now since the system is strictly substochastic, it follows from the fact (shown in *Lectures ...*) that the maximum N-period value converges to the maximum value that $\lim_{N\to\infty} V_s^N = V_s$ is also decreasing in $s = 0, \ldots, S$. Consequently, the second term in parentheses on the right-hand side of $(*)$ is decreasing in $s = 0, \ldots, S-1$ and the control limit policy defined by letting s^* be the smallest integer s, $0 \leq s < S$, such that $rp_0 - K + p_0 V_0 + (1-p_0)V_1 \geq rp_s + p_s V_0 + (1-p_s)V_{s+1}$ if there is such an integer, and letting $s^* \equiv S$ otherwise, has maximum value since that decision assures that equality occurs in $(*)$.

2. Minimum Expected Present Value of Sojourn Times.

(a) Factorial Moments of Sojourn Times. Since the system is transient, $P^* = 0$ and $D = (I-P)^{-1}$. We now show by induction that $\mathrm{E}\binom{T+k-1}{k} = D^k 1$. This is trivially so for $k = 0$. Suppose it is true for $k \geq 0$ and consider $k+1$. Then by the induction hypothesis,

$$\mathrm{E}\binom{T+k}{k+1} = \mathrm{E}\binom{T+k-1}{k} + \mathrm{E}\binom{T+k-1}{k+1} = D^k 1 + \mathrm{E}\binom{T+k-1}{k+1}.$$

Let R_1 be the event that the system reaches τ in the first step and $p_1 = \Pr(R_1)$. Then

$$\mathrm{E}\binom{T+k-1}{k+1} = \mathrm{E}\left(\binom{T+k-1}{k+1} \mid R_1\right) p_1 + \mathrm{E}\left(\binom{T+k-1}{k+1} \mid R_1^c\right)(1-p_1),$$

Observe that, in the above equation, the first term is 0 because $\binom{k}{k+1} = 0$, and the second term is $P\mathrm{E}\binom{T+k}{k+1}$. Thus, $\mathrm{E}\binom{T+k}{k+1} = D^k 1 + P\mathrm{E}\binom{T+k}{k+1}$, so $\mathrm{E}\binom{T+k}{k+1} = (I-P)^{-1} D^k 1 = D^{k+1} 1$.[2]

(b) Minimum Limiting Expected Present Value of Sojourn Times of Individuals and Populations, and Moment-Optimality.

$1^\circ \Leftrightarrow 2^\circ$. The present value of the number of exogenous individuals entering the system in a given state is $\sum_{i=1}^\infty \beta^i = \rho^{-1}$ for $\rho > 0$ since exactly one exogenous individual enters the state in each period. Then, $\mathrm{E}\underline{T}_\delta^\rho = \rho^{-1}\mathrm{E}T_\delta^\rho$, which implies the claim.

$2^\circ \Leftrightarrow 3^\circ$. As shown in *Lectures ...*, for each γ and all $\rho > 0$, $\mathrm{E}T_\gamma^\rho = R_\gamma^\rho 1 = -R_\gamma^\rho(-1) = -V_\gamma^\rho$ since $r_\gamma = -1$, which establishes the claim.

$3^\circ \Leftrightarrow 4^\circ$. For small values of ρ and each γ, $V_\gamma^\rho = \sum_{n=-1}^\infty \rho^n v_\gamma^n = \sum_{n=0}^\infty \rho^n v_\gamma^n$ since $P_\gamma^* = 0$. Thus $\rho^{-1}(V_\delta^\rho - V_\gamma^\rho) = \rho^{-1}(v_\delta^0 - v_\gamma^0) + (v_\delta^1 - v_\gamma^1) + o(1)$. On taking the limit inferior as $\rho \downarrow 0$ on both

[1] This monotonicity argument is due to Gareth James.

[2] This induction step is due to Alamuru Krishna.

sides, one sees that the limit inferior on the left is nonnegative if and only if $(v_\delta^0, v_\delta^1) \succeq (v_\gamma^0, v_\gamma^1)$ where $\succeq$ is the usual lexicographic order.

$4^\circ \Leftrightarrow 5^\circ$. It is enough to observe that $\mathrm{E}T_\delta = D_\delta 1 = -D_\delta r_\delta = -v_\delta^0$ and

$$\operatorname{Var} T_\delta = 2\mathrm{E}\binom{T_\delta+1}{2} - \mathrm{E}\binom{T_\delta}{1} - \mathrm{E}\binom{T_\delta}{1}^2 = 2D_\delta^2 1 - D_\delta 1 - (D_\delta 1)^2 = 2v_\delta^1 - v_\delta^0 - (v_\delta^0)^2.$$

No modifications are required in either case.

(c) Risk Posture. Since the present value of the individual's sojourn time is T_δ^ρ, minimizing $\mathrm{E}T_\delta^\rho$ is equivalent to maximizing $-\mathrm{E}T_\delta^\rho$, and $-T_\delta^\rho$ is a convex function of T_δ, it follows that the individual exhibits risk preference toward his sojourn time T_δ. By the equivalence of 2° and 5°, minimizing the sum of the first two terms of the expansion of $\rho^{-1}\mathrm{E}T_\delta^\rho$ for small $\rho > 0$ is equivalent to lexicographically minimizing $(\mathrm{E}T_\delta, -\operatorname{Var} T_\delta)$. Thus the individual prefers policies that minimize $\mathrm{E}T_\delta$, and among policies that achieve this minimum, those that maximize $\operatorname{Var} T_\delta$. This is consistent with risk preference.

(d) Limiting Maximum Expected Sojourn Times. In (b), modify 1° and 2° by replacing lim sup by lim inf, and $\leq$ by $\geq$; modify 3° and 4° by replacing -1 by $+1$; and modify 5° by replacing minimizes by maximizes. The modified version of (c) involves an individual who wishes to maximize $\mathrm{E}T_\delta^\rho$, and so exhibits risk aversion towards his sojourn time T_δ. By the equivalence of the modified versions of 2° and 5°, maximizing the sum of the first two terms of the expansion of $\rho^{-1}\mathrm{E}T_\delta^\rho$ for small $\rho > 0$ is equivalent to lexicographically maximizing $(\mathrm{E}T_\delta, -\operatorname{Var} T_\delta)$. Thus the individual prefers policies that maximize $\mathrm{E}T_\delta$, and among policies that achieve this maximum, those that minimize $\operatorname{Var} T_\delta$. This is consistent with risk aversion.

3. Optimal Control of Tandem Queues.

(a) Strong-Maximum-Present-Value Policies by Strong Policy Improvement Method. The state space consists of two states 1 and 2, viz., the station at which a job is located. There are four possible decisions: $\Delta = \{uu, us, su, ss\}$, where us represents using unskilled workers at station 1 and and skilled workers at station 2, etc. The transition matrices associated with each decision are:

$$P_{uu} = \begin{pmatrix} p & 1-p \\ 0 & p \end{pmatrix} \qquad P_{us} = \begin{pmatrix} p & 1-p \\ 0 & 2p-1 \end{pmatrix}$$

$$P_{su} = \begin{pmatrix} 2p-1 & 2(1-p) \\ 0 & p \end{pmatrix} \qquad P_{ss} = \begin{pmatrix} 2p-1 & 2(1-p) \\ 0 & 2p-1 \end{pmatrix}.$$

Notice that each is transient and hence the system is transient, i.e., $d = 0$. Thus $P_\delta^* = 0$ and $D_\delta = (I - P_\delta)^{-1}$ for all $\delta \in \Delta$ and, if T is the minimum of the ranks of the stationary matrices over all

decisions, then $T = 0$. Theorem 22 on p. 54 implies that a stationary strong-maximum-present-value policy exists and Theorem 25 on p. 57 implies that $\Delta \supseteq \Delta_0 \supseteq \Delta_1 \supseteq \Delta_2 = \Delta_\infty$.

Next show that $\Delta = \Delta_0 \supset \Delta_1 = \Delta_2 = \Delta_\infty$ and $\Delta_1 = \{ss\}$, i.e., that using skilled workers at both stations is the unique stationary strong-maximum-present-value policy. Saying that $\Delta_0 = \Delta$ is equivalent to saying that all stationary policies have the same value. That this is so should be clear from the results of the single station problem. Since the downstream station (station 2) may be considered as a single station in isolation, whatever action one takes in state 2 must give the same value, i.e., $v^0_{u2} = v^0_{s2} \equiv v^0_2 = r_2 - \frac{h_2}{1-p}$ $\left(= 34 - \frac{9}{0.3} = 4 \right)$. Then for station 1, as far as total expected value is concerned, the expected value earned in state 2 is an additional reward for shipping (so $r'_1 = r_1 + v^0_2 = 17 + 4 = 21$), whence the action taken in state 1 does not affect the expected value obtained in state 2, i.e., $v^0_{u1} = v^0_{s1} \equiv v^0_1 = r'_1 - \frac{h_1}{1-p}$ $(= 21 - \frac{6}{0.3} = 1)$. Thus the value of all stationary policies is $V^* = \binom{1}{4}$.

It is possible to show this directly by calculating $V_\delta = D_\delta r_\delta = (I - P_\delta)^{-1} r_\delta$ for each $\delta \in \Delta$, where the reward vectors associated with each decision are

$$r_{uu} = \begin{pmatrix} (1-p)r_1 - h_1 \\ (1-p)r_2 - h_2 \end{pmatrix} \qquad r_{us} = \begin{pmatrix} (1-p)r_1 - h_1 \\ 2(1-p)r_2 - 2h_2 \end{pmatrix}$$

$$r_{su} = \begin{pmatrix} 2(1-p)r_1 - 2h_1 \\ (1-p)r_2 - h_2 \end{pmatrix} \qquad r_{ss} = \begin{pmatrix} 2(1-p)r_1 - 2h_1 \\ 2(1-p)r_2 - 2h_2 \end{pmatrix}.$$

Now $v^0_\delta = \lim_{\rho \downarrow 0} V^\rho_\delta = V_\delta$ for each $\delta \in \Delta$, and so $v^0 = \begin{pmatrix} 1 \\ 4 \end{pmatrix}$ and $\Delta_0 = \Delta$. Now determine $v^1 = \max_{\gamma \in \Delta_0}(r^1_\gamma - v^0 + P_\gamma v^1) = \max_{\gamma \in \Delta}(-v^0 + P_\gamma v^1)$. As explained on p. 59 of *Lectures*..., the problem of finding v^1 is precisely that of finding a stationary maximum-value policy for the transient system with a restricted decision set Δ_0 and a new reward vector $r^1_\gamma - v^0 = -v^0$. To do this, use policy improvement and, as suggested, begin with the policy uu. Since $v^1_{uu} = -v^0 + P_{uu} v^1_{uu}$, it follows that

$$v^1_{uu1} = -1 + .7 v^1_{uu1} + .3 v^1_{uu2}$$

and

$$v^1_{uu2} = -4 \qquad\qquad + .7 v^1_{uu2}$$

from which one concludes

$$v^1_{uu} = \begin{pmatrix} -\frac{50}{3} \\ -\frac{40}{3} \end{pmatrix}.$$

Now consider improving uu with the aid of the comparison function

$$G_\gamma v^1_{uu} = -v^0 + P_\gamma v^1_{uu} - v^1_{uu}.$$

Set $\gamma = ss$. Then

$$G_{ss}v^1_{uu} = -\begin{pmatrix}1\\4\end{pmatrix} + \begin{pmatrix}.4 & .6\\0 & .4\end{pmatrix}\begin{pmatrix}-\frac{50}{3}\\-\frac{40}{3}\end{pmatrix} + \begin{pmatrix}-\frac{50}{3}\\-\frac{40}{3}\end{pmatrix} = \begin{pmatrix}1\\4\end{pmatrix},$$

and so ss improves uu. But ss is, in fact, the decision that gives the "greatest" improvement of uu, i.e., ss is the decision that Newton's method selects. This is clear since $(G_\gamma v^1_{uu})_t$ depends on γ^t, but not on γ^s for $s \neq t$, so that one has

$$G_{uu}v^1_{uu} = \begin{pmatrix}0\\0\end{pmatrix},\ G_{us}v^1_{uu} = \begin{pmatrix}0\\4\end{pmatrix} \text{ and } G_{su}v^1_{uu} = \begin{pmatrix}1\\0\end{pmatrix}.$$

Now since $v^1_{ss} = -v^0 + P_{ss}v^1_{ss}$, it follows that

$$v^1_{ss1} = -1 + .4v^1_{ss1} + .6v^1_{ss2}$$

and

$$v^1_{ss2} = -4 \qquad\qquad + .4v^1_{ss2}$$

from which follows

$$v^1_{ss} = \begin{pmatrix}-\frac{25}{3}\\-\frac{20}{3}\end{pmatrix}.$$

Now we claim that $G_\gamma v^1_{ss} \leq 0$ for all $\gamma \in \Delta_0$, which implies that $ss \in \Delta_1$. Further, since $G_\gamma v^1_{ss} \neq 0$ unless $\gamma = ss$, it follows that $\Delta_1 = \{ss\}$. Since Δ_∞ is non-empty and is a subset of Δ_1, conclude that $\Delta_\infty = \{ss\}$ and hence that ss is the unique stationary strong-maximum-present-value policy.

(b) Nonstationary Analysis. Under the stationary policy δ, the N^{th} cohort (i.e., the jobs that arrive in hour N+1) earn a value discounted to time N of $V^\rho_{\delta 1}$. Discounting this to time 0, the N^{th} cohort earns the present value $\beta^N V^\rho_{\delta 1}$. Thus the present value of the policy δ^∞ is

$$\sum_{N=0}^{\infty} w^N \beta^N V^\rho_{\delta 1} = V^\rho_{\delta 1} \sum_{N=0}^{\infty} \beta^N w^N.$$

But $\sum_{N=0}^{\infty} \beta^N w^N$ is the present value of the arrival stream. This demonstrates the hint. It follows immediately that a policy that has strong-maximum-present-value for a single job, also has strong-maximum-present-value for the problem with non-stationary arrivals. Thus ss is the unique stationary strong-maximum-present-value policy for this problem.

Homework 6 Due May 16

1. Limiting Present-Value Optimality with Binomial Immigration. Consider a bounded S-state discrete-time-parameter Markov population decision chain with immigration stream $w = (w^N)$, interest rate $100\rho\% > 0$ and discount factor is $\beta = \dfrac{1}{1+\rho}$. Assume throughout that the series $w^\rho \equiv \sum_{N=0}^{\infty} \beta^N w^N$ is absolutely-convergent for all $\rho > 0$. (For example, that is so if $w^N = O(N^q)$ for some $q \geq 0$.) The s^{th} diagonal element of w^ρ is the present value of the stream of immigrants into state s discounted to period one. Let $V_\pi^{\rho w}$ be the present value of the (cohort) policy π with the immigration stream w. Call λ *limiting present-value optimal* for w if $\liminf_{\rho \downarrow 0} (V_\lambda^{\rho w} - V_\pi^{\rho w}) \geq 0$ for all π. Let $\Delta_n \equiv \{\delta \in \Delta : V_\delta^n \succeq V_\gamma^n, \text{ all } \gamma \in \Delta\}$ and $\Delta_\infty \equiv \{\delta \in \Delta : V_\delta \succeq V_\gamma, \text{ all } \gamma \in \Delta\}$.

(a) Present Values of Convolutions are Products of Present Values. Suppose that w and $\widehat{w}$ are immigration streams for which the series defining w^ρ and $\widehat{w}^\rho$ are absolutely convergent. Show that $(w * \widehat{w})^\rho = w^\rho \widehat{w}^\rho$. Show by induction on n that $1_n^\rho = (1-\beta)^{-n} = (1+\rho)^n \rho^{-n}$ for $|n| = 0, 1, \ldots$ where 1_n^ρ is the present value of the binomial sequence of order n discounted to period one.

(b) Limiting Present-Value Optimality for Binomial Immigration Stream of Order n. Show that $V_\pi^{\rho w} = w^\rho V_\pi^\rho$. Show also that λ is limiting present-value optimal for the binomial immigration stream of order n (≥ -1) if and only if $\liminf_{\rho \downarrow 0} \rho^{-n}(V_\lambda^\rho - V_\pi^\rho) \geq 0$ for all π.

(c) Δ_n is the Set of Decisions that are Limiting Present-Value Optimal for the Binomial Immigration Stream of Order n. Show that δ^∞ is limiting present-value optimal for the binomial immigration stream of order n (≥ -1) if and only if $\delta \in \Delta_n$.

(d) Machine Sequencing. Let $\alpha_N \geq 0$ be the expected number of jobs arriving at a firm's machine shop in week $N+1$, $N = 0, 1, \ldots$. The machine shop has three types of machines labeled $1, 2, 3$. Each job must be processed through one of the four machine sequences $a, b, c,$ or d given in the table below. Each job takes three weeks to process through the shop with each sequence, one week on each machine in the sequence. There are enough machines of each type so that jobs are not delayed at any machine to which they are assigned by the machine shop. The firm incurs a cost of i thousand dollars in each week a job is processed by machine i. Which sequences are (Cesàro) overtaking optimal for the firm when the expected numbers of jobs arriving each week are respectively $\alpha_0 > 0$ and $\alpha_N = 0$ for $N > 0$, $\alpha_N = \alpha_0 > 0$ for all $N \geq 0$, and

Machines in each Sequence

	Week		
Sequence	1st	2nd	3rd
a	1	3	2
b	2	1	3
c	2	3	1
d	1	3	3

$\alpha_N = \alpha_0(N+1) > 0$ for all $N \geq 0$? Which sequences are limiting present-value optimal for each of the above three job-arrival streams? Which sequences have strong maximum present value for each of the three job-arrival streams. [*Hint*: Tabulate $\mathbb{V}_\delta^\alpha - \mathbb{V}_\gamma^\alpha$ for each sequence $\alpha = (\alpha_N)$ and relevant pairs γ, δ of decisions directly from the definitions without first finding the terms of the expansions.]

(e) Social Security: Aggregating Losses to Produce a Profit. A country that does not permit immigration divides its adult population into two groups, workers (W) and retirees (R). The country has proposed a new social security system that will cover all current and future inhabitants (except retirees in the initial population who will be supported in a different way). Individuals in the country spend one period of 30 years in each group. Let $w^0 = \text{diag}(w_W^0 \;\; 0)$ be the initial population in the two groups to be covered by social security. The vector of exogenous individuals (children) entering the population in period $N+1$ is $p^N w^0$, $p > 0$. Workers in one period have an even chance of surviving to retire in the next period and retirees in a period have no chance of surviving to retire again in the next period. Each worker earns an average income $I > 0$ during his working years and pays 10% thereof into social security. Each retiree receives $20p$% of I from social security during retirement.

(*i*) Show that the expected social security payments to each individual (except the initial retirees) is greater than, equal to, or less than the amount collected from the individual respectively according as $p > 1$, $p = 1$, or $p < 1$, and yet social security is always in the black!

(*ii*) Show that the ratio of a worker's retirement income from social security to his income per person before retirement after social security payments is $\frac{2}{9}p(p+1)$, and so is an increasing function of p (assuming that a worker's income is divided equally between himself and his p children). In particular that ratio is .17, .44 or .83 respectively according as $p = .5$, 1.0 or 1.5.

(*iii*) In this example with $p > 1$, whence the number of immigrants increases geometrically, the government pays out more to each individual than it collects (i.e., $V_\delta \ll 0$) and yet collects more money from the population as a whole than it pays out (i.e., $\liminf_{N\to\infty} V_\delta^{Nw} \gg 0$ (C, 1)). Show that this phenomenon cannot occur in a transient Markov population decision chain (as in this example) with a binomial immigration stream w.

2. Maximizing Reward Rate by Linear Programming. Assume that an S-state discrete-time-parameter Markov population decision chain has degree one. Let $w \geq 0$ be a row S-vector of the numbers of individuals initially in each state. Consider the following redundant dual pair of linear programs. The primal is to find row S-vectors x_γ^{-1} and x_γ^0 for $\gamma \in \Delta$ that

$$\text{maximize} \quad \sum_\gamma x_\gamma^{-1} r_\gamma$$

subject to

$$(1) \qquad \sum_\gamma x_\gamma^{-1} \quad - \sum_\gamma x_\gamma^0 Q_\gamma = w$$

$$(2) \qquad -\sum_\gamma x_\gamma^{-1} Q_\gamma \qquad\qquad = 0$$

$$(3) \qquad\qquad x_\gamma^{-1}, x_\gamma^0 \geq 0 \text{ for } \gamma \in \Delta.$$

The dual is to find column S-vectors v^{-1} and v^0 that

$$\text{minimize } w v^{-1}$$

subject to

$$(4) \qquad v^{-1} - Q_\gamma v^0 \geq r_\gamma$$

and

$$(5) \qquad -Q_\gamma v^{-1} \qquad\quad \geq 0$$

for all $\gamma \in \Delta$.

(a) Solution of Linear Programs. Show that there is a $\delta \in \Delta$ not depending on $w \geq 0$ and optimal solutions $(\overline{x}_\gamma^{-1} \;\; \overline{x}_\gamma^0)$ and $(\overline{v}^{-1} \;\; \overline{v}^0)$ respectively of the primal and dual such that $\overline{x}_\gamma^{-1} = \overline{x}_\gamma^0 = 0$ for all $\gamma \neq \delta$, $(\overline{v}^{-1} \;\; \overline{v}^0)$ is independent of $w \geq 0$, and $\overline{v}^{-1} = \max_{\gamma \in \Delta} v_\gamma^{-1} = v_\delta^{-1}$ is the least feasible value of v^{-1} for the dual. [*Hint*: Choose δ so that $G_{\gamma\delta}^0 \preceq 0$ for all $\gamma \in \Delta$, or equivalently $g_{\gamma\delta}^{-1} \leq 0$ and $N g_{\gamma\delta}^{-1} + g_{\gamma\delta}^0 \leq 0$ for all γ and all large enough N. Next show that $(\overline{v}^{-1} \;\; \overline{v}^0) \equiv (v_\delta^{-1} \;\; N v_\delta^{-1} + v_\delta^0)$ is feasible for the dual for all large enough N. Then show that for all large enough N, $(\overline{x}_\gamma^{-1} \;\; \overline{x}_\gamma^0)$ is feasible for the primal where $\overline{x}_\gamma^{-1} = \overline{x}_\gamma^0 = 0$ for all $\gamma \neq \delta$ and, for $\gamma = \delta$, $(\overline{x}_\delta^{-1} \;\; \overline{x}_\delta^0) \equiv (w P_\delta^* \;\; w(N P_\delta^* + D_\delta))$. Do this by using the fact that $0 \leq R_\delta^\rho = \rho^{-1} P_\delta^* + D_\delta + o(1)$ for all small enough $\rho > 0$. Next show that for all large enough N, $(\overline{x}_\gamma^{-1} \;\; \overline{x}_\gamma^0)$ and $(\overline{v}^{-1} \;\; \overline{v}^0)$ are respectively optimal for the primal and dual because their objective-function values are equal.]

(b) Average Maximum Equals Maximum Reward Rate. Show that $\lim_{N \to \infty} N^{-1} \mathcal{R}^N 0 = \max_{\gamma \in \Delta} v_\gamma^{-1}$. [*Hint*: Choose δ so $G_{\gamma\delta}^0 \preceq 0$ for all $\gamma \in \Delta$, or equivalently $\max_{\gamma \in \Delta} [N g_{\gamma\delta}^{-1} + g_{\gamma\delta}^0] = 0$ for all large enough N. Use this fact to show that for fixed large enough M, $\mathcal{R}^N(M v_\delta^{-1} + v_\delta^0) = (M+N) v_\delta^{-1} + v_\delta^0$, $N = 1, 2, \ldots$. The geometric interpretation of this result is that when $v_\delta^{-1} \neq 0$, $\mathcal{R}$ carries the half-line $\{N v_\delta^{-1} + v_\delta^0 : N \geq M\}$ into itself. And when $v_\delta^{-1} = 0$, v_δ^0 is a fixed point of $\mathcal{R}$. Indeed, Theorem 6 asserts that if the system is transient, then this fixed point is unique and is the maximum value.]

(c) Stochastic Systems with Strongly Connected System Graph. Suppose the system is stochastic and the system graph is strongly connected. The last means that for each pair of states $s \neq t$, there is a chain in the system graph from s to t. Show that this implies that there is a decision γ and a positive integer N (both depending on s and t) such that $P^N_{\gamma st} > 0$. Show that every stationary maximum-reward-rate policy δ^∞ has the property that $v^{-1}_{\delta s}$ is independent of s. [*Hint:* Set $v^{-1} = v^{-1}_\delta$, rewrite (5) as $v^{-1} \geq P_\gamma v^{-1}$ and iterate].

(d) Linear Programs for Stochastic Systems with Strongly Connected System Graph. Suppose the system is stochastic, the system graph is strongly connected and $\sum_s w_s = 1$. Use the result of (c) to reduce the pair of dual linear programs in the problem statement to the equivalent alternate pair of linear programs given below. The alternate primal is to find row S-vectors x^{-1}_γ for $\gamma \in \Delta$ that

$$\text{maximize} \sum_\gamma x^{-1}_\gamma r_\gamma$$

subject to

$$\sum_\gamma x^{-1}_\gamma 1 \quad = 1 \tag{6}$$

$$-\sum_\gamma x^{-1}_\gamma Q_\gamma = 0 \tag{7}$$

$$x^{-1}_\gamma \geq 0 \text{ for } \gamma \in \Delta. \tag{8}$$

The alternate dual is to find a number u^{-1} and a column S-vectors v^0 that

$$\text{minimize} \;\; u^{-1}$$

subject to

$$1u^{-1} - Q_\gamma v^0 \geq r_\gamma \tag{9}$$

for all $\gamma \in \Delta$ where 1 is here a column S-vector of ones.

Determine whether or not these alternate linear programs are dual to one another. Also show that there is a $\delta \in \Delta$ and optimal solutions $\overline{x}^{-1}_\gamma$ and $(\overline{u}^{-1} \;\; \overline{v}^0)$ respectively of the alternate primal and dual such that $\overline{x}^{-1}_\gamma = 0$ for all $\gamma \neq \delta$ and $1\overline{u}^{-1} = \max_{\gamma \in \Delta} v^{-1}_\gamma = v^{-1}_\delta$, and $\overline{u}^{-1}$ is the least feasible value of u^{-1} for the alternate dual. [*Hint:* Postmultiply (1) by the column S-vector 1 to form the alternate primal.]

Answers to Homework 6 Due May 16

1. Limiting Present-Value Optimality with Binomial Immigration.

(a) Present Values of Convolutions are Products of Present Values. We have

$$(w * \widehat{w}\,)^\rho = \sum_{N=0}^{\infty} \beta^N (w * \widehat{w}\,)_N = \sum_{N=0}^{\infty} \beta^N \sum_{i=0}^{N} w^i \widehat{w}^{N-i} = \sum_{i=0}^{\infty} \beta^i w^i \sum_{N=i}^{\infty} \beta^{N-i} \widehat{w}^{N-i} = w^\rho\, \widehat{w}^\rho.$$

We now show by induction that $1_n^\rho = (1-\beta)^{-n} = \rho^{-n}\,(1+\rho)^n$. The claim holds for $n = 0, 1, -1$ since $1_0^\rho = 1 = (1-\beta)^0$, $1_1^\rho = \sum_{i=0}^{\infty} \beta^k = (1-\beta)^{-1}$ and $1_{-1}^\rho = 1-\beta$. Suppose the claim is true for $n > 0$. Then $1_{n+1}^\rho = (\,1_1 * 1_n)^\rho = 1_1^\rho\, 1_n^\rho = (1-\beta)^{-1}(1-\beta)^{-n} = (1-\beta)^{-n-1}$. Suppose the claim is true for $n < 0$. Then $1_{n-1}^\rho = (\,1_{-1} * 1_n)^\rho = 1_{-1}^\rho\, 1_n^\rho = (1-\beta)\,(1-\beta)^{-n} = (1-\beta)^{-n+1}$.

(b) Limiting Present-Value Optimality for Binomial Immigration Streams of Order n. Evidently, $V_\pi^{\rho\omega} = \sum_{k=0}^{\infty} \beta^k w^k V_\pi^\rho = w^\rho V_\pi^\rho$. Then, λ is limiting present-value optimal for the binomial immigration stream 1_n^ρ of order n if $\liminf_{\rho\downarrow 0} 1_n^\rho\,(V_\lambda^\rho - V_\pi^\rho) \geq 0$ for all π, or equivalently, since $1_n^\rho = \rho^{-n}\,(1+\rho)^n$ and $\lim_{\rho\downarrow 0}(1+\rho)^n = 1 > 0$, if $\liminf_{\rho\downarrow 0} \rho^{-n}\,(V_\lambda^\rho - V_\pi^\rho) \geq 0$ for all π.

(c) Δ_n is the Set of Decisions that are Limiting Present-Value Optimal for the Binomial Immigration Stream of Order n. Since the system is bounded, there is a stationary policy that has maximum present value for all small enough $\rho > 0$. Thus, from (b), δ^∞ is limiting present-value optimal for the binomial immigration stream of order n if and only if the following holds: $\liminf_{\rho\downarrow 0} \rho^{-n}\,(V_\delta^\rho - V_\gamma^\rho) \geq 0$ for all $\gamma \in \Delta$. On substituting the relevant Laurent expansions, the last is so if and only if $\liminf_{\rho\downarrow 0} \rho^{-n} \sum_{i=-1}^{\infty} \rho^i\,(v_\delta^i - v_\gamma^i) = \liminf_{\rho\downarrow 0} \sum_{i=-1}^{n} \rho^{i-n}(v_\delta^i - v_\gamma^i) \geq 0$ for all $\gamma \in \Delta$, or equivalently $V_\delta^n - V_\gamma^n \succeq 0$ for all γ, i.e., $\delta \in \Delta_n$.

(d) Machine Sequencing. The following table contains all the data required to establish optimality of policies under different preference relations. By a suitable choice of units we can and do assume $\alpha_0 = 1$, so the three immigration streams are essentially equivalent to binomial streams of orders 0, 1 and 2 respectively. By examining the column corresponding to $N \geq 3$, observe that for binomial immigration of order 0, policies a, b and c are Cesàro overtaking optimal, and so by part (c), limiting present-value optimal. For the binomial immigration stream of order 1, note that policies a and b are Cesàro overtaking optimal since it suffices to tabulate the difference of two policies for policies a, b and c because they are the only policies in Δ_0. Therefore, a and b are in Δ_1 and, by (c), limiting present-value optimal. Thus, for the third immigration stream, which is binomial of order 2, it suffices to compute the difference of values of policies a and b. The table shows that policy a is Cesàro overtaking optimal and, by (c), limiting present-value optimal.

	$N=1$	$N=2$	$N\geq 3$
$V_c^N - V_d^N$	-1	-1	1
$V_b^N - V_c^N$	0	2	0
$V_a^N - V_b^N$	1	-1	0
$V_b^{N1} - V_c^{N1}$	0	2	2
$V_a^{N1} - V_b^{N1}$	1	0	0
$V_a^{N2} - V_b^{N2}$	1	1	1

The only policy that has maximum present value for all sufficiently small positive interest rates is a since it is the unique element of Δ_2 and so $\Delta_2 = \Delta_\infty \neq \emptyset$ by Theorem 25 of §1.8 and the fact that the system is bounded.

(e) Social Security: Aggregating Losses to Produce a Profit.

(i) The expected social security payments to an individual are $\left(\frac{1}{2}\right).2\,pI = .1pI$. However, each individual pays $.1I$ to social security during his working years, which is less than, equal to or greater than the expected social security payments to him according as $p > 1$, $p = 1$ or $p < 1$ respectively. Yet social security breaks even in every period except the first one when it collects money but does not make any payments as can be seen from the fact that in period N social security collects $0.1\,p^N w_W^0 I$ and pays $\frac{1}{2}.2\,pIp^{N-1}w_W^0$.

(ii) Observe that the income per person before retirement after social security payments is given by $\frac{.9I}{p+1}$ as the net income is split among an individual and his children. Then the ratio of a worker's retirement income from social security to his income before retirement after social security payments is $\frac{2}{9}\,p(p+1)$.

(iii) We claim that $\lim_{N\to\infty} V_\delta^{Nn} \leq 0$ and all $n = -1, 0, \ldots$. This is obvious for $n = -1$ since $V_\delta^{N,-1} \to v_\delta^{-1} = 0$. Now suppose that $n \geq 0$. Choose $\epsilon > 0$ so that $V_\delta \leq -2\epsilon 1$. It is enough to show by induction on n that $V_\delta^{Nn} \leq -\epsilon 1$ for all large enough N and all $n \geq 0$. To that end, observe that since δ is transient, $V_\delta^N \to V_\delta$ so $V_\delta^N \leq -\epsilon 1$ for all large enough N. Now suppose the claim is so for n and consider $n+1$. Then $V_\delta^{N,n+1} = \sum_{i=0}^N V_\delta^{in} \leq -\epsilon 1$ for all large enough N as claimed.

2. Maximizing Reward Rate by Linear Programming. Consider the primal program of finding row S-vectors x_γ^{-1} and x_γ^0 for $\gamma \in \Delta$ that

$$\text{maximize} \sum_\gamma x_\gamma^{-1} r_\gamma$$

subject to

$$\sum_\gamma x_\gamma^{-1} - \sum_\gamma x_\gamma^0 Q_\gamma = w \tag{1}$$

$$-\sum_\gamma x_\gamma^{-1} Q_\gamma = 0 \tag{2}$$

$$x_\gamma^{-1}, x_\gamma^0 \geq 0 \text{ for } \gamma \in \Delta. \tag{3}$$

Then the dual program is to find column S-vectors v^{-1} and v^0 that

$$\text{minimize } wv^{-1}$$

subject to

$$v^{-1} - Q_\gamma v^0 \geq r_\gamma \tag{4}$$

and

$$-Q_\gamma v^{-1} \geq 0 \tag{5}$$

for all $\gamma \in \Delta$.

(a) Solution of Linear Programs. The first step is to observe from Theorem 22 of §1.8 that there is a δ such that $G^0_{\gamma\delta} \preceq 0$ for all $\gamma \in \Delta$, or equivalently,

$$g^{-1}_{\gamma\delta} \leq 0 \text{ and } Ng^{-1}_{\gamma\delta} + g^0_{\gamma\delta} \leq 0 \text{ for all } \gamma \in \Delta \text{ and all large enough } N. \tag{6}$$

Now from Theorem 26 of §1.8, $v^{-1}_\delta = \max_{\gamma\in\Delta} v^{-1}_\gamma$ independently of $w \geq 0$. Also (6) holds if and only if $(\overline{v}^{-1} \;\; \overline{v}^0) \equiv (v^{-1}_\delta \;\; Nv^{-1}_\delta + v^0_\delta)$ is feasible for the dual for all large enough N.

The next step is to observe from Theorem 21 of §1.8 that $0 \leq R^\rho_\delta = \rho^{-1}P^*_\delta + D_\delta + o(1)$ for all small enough $\rho > 0$. Thus, $\overline{x}^0_\delta \equiv Nx^{-1}_\delta + x^0_\delta \geq 0$ for all large enough N and all $w \geq 0$ where $x^{-1}_\delta \equiv wP^*_\delta$ and $x^0_\delta \equiv wD_\delta$. Let $\overline{x}^{-1}_\delta \equiv x^{-1}_\delta$ and $\overline{x}^{-1}_\gamma = \overline{x}^0_\gamma = 0$ for all $\gamma \neq \delta$. We claim that $(\overline{x}^{-1}_\gamma \;\; \overline{x}^0_\gamma)$, $\gamma \in \Delta$, is feasible for the primal for all large enough N. The claim follows from $\overline{x}^{-1}_\delta Q_\delta = wP^*_\delta Q_\delta = 0$ and

$$\overline{x}^{-1}_\delta - \overline{x}^0_\delta Q_\delta = x^{-1}_\delta - (Nx^{-1}_\delta + x^0_\delta)Q_\delta = x^{-1}_\delta - x^0_\delta Q_\delta = w(P^*_\delta - D_\delta Q_\delta) = w$$

since $-D_\delta Q_\delta = D_\delta(P^*_\delta - Q_\delta) = I - P^*_\delta$, the last from Theorem 20 of §1.8.

The final step is to observe that $\overline{x}^{-1}_\delta r_\delta = wP^*_\delta r_\delta = wv^{-1}_\delta = w\overline{v}^{-1}$ for all $w \geq 0$. This implies that for all large enough N, $(\overline{x}^{-1}_\gamma \;\; \overline{x}^0_\gamma)$ and $(\overline{v}^{-1} \;\; \overline{v}^0)$ are respectively optimal for the primal and dual, and v^{-1}_δ is the least feasible value of v^{-1} for the dual.

(b) Average Maximum Equals Maximum Average Reward Rate. Choose δ so $G^0_{\gamma\delta} \preceq 0$ for all $\gamma \in \Delta$. Thus $\max_{\gamma\in\Delta}(Ng^{-1}_{\gamma\delta} + g^0_{\gamma\delta}) = 0$ for all large enough N, say all $N \geq M$, or equivalently, $\mathcal{R}(Nv^{-1}_\delta + v^0_\delta) = (N+1)v^{-1}_\delta + v^0_\delta$ for $N \geq M$. Thus, by repeated application of the last equation, $\mathcal{R}^N(Mv^{-1}_\delta + v^0_\delta) = \mathcal{R}^{N-1}((M+1)v^{-1}_\delta + v^0_\delta) = \cdots = (M+N)v^{-1}_\delta + v^0_\delta$, $N = 1, 2, \ldots$. Now Theorem 10 of §1.7 implies that $\|\mathcal{R}^N(Mv^{-1}_\delta + v^0_\delta) - \mathcal{R}^N 0\| = O(1)$. Therefore, on combining these facts it follows that $\lim_{N\to\infty} N^{-1}\mathcal{R}^N 0 = \lim_{N\to\infty} N^{-1}\mathcal{R}^N(Mv^{-1}_\delta + v^0_\delta) = v^{-1}_\delta$ as claimed.

(c) Stochastic Systems with Strongly Connected System Graph. Let $s \neq t$ be states. Since the system graph is strongly connected, there is a chain $\mathcal{C}$ in the system graph from s to t. For each state $e \in \mathcal{C} \setminus \{t\}$, let $f_e \in \mathcal{C}$ be such that (e, f_e) is an arc in the system graph. Thus there is a decision γ such that $p(f_e|e, \gamma^e) > 0$ for each such state e. Hence $P^N_{\gamma st} > 0$ where N is the number of arcs in $\mathcal{C}$, establishing the first claim.

Now let δ^∞ be a stationary maximum-reward-rate policy, $v^{-1} \equiv v^{-1}_\delta$, s be a state that minimizes v^{-1}_s and $t \neq s$ be any other state. As shown above, since the system graph is strongly con-

nected, there is a decision γ and a positive integer N such that $P^N_{\gamma st} > 0$. Then from (5), $v^{-1} \geq P_\gamma v^{-1} \geq \cdots \geq P^N_\gamma v^{-1}$, so by the definition of s and the fact that P^N_γ is stochastic, $v_s^{-1} \geq P^N_{\gamma st} v_t^{-1} + (1 - P^N_{\gamma st}) v_s^{-1}$. Hence since $P^N_{\gamma st} > 0$, $v_s^{-1} \geq v_t^{-1}$. However, by definition of s, $v_s^{-1} \leq v_t^{-1}$, so $v_s^{-1} = v_t^{-1} = v_{\delta s}^{-1}$ is independent of s as claimed.

(d) Linear Programs for Stochastic Systems with Strongly Connected System Graph. Suppose the system is stochastic, the system graph is strongly connected and $\sum_s w_s = 1$. Now on postmultiplying (1) by the S-column vector 1 of ones and using the facts that $Q_\gamma 1 = 0$ (because P_γ is stochastic) and $w1 = 1$, yields (6) of the alternate primal. Thus, the constraint set of the alternate primal contains that of the original primal. Also, since one optimal solution $(\overline{v}^{-1} \;\; \overline{v}^0)$ of the dual takes the form $\overline{v}^{-1} = 1u^{-1}$, the last inequality (5) of the dual is redundant and the dual objective function reduces to $w1u^{-1} = u^{-1}$. Hence the dual is equivalent to the alternate dual. Further, the alternate dual is the dual of the alternate primal. Now choose δ as in (a) above. Then there are optimal solutions $(\overline{x}_\gamma^{-1} \;\; \overline{x}_\gamma^0)$ and $(\overline{v}^{-1} \;\; \overline{v}^0)$ respectively of the original primal and dual such that $\overline{x}_\gamma^{-1} = 0$ for all $\gamma \neq \delta$ and $\max_{\gamma \in \Delta} v_\gamma^{-1} = v_\delta^{-1} = 1\overline{u}^{-1}$. Thus, $\overline{x}_\gamma^{-1}$ is optimal for the alternate primal and $\overline{u}^{-1}$ is optimal and the least feasible value of u^{-1} for the alternate dual.

Homework 7 Due May 23

1. Discovering System Boundedness. Consider the following collections of linear inequalities and nonlinear equations for a discrete-time-parameter finite Markov population decision chain with $r_\delta = 1$ for all δ, viz.,

$$P_\delta v \le v \text{ and } r_\delta + P_\delta u \le u + v \text{ for all } \delta \in \Delta, \text{ and } (v,u) \ge 0 \tag{1}$$

and

$$\max_{\delta\in\Delta} P_\delta v = v,\ \max_{\delta\in\Delta}(r_\delta + P_\delta u) = u + v \text{ and } (v,u) \ge 0. \tag{2}$$

Parts (a)-(c) below establish that system boundedness is equivalent to the existence of a (v,u) satisfying either of the conditions (1) or (2). Thus system boundedness can be checked by linear programming.

(a) (1) Implies System Boundedness. Show that if there is a pair (v,u) satisfying (1), then the system is bounded. [*Hint*: Argue as in the proof of the System-Degree Theorem 9 of §1.7 for $d = 1$. In particular, let J, K be a partition of the states for which $v_J \gg 0$ and $v_K = 0$. Show that $P_{\delta KJ} = 0$ for all $\delta \in \Delta$. Then show the restriction of the system to states in K is transient by considering the second system of inequalities.]

(b) System Boundedness Implies (2). Show that if the system is bounded, then there is a (v,u) that satisfies (2). [*Hint*: There is a δ^∞ with strong maximum present value, so $G^0_{\gamma\delta} \preceq 0$ for all $\gamma \in \Delta$. Then show that $(v,u) = (v_\delta^{-1}, Mv_\delta^{-1} + v_\delta^0)$ satisfies (2) for large enough $M > 0$.]

(c) (2) Implies (1). Show that if (v,u) satisfies (2), then (v,u) satisfies (1).

2. Finding the Maximum Spectral Radius. Consider a discrete-time-parameter finite Markov population decision chain.

(a) Monotonicity of Spectral Radius. Show that if P and Q are square real matrices of the same order with $0 \le P \le Q$, then $\sigma_P \le \sigma_Q$. [*Hint*: Use Corollary 8 of §1.7.]

(b) Bounds on Spectral Radius. Show that if $P = (p_{ij})$ is a square real nonnegative matrix, then

$$m \equiv \min_i \sum_j p_{ij} \le \sigma_P \le \max_i \sum_j p_{ij} \equiv M.$$

[*Hint*: Construct matrices $\underline{P}$ and $\overline{P}$ for which $0 \le \underline{P} \le P \le \overline{P}$, the row sums of $\underline{P}$ all equal m and the row sums of $\overline{P}$ all equal M.]

(c) Bounds on Maximum Spectral Radius. Show that $m \le \sigma \le M$ where

$$m \equiv \min_s \max_a \sum_t p(t|s,a),\ M \equiv \max_{a,s} \sum_t p(t|s,a) \text{ and } \sigma \equiv \max_\delta \sigma_\delta$$

is the maximum of the spectral radii of the P_δ.

(d) Finding the Maximum Spectral Radius. Show that if $0 \le m < M$, $m \le \sigma \le M$, and θ is the midpoint of the interval $[m, M]$, then $\theta > \sigma$ if and only if the system with transition matrices $\theta^{-1}P_\delta$ for all δ is transient. Use this fact to give an iterative method for estimating σ.

3. Irreducible Systems and Cesàro Geometric-Overtaking Optimality. Consider a discrete-time-parameter finite Markov population decision chain. Let $\mathcal{R}v \equiv \max_{\delta\in\Delta} P_\delta v$. Call a real number μ an *eigenvalue* of $\mathcal{R}$ if there is a real vector v, called an associated *eigenvector*, such that $\mathcal{R}v = \mu v$. Call the system *irreducible* if the system graph $\mathcal{G}$ is *strongly connected*, i.e., there is a chain (directed path) from each state to each other state. Let $\sigma \equiv \max_{\delta\in\Delta} \sigma_\delta$ be the maximum spectral radius.

(a) System Boundedness. Show that if $\mathcal{R}$ has a unit eigenvalue ($\mu = 1$) and associated eigenvector $v \gg 0$, then the system is bounded.

(b) Unit Eigenvalue of $\mathcal{R}$ with Positive Eigenvector. Show that if the system is irreducible and bounded, but not transient, then $\mathcal{R}$ has a unit eigenvalue and associated eigenvector $v \gg 0$. [*Hint*: Use the results of parts (a)-(c) of Problem 1. Show that since the system is not transient, the set J in the hint for Problem 1(a) is not empty. Then use the irreducibility to show that the set K is empty.]

(c) Positive Maximum Spectral Radius an Eigenvalue of $\mathcal{R}$ with Positive Eigenvector. Show that if the system is irreducible, then $\sigma > 0$, σ is an eigenvalue of $\mathcal{R}$ and that there is an associated eigenvector $v \gg 0$. [*Hint*: Use the irreducibility and Problem 2(c) to show that $\sigma > 0$. Next consider the normalized system in which P_δ is replaced by $\overline{P}_\delta \equiv \sigma^{-1} P_\delta$ for each δ. Let $r_\delta = 1$ for all $\delta \in \Delta$ and V^ρ be the resulting maximum expected present value with interest rate $100\rho\%$ > 0, so that $V^\rho = \max_{\delta\in\Delta}(\beta 1 + \beta\overline{P}_\delta V^\rho) \geq \beta 1$. Let $\|u\|$ be the sum of the absolute values of the elements of u. Show that $\|V^\rho\|^{-1} V^\rho$ has a limit point $v \geq 0$ as $\rho \downarrow 0$ with $\|v\| = 1$, and $\overline{\mathcal{R}} \equiv \sigma^{-1}\mathcal{R}$ has a unit eigenvalue with v being an associated eigenvector. Then use the irreducibility to show that $v \gg 0$.]

(d) Perron-Frobenius Theorem. Show that if P_δ is irreducible, then $\sigma_\delta > 0$, σ_δ is an eigenvalue of P_δ and there is an associated eigenvector $v \gg 0$, i.e., $P_\delta v = \sigma_\delta v$.

(e) Maximum Spectral Radius is Maximum Population Growth Factor. Show that if the system is irreducible, then $\max_{\delta\in\Delta} \|P_\delta^N\| = O(\sigma^N)$. Then show that the normalized system is bounded. [*Hint*: Apply the result of part (c).]

(f) Existence of Stationary Cesàro Geometric-Overtaking Optimal Policies. Suppose that the system is irreducible. Call λ *Cesàro geometrically-overtaking optimal* if

$$\liminf_{N\to\infty} \sigma^{-N}(V_\lambda^{N,-1} - V_\pi^{N,-1}) \geq 0 \ (\mathrm{C},1) \text{ for all } \pi.$$

Show that there is a stationary Cesàro geometrically-overtaking optimal policy, viz., any stationary maximum-reward-rate policy for the normalized system.

(g) Maximum Long-Run Growth of Expected Symmetric Multiplicative Utility. Discuss the application of the result of part (f) to the problem of maximizing the long-run growth of expected symmetric multiplicative utility for an irreducible finite Markov decision chain.

Answers to Homework 7 Due May 23

1. Discovering System Boundedness.

(a) (1) Implies System Boundedness. Let J, K be a partition of the states such that $v_J \gg 0$ and $v_K = 0$, so from $P_\delta v \leq v$ it follows that $P_{\delta JJ} v_J \leq v_J$ and $P_{\delta KJ} v_J \leq 0$. Then, by induction $P^N_{\delta JJ} v_J \leq v_J$ and so $P^N_{\delta JJ} = O(1)$. Conclude from $P_{\delta KJ} v_J \leq 0$, $v_J \gg 0$ and $P_\delta \geq 0$ that $P_{\delta KJ} = 0$ and thus $P^N_{\delta KJ} = 0$ for every N. Also, since (v, u) satisfies (1), $1 + P_{\delta KK} u_K \leq u_K$, so by Problem 2(a) of Homework 3, the restriction of the system to states in K is transient, whence $P^i_{\delta KK} = O(\alpha^i)$ for some α, $0 < \alpha < 1$. Observe that, since δ is arbitrary, it follows from the System-Degree Theorem that for each policy $\pi = (\delta_1, \delta_2, \dots)$, $\max_\pi \|P^N_{\pi JJ}\| = O(1)$, $\max_\pi \|P^N_{\pi KJ}\| = 0$ for every N and $\max_\pi \|P^i_{\pi KK}\| = O(\alpha^i)$. Thus $P^N_{\pi JK} = \sum_{i=1}^N P^{i-1}_{\pi JJ} P_{\delta_i JK} P^{N-i}_{\pi_i KK}$, $\max_\pi \|P^{i-1}_{\pi JJ}\| = O(1)$ and $\max_\pi \|P^{N-i}_{\pi_i KK}\| = O(\alpha^{N-i})$, so $\max_\pi \|P^N_{\pi JK}\| = O(1)$. Hence $\max_\pi \|P^N_\pi\| = O(1)$.

(b) System Boundedness Implies (2). Since system is bounded, by Theorem 22 there is a stationary strong-maximum-present-value policy δ^∞. Then, $G_{\gamma\delta} \preceq 0$ for every γ, so $(g^{-1}_{\gamma\delta}, g^0_{\gamma\delta}) \preceq 0$, and $G_{\delta\delta} = 0$. Observe that since the rewards are positive, $V^\rho_\delta \geq 0$ for all small enough $\rho > 0$. Thus since $V^\rho_\delta = \sum_{n=-1}^\infty \rho^n v^n_\delta$, $(v^{-1}_\delta, v^0_\delta) \succeq 0$. These facts imply that $v^{-1}_\delta \geq 0$ and that there exists $M > 0$ with $\max_\gamma g^{-1}_{\gamma\delta} = 0$, $\max_\gamma (g^0_{\gamma\delta} + M g^{-1}_{\gamma\delta}) = 0$ and $M v^{-1}_\delta + v^0_\delta \geq 0$. These inequalities restate the fact that $(v, u) \equiv (v^{-1}_\delta, M v^{-1}_\delta + v^0_\delta)$ satisfies (2).

(c) (2) Implies (1). Immediate.

2. Finding the Maximum Spectral Radius.

(a) Monotonicity of Spectral Radius. The proof is by contradiction. Suppose $0 \leq P \leq Q$, but $\sigma_Q < \sigma_P$. Then $\sigma_P^{-1} Q$ has spectral radius $\sigma_P^{-1}\sigma_Q < 1$. Thus, $\sigma_P^{-1} Q$ is transient, and so $\sigma_P^{-1} P$ is also transient since $\sigma_P^{-1} P \leq \sigma_P^{-1} Q$. But $1 = \sigma_{\sigma_P^{-1} P}$, contradicting the transience of $\sigma_P^{-1} P$.

(b) Bounds on Spectral Radius. Since $m \leq \sigma_P \leq M$ is trivial if $M = 0$ and the left-hand inequality is trivial if $m = 0$, it suffices to consider the case that $0 < m \leq M$. Now observe that if Q is a square stochastic matrix, then $Q^N 1 = 1$ for $N = 1, 2, \dots$, so $d_Q = 1$, and hence $\sigma_Q = 1$ by Lemma 7. Thus by respectively reducing the elements of rows of P with sums above m and then increasing the elements of rows of P with sums below M, produces the matrices $\underline{P}$ and $\overline{P}$ such that $0 \leq \underline{P} \leq P \leq \overline{P}$, $\underline{P} 1 = m1$ and $\overline{P} 1 = M1$. Hence, $m^{-1}\underline{P}$ and $M^{-1}\overline{P}$ are stochastic and so have spectral radius one, whence $\sigma_{\underline{P}} = m$ and $\sigma_{\overline{P}} = M$. Thus, from part (a), $m \leq \sigma_P \leq M$.

(c) Bounds on the Maximum Spectral Radius. Choose γ so that $m = \min_s \sum_t p(t\,|s, \gamma^s)$. Then, by part (b) and Lemma 11 of §1, it follows that $m \leq \sigma_\gamma \leq \sigma \leq \max_\delta \|P_\delta\| = M$ where $\|\cdot\|$ is the Chebyshev norm.

(d) Finding the Maximum Spectral Radius. The θ-*normalized* system is the one with transition matrices $\theta^{-1} P_\delta$ for all δ. Observe from Lemma 7 of §1 that $\sigma < \theta$ if and only if the θ-nor-

malized system is transient, a fact that can be checked by the method of problem 2(b) of the Homework 3. If the θ-normalized system is transient, then reset $M = \theta$; otherwise, reset $m = \theta$. The resulting interval $[m, M]$ contains σ by what was just shown, and the interval is one-half the length of the original interval. Now repeat this construction iteratively until $M - m$ is small enough.

3. Irreducible Systems and Cesàro Geometric-Overtaking Optimality.

(a) System Boundedness. Let v be a positive eigenvector associated with the unit eigenvalue of $\mathcal{R}$. Then, $v = \mathcal{R}v = \cdots = \mathcal{R}^N v = \max_\pi P_\pi^N v$, so $P_\pi^N = O(1)$.

(b) Unit Eigenvalue of $\mathcal{R}$ with Positive Eigenvector. Suppose the system is bounded and irreducible, but not transient. From Problem 1, there exists $v \geq 0$ with $\max_{\delta\in\Delta} P_\delta v = v$. Thus, v is a nonnegative eigenvector associated with a unit eigenvalue. We claim v is positive. Consider a partition of the state space J, K as in 1(a). Since the system is not transient, $J \neq \emptyset$ because, as shown in 1(a), the restriction of the system to states in K is transient. Furthermore, $K = \emptyset$ for if not, $P_{\pi KJ}^N = 0$ for every N as in 1(a). But this contradicts the irreducibility of the system. Thus, $v = v_J \gg 0$.

(c) Positive Maximum Spectral Radius an Eigenvalue of $\mathcal{R}$ with Positive Eigenvector. Since the system is irreducible, for every state s there exists an action a such that $\sum_t p(t \mid s, a) > 0$. Thus by 2(c), $\sigma \geq m = \min_s \max_a \sum_t p(t \mid s, a) > 0$. Consider the normalized system $\overline{P}_\delta = \sigma^{-1} P_\delta$. This system is not transient because $\bar{\sigma} = 1$. Let $\bar{r}_\delta = 1$ for all $\delta \in \Delta$. Then the maximum present value V^ρ with interest rate $100\rho\% > 0$ is such that $\lim_{\rho\downarrow 0} \|V^\rho\| = \infty$ because in the contrary event V^ρ is bounded and so the system is transient. But since all rewards are nonnegative, $V^\rho = \max_{\delta\in\Delta} \{\beta 1 + \beta \overline{P}_\delta V^\rho\} \geq \beta 1 \gg 0$. Now multiplying both sides by $\|V^\rho\|^{-1}$ yields $V^\rho \|V^\rho\|^{-1} = \max_{\delta\in\Delta} [\beta 1 \|V^\rho\|^{-1} + \beta \overline{P}_\delta V^\rho \|V^\rho\|^{-1}]$. Thus since $V^\rho \|V^\rho\|^{-1}$ has unit norm, it has a limit point $v \geq 0$ as ρ converges to zero. Also $\|v\| = 1$ and $v = \max_{\delta\in\Delta} \overline{P}_\delta v = \max_{\delta\in\Delta} \sigma^{-1} P_\delta v$ or equivalently, $\sigma v = \mathcal{R} v$. This shows that σ is an eigenvalue of $\mathcal{R}$ and v is an associated eigenvector. Observe that this implies $P_\delta v \leq \sigma v$ for all δ. Let J, K partition the state space so that $v_K = 0$ and $v_J \gg 0$. Since $\|v\| = 1$, necessarily $J \neq \emptyset$. And, if $K \neq \emptyset$, then it follows from $P_\delta v \leq \sigma v$ that $P_{\delta KJ} v_J \leq 0$ for all δ, which implies $P_{\delta KJ} = 0$, contradicting the irreducibility of the system. Thus $v \gg 0$.

(d) Perron-Frobenius Theorem. This is part (c) when $\Delta = \{\delta\}$.

(e) Maximum Spectral Radius is Maximum Population Growth Rate. Since the system is irreducible, $\sigma > 0$ (from (c)) and there exists $v \gg 0$ such that $\sigma v = \mathcal{R} v \geq P_\delta v$ for all $\delta \in \Delta$, so $\max_\pi P_\pi^N v \leq \sigma^N v$, whence $\max_\pi \|P_\pi^N\| = O(\sigma^N)$. Furthermore, $\max_\pi \|\overline{P}_\pi^N\| = \sigma^{-N} \max_\pi \|P_\pi^N\| = O(1)$ and so the normalized system is bounded.

(f) Existence of Stationary Cesàro Geometric-Overtaking Optimal Policies. By (e) the normalized system is bounded. Then by Theorem 32 there exists a stationary Cesàro-overtaking-optimal policy δ^∞ for the binomial immigration stream of order -1, i.e.,

$$\liminf_{N\to\infty} (\overline{V}_\delta^{N,-1} - \overline{V}_\pi^{N,-1}) \geq 0 \ (\mathrm{C},1)$$

for all $\pi = (\gamma_1, \gamma_2, \dots)$ where $\overline{V}_\pi^{N,-1} = \overline{P}_\pi^{N-1} r_{\gamma_N} = \sigma^{-N+1} P_\pi^{N-1} r_{\gamma_N} = \sigma^{-N+1} V_\pi^{N,-1}$. Thus,

$$\liminf_{N\to\infty} \sigma^{-N}(V_\delta^{N,-1} - V_\pi^{N,-1}) \geq 0 \ (\mathrm{C},1) \text{ for all } \pi.$$

(g) Maximum Long-Run Growth of Expected Symmetric Multiplicative Utility. From equation (6) of §1.4 of *Lectures*..., the maximum expected utility V^N in periods $1, \dots, N$ satisfies $V^N = \max_{\delta\in\Delta} P_\delta V^{N-1} = \max_{\pi\in\Delta^\infty} P_\pi^N V^0$ where $V^0 = 1$. Let $r_\delta = 1$ for all δ. Then, $V_\pi^{N,-1} = P_\pi^{N-1} 1$. Since the system is irreducible, apply the result of (f) to find a policy that maximizes the long-run growth of expected symmetric utility.

Homework 8 Due May 30

1. Element-Wise Product of Symmetric Positive Semi-Definite Matrices. Show that if C is an $m \times m$ symmetric positive semi-definite matrix that is partitioned symmetrically into blocks C_{ij} with C_{ii} symmetric positive definite for $i, j = 1, \ldots, n \leq m$, and if $Q = (q_{ij})$ is an $n \times n$ symmetric positive definite matrix, then the matrix H composed of blocks $q_{ij}C_{ij}$ is symmetric and positive definite. [*Hint*: Observe that since C is a symmetric positive semi-definite matrix, there is a diagonal matrix $\Lambda = \text{diag}\,(\lambda_1, \ldots, \lambda_m) \geq 0$ of eigenvalues of C and a matrix V with column m-vectors $v_1, \ldots, v_m$ such that $C = V\Lambda V^{\mathrm{T}} = \sum_{k=1}^{m} \lambda_k v_k v_k^{\mathrm{T}}$. One can partition each m-element row vector $x^{\mathrm{T}} = (x_1^{\mathrm{T}}, \ldots, x_n^{\mathrm{T}})$ corresponding to the partitioning of C, and in particular $v_k^{\mathrm{T}} = (v_{k1}^{\mathrm{T}}, \ldots, v_{kn}^{\mathrm{T}})$. Let $w_k = (x_i^{\mathrm{T}} v_{ki})$ be an n-element column vector for each k. Show that $x^{\mathrm{T}} H x = \sum_{k=1}^{m} \lambda_k w_k^{\mathrm{T}} Q w_k$.]

2. Quadratic Unconstrained Team-Decision Problem with Normally Distributed Observations. Consider the quadratic unconstrained team-decision problem in which Q is symmetric and positive definite, and the random column vectors $Z_1, \ldots, Z_n$ have a joint normal distribution with $\mathrm{E}Z_i = 0$, C_{ij} denoting the matrix of covariances between the components of Z_i and Z_j, and C_{ii} being an identity matrix for $i, j = 1, \ldots, n$. You may assume the known (e.g., Feller, Vol. II, 80-87) facts that the covariance matrix $C = (C_{ij})$ is symmetric and positive semi-definite, and that $\mathrm{E}^{Z_i} Z_j = C_{ji} Z_i$ for $i, j = 1, \ldots, n$. Also assume that $\mathrm{E}^{Z_i} \delta_i = d_i Z_i + \mathrm{E}\delta_i$ for some row vector d_i, $i = 1, \ldots, n$. Show that there is an optimal team decision function X^* that is affine, i.e., $X_i^*(Z_i) = b_i Z_i + c_i$ where the row vectors b_i and numbers c_i, $i = 1, \ldots, n$, are the unique solutions of certain systems of linear equations. You may assume that X^* is optimal if, with probability one, $\mathrm{E}^{Z_i} \delta_i = q_i \mathrm{E}^{Z_i} X^*$, $i = 1, \ldots, n$. [*Hint*: Apply the result of Problem 1 about element-wise products of symmetric positive semi-definite matrices.]

3. Optimal Baking. At the beginning of a day, n bakeries make forecasts $Z_1, \ldots, Z_n$ of their respective demands for bread during the day. The forecasts are random variables with known means. After observing only its own forecast, bakery i bakes X_i loaves of bread, $i = 1, \ldots, n$. The actual demands $D_1, \ldots, D_n$ for bread at bakeries $1, \ldots, n$ are random variables. Assume that D_i and Z_i have finitely many values, that forecasts are unbiased, i.e., $\mathrm{E}(D_i | Z_i) = Z_i$ for all i, and that (D_i, Z_i) and (D_j, Z_j) are independent for each $i \neq j$. The loss $L(X, Z, D)$, where $X = (X_i)$, $Z = (Z_i)$, and $D = (D_i)$, is

$$L(X, Z, D) = \frac{1}{2} \sum_{i=1}^{n} a_i (X_i - D_i)^2 + \frac{1}{2} (\sum_{i=1}^{n} (X_i - D_i) - b)^2$$

where $a_i > 0$, $i = 1, \ldots, n$ and $b > 0$. The problem is to choose a baking policy from the above class that minimizes the expected one-period loss. Solve the problem explicitly. [*Hint*: Consider affine functions of the form $X_i = \alpha_i Z_i + \beta_i$ and find the coefficients α_i and β_i.]

Answers to Homework 8 Due May 30

1. Element-Wise Product of Symmetric Positive Semi-Definite Matrices. Observe that since C is an $m \times m$ symmetric positive semi-definite matrix, there exists a matrix V whose column m-vectors are $v_1, \ldots, v_m$ and a diagonal matrix $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_m) \geq 0$ of eigenvalues of C such that $C = V\Lambda V^T = \sum_{k=1}^m \lambda_k v_k v_k^T$. On partitioning each m-element row vector $x^T = (x_1^T, \ldots, x_n^T)$, $m \geq n$, corresponding to the partitioning of C, and in particular doing so for $v_k^{\mathrm{T}} = (v_{k1}^{\mathrm{T}}, \ldots, v_{kn}^{\mathrm{T}})$, then each block $C_{ij} = \sum_{k=1}^m \lambda_k v_{ki} v_{kj}^T$. Let $w_k = (x_i^{\mathrm{T}} v_{ki})$ be an n-element column vector for each k. Therefore, since $H_{ij} = q_{ij} C_{ij}$, it follows that

$$x^T H x = \sum_{i,j} x_i^T q_{ij} \left(\sum_{k=1}^m \lambda_k v_{ki} v_{kj}^T \right) x_j = \sum_{k=1}^m \lambda_k \sum_{i,j} w_k^i q_{ij} w_k^j = \sum_{k=1}^m \lambda_k w_k^T Q w_k \geq 0$$

since Q is positive definite. But from the fact that C_{ii} is symmetric positive definite, it follows that for every $x_i \neq 0$,

$$0 < x_i^T C_{ii} x_i = \sum_{k=1}^m \lambda_k x_i^T v_{ki} v_{ki}^T x_i = \sum_{k=1}^m \lambda_k (w_k^i)^2.$$

Thus, there exists at least one value of k for which $\lambda_k > 0$ and $w_k \neq 0$. Hence, for $x \neq 0$, there is an i such that $x_i \neq 0$ and so $x^T H x = \sum_{k=1}^m \lambda_k w_k^T Q w_k > 0$.

2. Quadratic Unconstrained Team-Decision Problem with Normally Distributed Observations. The optimal solution X^* must satisfy $\mathrm{E}^{Z_i} \delta_i = q_i \mathrm{E}^{Z_i} X^*$ with probability one for $i = 1, \ldots, n$. But by assumption, $\mathrm{E}^{Z_i} \delta_i = d_i Z_i + \mathrm{E}\delta_i$. If $X_i^*(Z_i) = b_i Z_i + c_i$, then

$$d_i Z_i + \mathrm{E}\delta_i = \sum_{j=1}^n q_{ij} \mathrm{E}^{Z_i} X_j^* = \sum_{j=1}^n q_{ij}(b_j C_{ji} Z_i + c_j) = \sum_{j=1}^n (b_j q_{ji} C_{ji} Z_i + q_{ij} c_j)$$

since Q is symmetric. Let H be as in Problem 1. Then, $d_i Z_i + \mathrm{E}\delta_i = b H_i Z_i + q_i c$ for all i and all Z_i where $b = (b_1, \ldots, b_n)$, $c^T = (c_1, \ldots, c_n)$ and $H_i^T = (H_{i1}, \ldots, H_{in})$. Thus it must be that $\mathrm{E}\delta = Qc$ and $d = bH$. These systems have the unique solutions $c = Q^{-1}\mathrm{E}\delta$ and $b = dH^{-1}$ since Q is symmetric positive definite and, by Problem 1, so is H.

3. Optimal Baking. Write the problem in the form of a Quadratic Unconstrained Team Decision Problem by defining $Q = \operatorname{diag} a + 11^T$ and writing the objective function as $r(X, Z) = -\frac{1}{2} X^T Q X + \delta X +$ constant term where $\delta_i = a_i D_i + b + \sum_{j=1}^n D_j$. Observe that since $a_i > 0$ for all i, Q is positive definite. By Theorem 1 of §2.3, X^* is optimal if and only if $\mathrm{E}^{Z_i} \delta_i = \mathrm{E}^{Z_i} q_i X^*$, i.e.,

$$\mathrm{E}^{Z_i}(a_i D_i + b + \sum_{j=1}^n D_j) = \mathrm{E}^{Z_i}\Big\{ (\sum_{j \neq i} X_j) + (1 + a_i) X_i \Big\},$$

and this is equivalent to

$$a_i Z_i + b + Z_i + \sum_{j \neq i} \mathrm{E} D_j = (\sum_{j \neq i} \mathrm{E} X_j) + (1 + a_i) X_i$$

since $\mathrm{E}^{Z_i} D_i = Z_i$, $\mathrm{E}^{Z_i} D_j = \mathrm{E} D_j$, $\mathrm{E}^{Z_i} X_i = X_i$ and $\mathrm{E}^{Z_i} X_j = \mathrm{E} X_j$. If $X_i = \alpha_i Z_i + \beta_i$, find α_i, β_i such that X_i is the solution to the system above. Do this by substituting X_i in the right-hand side and observing that $\mathrm{E} D_i = \mathrm{E} Z_i$ for all i, since $\mathrm{E} D_i = \mathrm{E}(\mathrm{E}^{Z_i} D_i) = \mathrm{E} Z_i$. Then, this new equivalent system has solution $\alpha_i = 1$ and $\beta_i = \dfrac{b}{a_i(1 + \sum_k a_k^{-1})}$ and thus $X_i^* = Z_i + \beta_i$ is the optimal solution.

Homework 9 Due June 4

1. Revenue Management: Pricing a House for Sale. An executive must sell her house in an interval $[0,t]$. Her goal is to maximize her expected sale price. She is willing to vary her asking price as a piecewise-constant function of time, with her asking price at any moment restricted to a given finite set P of positive numbers. Whenever her asking price is $p \in P$, buyers willing to pay that price arrive according to a poisson process with mean rate $q_p > 0$ per unit time, with q_p being strictly decreasing in p. She sells the house to the first buyer arriving during $[0,t]$ at her then current asking price, or if no buyer arrives, to her company at the guaranteed price p^* with $0 < p^* < \underline{p} \equiv \min P$. Let V^t be the maximum expected sale price in $[0,t]$.

(a) Optimality Equation. Give an equation whose solution is (V^t).

(b) Existence of Optimal Asking-Price Policy. Explain why there exists a maximum-expected-asking-price policy.

(c) Monotonicity and Concavity of Maximum Expected Sale Price in Time Remaining. Show that V^t is positive, strictly increasing and concave in $t \geq 0$.

(d) Monotonicity of Optimal Asking-Price in Time Remaining. Show that with one optimal asking-price policy, the asking price p_t used when t time units remain has the properties that:

1° $p_t = \overline{p} \equiv \max P$ for all large enough $t > 0$ and

2° p_t is increasing in $t > 0$.

(e) Limit of Maximum t-Period Expected Sale Price. Show that $\lim_{t\to\infty} V^t = \overline{p}$.

2. Transient Systems in Continuous Time. Show that if every stationary policy is transient in a continuous-time-parameter finite Markov population decision chain, then every policy is transient and there is a stationary maximum-value policy. Explain briefly how successive approximations can be used to find a stationary policy having value within $\epsilon > 0$ of the maximum. [*Hint*: For the first part, adapt the proof of the corresponding result for the discrete-time-parameter problem.]