Stochastic Optimization

Anton J. Kleywegt and Alexander Shapiro

August 24, 2003

School of Industrial and Systems Engineering,
Georgia Institute of Technology,
Atlanta, Georgia 30332-0205, USA

Contents

1 Introduction

Decision makers often have to make decisions in the presence of uncertainty. Decision problems are often formulated as optimization problems, and thus in many situations decision makers wish to solve optimization problems that depend on parameters which are unknown. Typically it is quite difficult to formulate and solve such problems, both conceptually and numerically. The difficulty already starts at the conceptual stage of modeling. Often there are a variety of ways in which the uncertainty can be formalized. In the formulation of optimization problems, one usually attempts to find a good trade-off between the realism of the optimization model, which affects the usefulness and quality of the obtained decisions, and the tractability of the problem, so that it can be solved analytically or numerically. As a result of these considerations there are a large number of different approaches for formulating and solving optimization problems under uncertainty. It is impossible to give a complete survey of all such methods in one article. Therefore this article aims only to give a flavor of prominent approaches to optimization under uncertainty.

2 Optimization under Uncertainty

To describe some issues involved in optimization under uncertainty, we start with a static optimization problem. Suppose we want to maximize an objective function $G(x, \omega)$, where x denotes the decision to be made, $\mathcal{X}$ denotes the set of all feasible decisions, ω denotes an outcome that is unknown at the time the decision has to be made, and Ω denotes the set of all possible outcomes.

There are several approaches for dealing with optimization under uncertainty. In Section 2.1, some of these approaches are illustrated in the context of an example.

2.1 Example

Example 2.1 (Newsvendor Problem) Many companies sell seasonal products, such as fashion articles, airline seats, Christmas decorations, magazines and newspapers. These products are characterized by a relatively short selling season, after which the value of the products decrease substantially. Often, a decision has to be made how much of such a product to manufacture or purchase before the selling season starts. Once the selling season has started, there is not enough time remaining in the season to change this decision and implement the change, so that at this stage the quantity of the product is given. During the season the decision maker may be able to make other types of decisions to pursue desirable results, such as to change the price of the product as the season progresses and sales of the product takes place. Such behavior is familiar in many industries. Another characteristic of such a situation is that the decisions have to be made before the eventual outcomes become known to the decision maker. For example, the decision maker has to decide how much of the product to manufacture or purchase before the demand for the product becomes known. Thus decisions have to be made without knowing which outcome will take place.

Suppose that a manager has to decide how much of a seasonal product to order. Thus the decision variable $x \in \mathbb{R}_+$ is the order quantity. The cost of the product to the company is c per unit of the product. During the selling season the product can be sold at a price (revenue) of r per unit of the product. After the selling season any remaining product can be disposed of at a salvage value of s per unit of the product, where typically $s < r$. The demand D for the product is unknown at the time the order decision x has to be made. If the demand D turns out to be greater than the order quantity x, then the whole quantity x of the product is sold during the season, and no product remains at the end of the season, so that the total revenue turns out to be rx. If the demand D turns out to be less than the order quantity x, then quantity D of the product is sold during the season, and the remaining amount of product at the end of the season is $x - D$, so that the total revenue turns out to be $rD + s(x - D)$. Thus the profit is given by

$$G(x, D) = \begin{cases} rD + s(x - D) - cx & \text{if} \quad D \leq x \\ rx - cx & \text{if} \quad D > x \end{cases} \tag{2.1}$$

The manager would like to choose x to maximize the profit $G(x, D)$, but the dilemma is that D is unknown at the time the decision should be made. This problem is often called the *newsvendor problem*.

Note that if $r \leq c$, then the company can make no profit from buying and selling the product, so that the optimal order quantity is $x^* = 0$, irrespective of what the demand D turns out to be. Also, if $s \geq c$, then any unsold product at the end of the season can be disposed of at a value at least equal to the cost of the product, so that it is optimal to order as much as possible, irrespective of what the demand D turns out to be. For the remainder of this example, we assume that $s < c < r$.

Under this assumption, for any given $D \geq 0$, the function $G(\cdot, D)$ is a piecewise linear function with positive slope $r - c$ for $x < D$ and negative slope $s - c$ for $x > D$. Therefore, if the demand D is known at the time the order decision has to be made, then the best decision is to choose order quantity $x^* = D$.

However, if D is not known at the time the decision should be made, then the problem becomes more difficult. There are several approaches to decision making in the usual case where the demand is not known. Sometimes a manager may want to hedge against the worst possible outcome. Suppose the manager thinks that the demand D will turn out to be some number in the interval $[a, b] \subset \mathbb{R}_+$, with $a < b$, i.e., the lower and upper bounds for the demand are known to the manager. In that case, in order to hedge against the worst possible scenario, the manager chooses the value of x that gives the best profit under the worst possible outcome. For any decision x, the worst possible outcome is given by

$$g_1(x) \equiv \min_{D \in [a,b]} G(x, D) = G(x, a) = \begin{cases} (r-s)a - (c-s)x & \text{if} \quad a \leq x \\ (r-c)x & \text{if} \quad a > x \end{cases}$$

Because the manager wants to hedge against the worst possible outcome, the manager chooses the value of x that gives the best profit under the worst possible outcome, that is, the manager chooses the value of x that maximizes $g_1(x)$, which is $x_1 = a$. Clearly, in many cases this will be an overly conservative decision.

Sometimes a manager may want to make the decision that under the worst possible outcome will still appear as good as possible compared with what would have been the best decision with hindsight, that is after the outcome becomes known. For any outcome of the demand D, let

$$g^*(D) \equiv \max_{x \in \mathbb{R}_+} G(x, D) = (r - c)D$$

denote the optimal profit with hindsight, also called the optimal value with perfect information. The optimal decision with perfect information, $x^* = D$, is sometimes called the wait-and-see solution. Suppose the manager chose to order quantity x, so that the actual profit turned out to be $G(x, D)$. The amount of profit that the company missed out on because of a suboptimal decision is given by $g^*(D) - G(x, D)$. This quantity,

$$A(x, D) \equiv g^*(D) - G(x, D) = \begin{cases} (c-s)(x-D) & \text{if} \quad D \leq x \\ (r-c)(D-x) & \text{if} \quad D > x \end{cases}$$

is often called the absolute regret. The manager may want to choose the value of x that minimizes the absolute regret under the worst possible outcome. For any decision x, the worst possible outcome is given by

$$\begin{aligned} g_2(x) &\equiv \max_{D \in [a,b]} A(x, D) = \max\{(c-s)(x-a), (r-c)(b-x)\} \\ &= \max\{A(x, a), A(x, b)\} = \begin{cases} (r-c)(b-x) & \text{if} \quad x \leq \frac{(c-s)}{(r-s)}a + \frac{(r-c)}{(r-s)}b \\ (c-s)(x-a) & \text{if} \quad x > \frac{(c-s)}{(r-s)}a + \frac{(r-c)}{(r-s)}b \end{cases} \end{aligned}$$

Because the manager wants to choose the value of x that minimizes the absolute regret under the worst possible outcome, the manager chooses the value of x that minimizes $g_2(x)$, which is $x_2 = [(c-s)a + (r-c)b]/(r-s)$. Note that x_2 is a convex combination of a and b, and thus $a < x_2 < b$. The larger the salvage loss per unit $c - s$, the closer x_2 is to a, and the larger the profit per unit $r - c$, the closer x_2 is to b. That seems to be a more reasonable decision than $x_1 = a$, but it will be shown that in many cases one can easily obtain a better solution than x_2.

A similar approach is to choose the value of x that minimizes the relative regret $R(x, D)$ under the worst possible outcome, where

$$R(x, D) \equiv \frac{g^*(D) - G(x, D)}{g^*(D)} = \begin{cases} \frac{(c-s)(x-D)}{(r-c)D} = \frac{(c-s)}{(r-c)}\left(\frac{x}{D} - 1\right) & \text{if} \quad D \leq x \\ \frac{(r-c)(D-x)}{(r-c)D} = 1 - \frac{x}{D} & \text{if} \quad D > x \end{cases}$$

For any decision x, the worst possible outcome is given by

$$\begin{aligned} g_3(x) &\equiv \max_{D\in[a,b]} R(x,D) = \max\left\{\frac{(c-s)}{(r-c)}\left(\frac{x}{a}-1\right), 1-\frac{x}{b}\right\} \\ &= \max\{R(x,a),R(x,b)\} = \begin{cases} 1-\frac{x}{b} & \text{if} \quad x \le \frac{ab}{\frac{(r-c)}{(r-s)}a+\frac{(c-s)}{(r-s)}b} \\ \frac{(c-s)}{(r-c)}\left(\frac{x}{a}-1\right) & \text{if} \quad x > \frac{ab}{\frac{(r-c)}{(r-s)}a+\frac{(c-s)}{(r-s)}b} \end{cases} \end{aligned}$$

The manager then chooses the value of x that minimizes $g_3(x)$, which is $x_3 = ab/\{[(r-c)a+(c-s)b]/(r-s)\}$. Note that $[(r-c)a+(c-s)b]/(r-s)$ in the denominator of the expression for x_3 is a convex combination of a and b, and thus $a < x_3 < b$. Similar to x_2, the larger the salvage loss per unit $c-s$, the closer x_3 is to a, and the larger the profit per unit $r-c$, the closer x_3 is to b.

A related approach is to choose the value of x that maximizes the competitive ratio $\rho(x,D)$ under the worst possible outcome, where

$$\rho(x,D) \equiv \frac{G(x,D)}{g^*(D)}$$

Because $\rho(x,D) = 1 - R(x,D)$, maximizing the competitive ratio $\rho(x,D)$ is equivalent to minimizing the relative regret $R(x,D)$, so that this approach leads to the same solution x_3 as the previous approach.

It was assumed in all the variants of the worst-case approach discussed above that no a priori information about the demand D was available to the manager except the lower and upper bounds for the demand. In some situations this may be a reasonable assumption and the worst-case approach could make sense if the range of the demand is known and is not "too large". However, in many applications the range of the unknown quantities is not known with useful precision, and other information, such as information about the probability distributions or sample data of the unknown quantities, may be available.

Another approach to decision making under uncertainty, different from the worst-case approaches described above, is the stochastic optimization approach, which is the approach that most of this article is focused on. Suppose that the demand D can be viewed as a random variable with a known, or at least well estimated, probability distribution. The corresponding cumulative distribution function (cdf) F can be estimated from historical data or by using a priori information available to the manager. Then one can try to optimize the objective function on average, i.e. to maximize the expected profit

$$\begin{aligned} g(x) &\equiv \mathbb{E}[G(x,D)] \\ &= \int_0^x [rw + s(x-w)]\, dF(w) + \int_x^\infty rx\, dF(w) - cx \end{aligned} \tag{2.2}$$

This optimization problem is easy to solve. For any $D \ge 0$, the function $G(x,D)$ is concave in x. Therefore, the expected value function g is also concave. First, suppose the demand D has a probability density function (pdf). Then

$$g'(x) = sF(x) + r(1-F(x)) - c \tag{2.3}$$

Recalling that g is concave, it follows that the expected profit $g(x)$ is maximized where $g'(x) = 0$, that is at x^*, where x^* satisfies

$$F(x^*) = \frac{r-c}{r-s}$$

Because $s < c < r$, it follows that $0 < (r-c)/(r-s) < 1$, so that a value of x^* that satisfies $F(x^*) = (r-c)/(r-s)$ can always be found. If the demand D does not have a pdf, a similar result still holds. In general

$$x^* = F^{-1}\left(\frac{r-c}{r-s}\right)$$

where

$$F^{-1}(p) \quad \equiv \quad \min\{x \,:\, F(x) \geq p\}$$

Another point worth mentioning is that by solving (2.2), the manager tries to optimize the profit on average. However, the realized profit $G(x^*, D)$ could be very different from the corresponding expected value $g(x^*)$, depending on the particular realization of the demand D. This may happen if $G(x^*, D)$, considered as a random variable, has a large variability which could be measured by its variance $\mathrm{Var}[G(x^*, D)]$. Therefore, if the manager wants to hedge against such variability he may consider the following optimization problem

$$\max_{x \geq 0} \left\{ g_\beta(x) \equiv I\!E[G(x, D)] - \beta \,\mathrm{Var}[G(x, D)] \right\} \tag{2.4}$$

The coefficient $\beta \geq 0$ represents the weight given to the conservative part of the decision. If β is large, then the above optimization problem tries to find a solution with minimal profit variance, while if $\beta = 0$, then problem (2.4) coincides with problem (2.2). Note that since the variance $\mathrm{Var}[G(x, D)] \equiv I\!E[(G(x, D) - I\!E[G(x, D)])^2]$ is itself an expected value, from a mathematical point of view problem (2.4) is similar to the expected value problem (2.2). Thus, the problem of optimizing the expected value of an objective function $G(x, D)$ is very general—it could include the means, variances, quantiles, and almost any other aspects of random variables of interest.

The following deterministic optimization approach is often used for decision making under uncertainty. The random variable D is replaced by its mean $\mu = I\!E[D]$, and then the following deterministic optimization problem is solved.

$$\max_{x \in I\!R_+} G(x, \mu)$$

A resulting optimal solution $\bar{x}$ is sometimes called an expected value solution. Of course, this approach requires that the mean of the random variable D be known to the decision maker. In the present example, the optimal solution of this deterministic optimization problem is $\bar{x} = \mu$. Note that the two solutions, the $(r - c)/(r - s)$-quantile x^* and the mean $\bar{x}$, can be very different. Also, it is well known that the quantiles are much more stable to variations of the cdf F than the corresponding mean value. It typically happens that an optimal solution x^* of the stochastic optimization problem is more robust with respect to variations of the probability distributions than an optimal solution $\bar{x}$ of the corresponding deterministic optimization problem. Also note that, for any x, $G(x, D)$ is concave in D. As a result it follows from Jensen's inequality that $G(x, I\!E[D]) \geq I\!E[G(x, D)]$, and thus the objective function of the deterministic optimization problem is biased upward relative to the objective function of the stochastic optimization problem, and the optimal value of the deterministic optimization problem is biased upward relative to the optimal value of the stochastic optimization problem, because $\max_{x \in I\!R_+} G(x, I\!E[D]) \geq \max_{x \in I\!R_+} I\!E[G(x, D)]$.

One can also try to solve the optimization problem

$$\max_{x \in I\!R_+} G(x, D)$$

for particular realizations of D, and then take the expected value of the obtained solutions as the final solution. In the present example, for any realization D, the optimal solution of this problem is $x = D$, and hence the expected value of these solutions, and final solution, is $\mu = \bar{x}$. Note that in many optimization problems it may not make sense to take the expected value of the obtained solutions. This is usually the case in optimization problems with discrete solutions, for example, when a solution is a path in a network, there does not seem to be a useful way to take the average of several different paths.

2.2 Summary of Approaches for Decision Making Under Uncertainty

In this section we summarize several approaches often used for decision making under uncertainty, as introduced in the example of Section 2.1.

2.2.1 Worst-case Approaches

Hedging Against the Worst-case Outcome The chosen decision x_1 is obtained by solving the following optimization problem.

$$\sup_{x \in \mathcal{X}} \inf_{\omega \in \Omega} G(x, \omega)$$

Minimizing the Absolute Regret The chosen decision x_2 is obtained by solving the following optimization problem.

$$\inf_{x \in \mathcal{X}} \sup_{\omega \in \Omega} \{g^*(\omega) - G(x, \omega)\}$$

Minimizing the Relative Regret The chosen decision x_3 is obtained by solving the following optimization problem.

$$\inf_{x \in \mathcal{X}} \sup_{\omega \in \Omega} \frac{g^*(\omega) - G(x, \omega)}{g^*(\omega)}$$

assuming $g^*(\omega) > 0$ for all $\omega \in \Omega$. An equivalent approach is to choose the solution x_3 that maximizes the competitive ratio, as given by the following optimization problem.

$$\sup_{x \in \mathcal{X}} \inf_{\omega \in \Omega} \frac{G(x, \omega)}{g^*(\omega)}$$

2.2.2 The Stochastic Optimization Approach

The chosen decision x^* is obtained by solving the following optimization problem.

$$\sup_{x \in \mathcal{X}} \{g(x) \equiv I\!E[G(x, \omega)]\} \tag{2.5}$$

2.2.3 The Deterministic Optimization Approach

The chosen decision $\bar{x}$ is obtained by solving the following optimization problem.

$$\sup_{x \in \mathcal{X}} G(x, I\!E[\omega]) \tag{2.6}$$

2.3 Evaluation Criteria for the Stochastic Optimization Approach

Next we introduce some criteria that are useful for evaluating the stochastic optimization approach to decision making under uncertainty. The optimal value with perfect information is given by

$$g^*(\omega) \equiv \sup_{x \in \mathcal{X}} G(x, \omega)$$

Thus the expected value with perfect information is given by $I\!E[g^*(\omega)]$. Also, the expected value of an optimal solution x^* of the stochastic optimization problem 2.5 is given by

$$g(x^*) \equiv \sup_{x \in \mathcal{X}} I\!E[G(x, \omega)]$$

Note that

$$g(x^*) \equiv \sup_{x \in \mathcal{X}} I\!E[G(x, \omega)] \leq I\!E\left[\sup_{x \in \mathcal{X}} G(x, \omega)\right] \equiv I\!E[g^*(\omega)]$$

The difference, $I\!E[g^*(\omega)] - g(x^*) = I\!E[A(x^*, \omega)]$, is often called the value of perfect information.

It is also interesting to compare $g(x^*)$ with the value obtained from the deterministic optimization problem 2.6. The expected value of an optimal solution $\bar{x}$ of the deterministic optimization problem is given by $g(\bar{x}) \equiv I\!E[G(\bar{x}, \omega)]$. Note that

$$g(x^*) \equiv \sup_{x \in \mathcal{X}} I\!E[G(x, \omega)] \geq I\!E[G(\bar{x}, \omega)] \equiv g(\bar{x})$$

The difference, $g(x^*) - g(\bar{x})$, is sometimes called the value of the stochastic solution.

2.4 Estimation of Probability Distributions

The stochastic optimization approach usually involves the assumption that the probability distribution of the unknown outcome is known. However, in practice, the probability distribution is usually not known. One way to deal with this situation is to estimate a distribution from data, assuming that the data is relevant for the decision problem, and then to use the estimated distribution in the stochastic optimization problem. There are several approaches to estimate probability distributions from data.

A simple and versatile estimate of a probability distribution is the empirical distribution. Suppose we want to estimate the cumulative distribution function (cdf) F of a random variable W, and we have a data set $W_1, W_2, \ldots, W_k$ of k observations of W. Let $N(w)$ denote the number of observations that have value less than or equal to w. Then the empirical cumulative distribution function is given by $\hat{F}_k(w) \equiv N(w)/k$. Let $W_{1:k}, W_{2:k}, \ldots, W_{k:k}$ denote the order statistics of the k observations of W, that is, $W_{1:k}$ is the smallest among $W_1, W_2, \ldots, W_k$; $W_{2:k}$ is the second smallest among $W_1, W_2, \ldots, W_k$; $\ldots$; $W_{k:k}$ is the largest among $W_1, W_2, \ldots, W_k$. Then, for any $i \in \{1, 2, \ldots, k\}$, and any $p \in ((i-1)/k, i/k]$, $\hat{F}_k^{-1}(p) = W_{i:k}$. Also, assuming that $W_1, W_2, \ldots, W_k$ are independent and identically distributed with cdf F, it follows that the cdf $F_{i:k}$ of $W_{i:k}$ is given by

$$F_{i:k}(w) = \sum_{j=i}^{k} \binom{k}{j} F(w)^j [1 - F(w)]^{k-j}$$

Further, if W has a probability density function (pdf) f, then it follows that the pdf $f_{i:k}$ of $W_{i:k}$ is given by

$$f_{i:k}(w) = i \binom{k}{i} f(w) F(w)^{i-1} [1 - F(w)]^{k-i}$$

Use of the empirical distribution, and its robustness, are illustrated in an example in Section 2.5.

If there is reason to believe that the random variable W follows a particular type of probability distribution, for example a normal distribution, with one or more unknown parameters, for example the mean μ and variance σ^2 of the normal distribution, then standard statistical techniques, such as maximum likelihood estimation, can be used to estimate the unknown parameters of the distribution from data. Also, in such a situation, a Bayesian approach can be used to estimate the unknown parameters of the distribution from data, to optimize an objective function that is related to that of the stochastic optimization problem. More details can be found in Berger (1985).

2.5 Example (continued)

In this section the use of the empirical distribution, and its robustness, are illustrated with the newsvendor example of Section 2.1.

Suppose the manager does not know the probability distribution of the demand, but a data set $D_1, D_2, \ldots, D_k$ of k independent and identically distributed observations of the demand D is available. As before, let $D_{1:k}, D_{2:k}, \ldots, D_{k:k}$ denote the order statistics of the k observations of D. Using the empirical distribution $\hat{F}_k$, the resulting decision rule is simple. If $(r-c)/(r-s) \in ((i-1)/k, i/k]$ for some $i \in \{1, 2, \ldots, k\}$, then

$$\hat{x} = \hat{F}_k^{-1}\left(\frac{r-c}{r-s}\right) = D_{i:k}$$

That is, the chosen order quantity $\hat{x}$ is the ith smallest observation $D_{i:k}$ of the demand.

To illustrate the robustness of the solution $\hat{x}$ obtained with the empirical distribution, suppose that, unknown to the manager, the demand D is exponentially distributed with rate λ, that is, the mean demand is $I\!E[D] = 1/\lambda$. The objective function is given by

$$g(x) \equiv I\!E[G(x,D)] = \frac{r-s}{\lambda}\left(1-e^{-\lambda x}\right) - (c-s)x$$

The pdf $f_{i:k}$ of $D_{i:k}$ is given by

$$\begin{aligned} f_{i:k}(w) &= i\binom{k}{i}\lambda e^{-(k-i+1)\lambda w}\left(1-e^{-\lambda w}\right)^{i-1} \\ &= i\binom{k}{i}\lambda\sum_{j=0}^{i-1}\binom{i-1}{j}(-1)^j e^{-(k-i+j+1)\lambda w} \end{aligned}$$

The expected objective value of the chosen order quantity $\hat{x} = D_{i:k}$ is given by (assuming that $D_1, D_2, \ldots, D_k$ and D are i.i.d. $\exp(\lambda)$)

$$\begin{aligned} I\!E[G(D_{i:k},D)] &= I\!E\left[\frac{r-s}{\lambda}\left(1-e^{-\lambda D_{i:k}}\right) - (c-s)D_{i:k}\right] \\ &= \int_0^\infty \left[\frac{r-s}{\lambda}\left(1-e^{-\lambda w}\right) - (c-s)w\right] i\binom{k}{i}\lambda\sum_{j=0}^{i-1}\binom{i-1}{j}(-1)^j e^{-(k-i+j+1)\lambda w}\,dw \\ &= \left[(r-s)\sum_{j=0}^{i}\binom{i}{j}(-1)^j\frac{1}{k-i+j+1}\right. \\ &\quad \left. -(c-s)\sum_{j=0}^{i-1}\binom{i-1}{j}(-1)^j\frac{1}{(k-i+j+1)^2}\right] i\binom{k}{i}I\!E[D] \end{aligned}$$

Next we compare the objective values of several solutions, including the optimal value with perfect information, $I\!E[G(D_{i:k},D)]$, $g(x^*)$, and $g(\bar{x})$. Recall that the optimal value with perfect information is given by

$$g^*(D) \equiv \max_{x\in I\!R_+} G(x,D) = (r-c)D$$

Thus the expected value with perfect information is given by

$$I\!E[g^*(D)] = (r-c)I\!E[D]$$

Also, the optimal solution x^* of the stochastic optimization problem is given by

$$x^* = F^{-1}\left(\frac{r-c}{r-s}\right) = -\ln\left(\frac{c-s}{r-s}\right)I\!E[D]$$

and the optimal objective function value is given by

$$g(x^*) = \left[(r-c) + (c-s)\ln\left(\frac{c-s}{r-s}\right)\right]I\!E[D]$$

Thus the value of perfect information is

$$I\!E[g^*(D)] - g(x^*) = -(c-s)\ln\left(\frac{c-s}{r-s}\right)I\!E[D] = -\frac{c-s}{r-s}\ln\left(\frac{c-s}{r-s}\right)(r-s)I\!E[D]$$

It is easy to obtain bounds on the value of perfect information. Consider the function $h(y) \equiv y \ln(y)$ for $y > 0$. Then $h'(y) = \ln(y) + 1$ and $h''(y) = 1/y > 0$, because $y > 0$. Thus h is convex on $(0, \infty)$, and $h(y)$ attains a minimum of $-1/e$ when $y = 1/e$. Also, $\lim_{y \to 0} h(y) = 0$, and $h(1) = 0$. Hence the value of perfect information attains a minimum of zero when $c = s$ and when $c = r$. This makes sense from previous results, since the optimal decisions when $c \leq s$ (x^* as large as possible) or when $c \geq r$ ($x^* = 0$) do not depend on information about the demand. Also, the value of perfect information attains a maximum of $(r - s)\mathbb{E}[D]/e$ when $(c - s)/(r - s) = 1/e$, i.e., when the ratio of profit per unit to the salvage loss per unit $(r - c)/(c - s) = e - 1$.

The optimal solution $\bar{x}$ of the deterministic optimization problem (2.6) is $\bar{x} = \mathbb{E}[D]$. The expected value of this solution is given by

$$g(\bar{x}) \equiv \mathbb{E}[G(\bar{x}, D)] = \left[(r - c) - \frac{r - s}{e}\right] \mathbb{E}[D]$$

Hence the value of the stochastic solution is given by

$$g(x^*) - \mathbb{E}[G(\bar{x}, D)] = \left[\frac{c - s}{r - s} \ln\left(\frac{c - s}{r - s}\right) + \frac{1}{e}\right] (r - s)\mathbb{E}[D]$$

It follows from the properties of $h(y) \equiv y \ln(y)$ that the value of the stochastic solution attains a minimum of zero when the value of perfect information attains a maximum, i.e., when $(c - s)/(r - s) = 1/e$. Also, the value of the stochastic solution attains a maximum of $(r - s)\mathbb{E}[D]/e$ when the value of perfect information attains a minimum, i.e., when $c = s$ and when $c = r$, that is, when using the expected demand $\mathbb{E}[D]$ gives the poorest results.

Next we evaluate the optimality gaps of several solutions. Let $\theta \equiv (r - c)/(r - s) \in ((i - 1)/k, i/k]$ for some $i \in \{1, 2, \ldots, k\}$. Then the optimality gap of the solution based on the empirical distribution is given by

$$\begin{aligned}
\gamma_k^e(\theta) &\equiv \frac{g(x^*) - \mathbb{E}[G(D_{i:k}, D)]}{(r - s)\mathbb{E}[D]} \\
&= \theta + (1 - \theta)\ln(1 - \theta) - \left[\sum_{j=0}^{i} \binom{i}{j} (-1)^j \frac{1}{k - i + j + 1}\right. \\
&\qquad \left. -(1 - \theta) \sum_{j=0}^{i-1} \binom{i - 1}{j} (-1)^j \frac{1}{(k - i + j + 1)^2}\right] i \binom{k}{i}
\end{aligned}$$

Note that the division by $\mathbb{E}[D]$ can be interpreted as rescaling the product units so that $\mathbb{E}[D] = 1$, and the division by $r - s$ can be interpreted as rescaling the money units so that $r - s = 1$. The optimality gap of the optimal solution $\bar{x}$ of the deterministic optimization problem is given by

$$\begin{aligned}
\gamma^d(\theta) &\equiv \frac{g(x^*) - g(\bar{x})}{(r - s)\mathbb{E}[D]} \\
&= (1 - \theta)\ln(1 - \theta) + \frac{1}{e}
\end{aligned}$$

To evaluate the worst-case solutions x_1, x_2, and x_3, suppose that the interval $[a, b]$ is taken as $[0, \beta\mathbb{E}[D]]$ for some $\beta > 0$. Then $x_1 = a = 0$, and thus $g(x_1) = 0$, and the optimality gap of the worst-case solution x_1 is given by

$$\begin{aligned}
\gamma_1(\theta) &\equiv \frac{g(x^*) - g(x_1)}{(r - s)\mathbb{E}[D]} \\
&= \theta + (1 - \theta)\ln(1 - \theta)
\end{aligned}$$

Also, $x_2 = [(c-s)a + (r-c)b]/(r-s) = (r-c)\beta I\!E[D]/(r-s) = \theta\beta I\!E[D]$, and thus

$$g(x_2) \;=\; \left[\left(1-e^{-\theta\beta}\right) - (1-\theta)\theta\beta\right](r-s)I\!E[D]$$

and the optimality gap of the absolute regret solution x_2 is given by

$$\begin{aligned}\gamma_2(\theta) &\equiv \frac{g(x^*)-g(x_2)}{(r-s)I\!E[D]} \\ &= \theta + (1-\theta)\ln(1-\theta) - \left[\left(1-e^{-\theta\beta}\right) - (1-\theta)\theta\beta\right]\end{aligned}$$

Also, $x_3 = ab/\{[(r-c)a + (c-s)b]/(r-s)\} = 0$, and thus $g(x_3) = 0$, and the optimality gap of the relative regret solution x_3 is $\gamma_3(\theta) = \gamma_1(\theta)$.

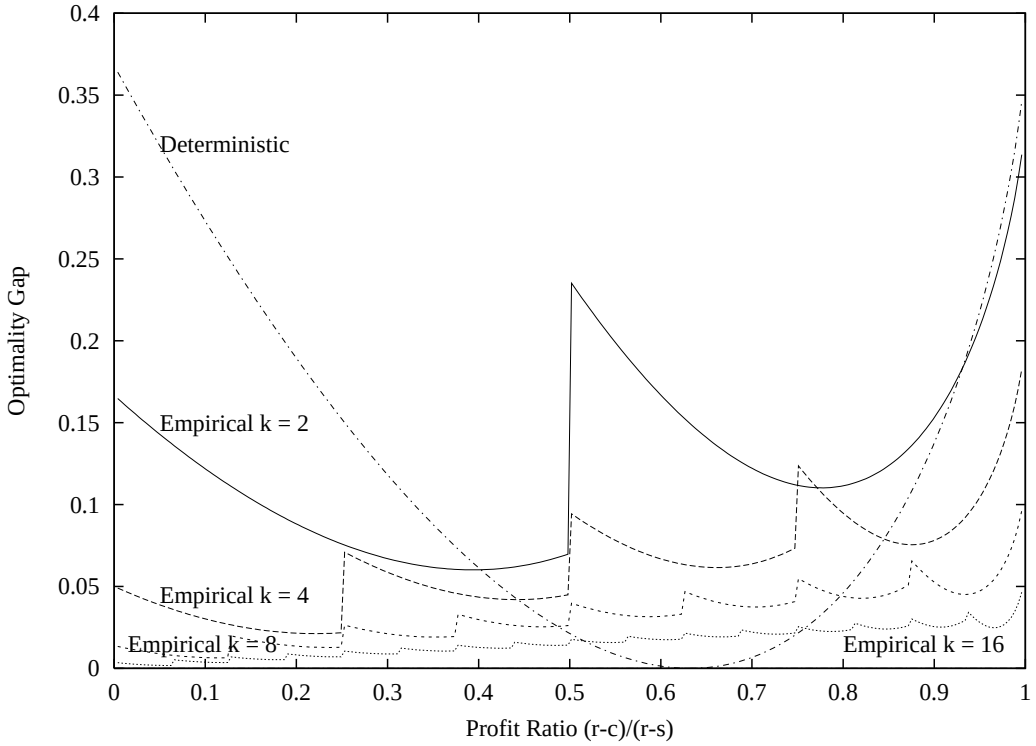

Figure 1: Optimality gaps $\gamma_k^e(\theta)$ of the empirical approach for $k = 2, 4, 8, 16$, as well as the optimality gap $\gamma^d(\theta)$ of the deterministic approach, as a function of $\theta \equiv (r-c)/(r-s)$.

Figure 1 shows the optimality gaps $\gamma_k^e(\theta)$ for $k = 2, 4, 8, 16$, as well as the optimality gap $\gamma^d(\theta)$ as a function of θ. It can be seen that the empirical solutions $\hat{x}$ tend to be more robust, in terms of the optimality gap, than the expected value solution $\bar{x}$, even if the empirical distribution is based on a very small sample size k. Only in the region where $\theta \approx 1 - 1/e$, i.e., where the value of the stochastic solution is small, does $\bar{x}$ give a good solution. It should also be kept in mind that the solution $\bar{x}$ is based on knowledge of the expected demand, whereas the empirical solutions do not require such knowledge, but the empirical solutions in turn require a data set of demand observations. Figure 2 shows the optimality gaps $\gamma_1(\theta)$, $\gamma_2(\theta)$, and $\gamma_3(\theta)$ for $\beta = 1, 2, 3, 4, 5$, as a function of θ. Solutions x_1 and x_3 do not appear to be very robust. Also, only when β is chosen to be close to 2 does the absolute regret solution x_2 appear to have robustness that compares well with the robustness of the empirical solutions. The performance of the absolute regret solution x_2 appears to be quite sensitive to the choice of β. Furthermore, a decision maker is not likely to know what is the best choice of β.

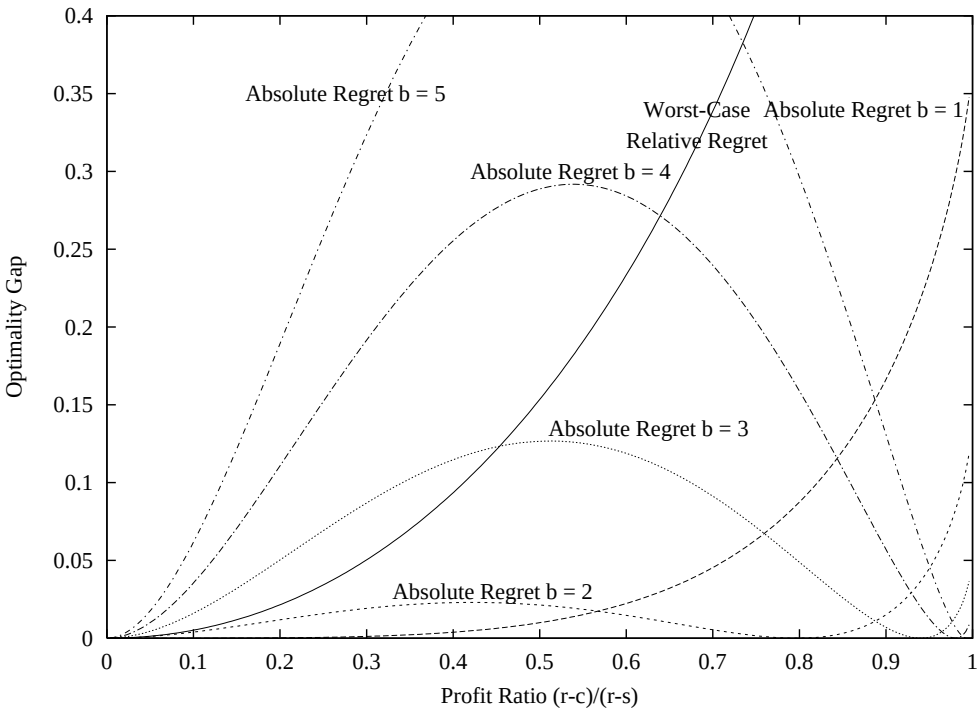

Figure 2: Optimality gaps $\gamma_1(\theta)$, $\gamma_2(\theta)$, and $\gamma_3(\theta)$ of the worst-case approaches, for $\beta = 1, 2, 3, 4, 5$, as a function of $\theta \equiv (r-c)/(r-s)$.

3 Stochastic Programming

The discussion of the above example motivates us to introduce the following model optimization problem, referred to as a *stochastic programming* problem,

$$\inf_{x \in \mathcal{X}} \big\{ g(x) \equiv I\!E[G(x, \omega)] \big\} \tag{3.1}$$

(We consider a minimization rather than a maximization problem for the sake of notational convenience.) Here $\mathcal{X} \subset I\!R^n$ is a set of permissible values of the vector x of decision variables, and is referred to as the feasible set of problem (3.1). Often $\mathcal{X}$ is defined by a (finite) number of smooth (or even linear) constraints. In some other situations the set $\mathcal{X}$ is finite. In that case problem (3.1) is called a *discrete* stochastic optimization problem (this should not be confused with the case of discrete probability distributions). Variable ω represents random (or stochastic) aspects of the problem. Often ω can be modeled as a finite dimensional random vector, or in more involved cases as a random process. In the abstract framework we can view ω as an element of the probability space $(\Omega, \mathcal{F}, P)$ with the known probability measure (distribution) P.

It is also possible to consider the following extensions of the basic problem (3.1).

- One may need to optimize a function of the expected value function $g(x)$. This happened, for example, in problem (2.4), where the manager wanted to optimize a linear combination of the expected value and the variance of the profit. Although important from a modeling point of view, such an extension usually does not introduce additional technical difficulties into the problem.

- The feasible set can also be defined by constraints given in a form of expected value functions. For example, suppose that we want to optimize an objective function subject to the constraint that the event $\{h(x, W) \geq 0\}$, where W is a random vector with a known probability distribution and $h(\cdot, \cdot)$ is a given function, should happen with a probability not bigger than a given number $p \in (0, 1)$. Probability

of this event can be represented as the expected value $I\!E[\psi(x, W)]$, where

$$\psi(x, w) \equiv \begin{cases} 1 & \text{if} \quad h(x, w) \geq 0 \\ 0 & \text{if} \quad h(x, w) < 0 \end{cases}$$

Therefore, this constraint can be written in the form $I\!E[\psi(x, W)] \leq p$. Problems with such probabilistic constraints are called *chance constrained* problems. Note that even if the function $h(\cdot, \cdot)$ is continuous, the corresponding indicator function $\psi(\cdot, \cdot)$ is discontinuous unless it is identically equal to zero or one. Because of that, it may be technically difficult to handle such a problem.

- In some cases the involved probability distribution P_θ depends on parameter vector θ, whose components also represent decision variables. That is, the expected value objective function is given in the form

$$g(x, \theta) \equiv I\!E_\theta[G(x, \omega)] = \int_\Omega G(x, \omega)\, dP_\theta(\omega) \tag{3.2}$$

By using a transformation it is sometimes possible to represent the above function $g(\cdot)$ as the expected value of a function, depending on x and θ, with respect to a probability distribution that is independent of θ. We shall discuss such likelihood ratio transformations in Section 3.4.

The above formulation of stochastic programs is somewhat too general and abstract. In order to proceed with a useful analysis we need to identify particular classes of such problems that on one hand are interesting from the point of view of applications and on the other hand are computationally tractable. In the following sections we introduce several classes of such problems and discuss various techniques for their solution.

3.1 Stochastic Programming with Recourse

Consider again problem (2.2) of the newsvendor example. We may view that problem as a two-stage problem. At the first stage a decision should be made about the quantity x to order. At this stage the demand D is not known. At the second stage a realization of the demand D becomes known and, given the first stage decision x, the manager makes a decision about the quantities y and z to sell at prices r and s, respectively. Clearly the manager would like to choose y and z in such a way as to maximize the profit. It is possible to formulate the second stage problem as the simple linear program

$$\max_{y,z} ry + sz \quad \text{subject to} \quad y \leq D,\ y + z \leq x,\ y \geq 0,\ z \geq 0 \tag{3.3}$$

The optimal solution of the above problem (3.3) is $y^* = \min\{x, D\}$, $z^* = \max\{x - D, 0\}$, and its optimal value is the profit $G(x, D)$ defined in (2.1). Now at the first stage, before a realization of the demand D becomes known, the manager chooses a value for the first stage decision variable x by maximizing the expected value of the second stage optimal profit $G(x, D)$.

This is the basic idea of a two-stage stochastic program with recourse. At the first stage, before a realization of the random variables ω becomes known, one chooses the first stage decision variables x to optimize the expected value $g(x) \equiv I\!E[G(x, \omega)]$ of an objective function $G(x, \omega)$ that depends on the optimal second stage objective function $Q(x, \xi(\omega))$.

A *two-stage stochastic linear program with fixed recourse* is a two-stage stochastic program with the form

$$\begin{array}{ll} \min_x & c^T x + I\!E[Q(x, \xi)] \\ \text{s.t.} & Ax = b,\ x \geq 0 \end{array} \tag{3.4}$$

where $Q(x, \xi)$ is the optimal value of the second stage problem

$$\begin{array}{ll} \min_y & q(\omega)^T y \\ \text{s.t.} & T(\omega)x + Wy = h(\omega),\ y \geq 0 \end{array} \tag{3.5}$$

The second stage problem depends on the data $\xi(\omega) \equiv (q(\omega), h(\omega), T(\omega))$, elements of which can be random, while the matrix W is assumed to be known beforehand. The matrices $T(\omega)$ and W are called the *technology* and *recourse* matrices, respectively. The expectation $I\!E[Q(x,\xi)]$ is taken with respect to the random vector $\xi = \xi(\omega)$, whose probability distribution is assumed to be known. The above formulation originated in the works of Dantzig (1955) and Beale (1955).

Note that the optimal solution $y^* = y^*(\omega)$ of the second stage problem (3.5) depends on the random data $\xi = \xi(\omega)$, and therefore is random. One can write $Q(x, \xi(\omega)) = q(\omega)^T y^*(\omega)$.

The next question is how one can solve the above two-stage problem numerically. Suppose that the random data have a *discrete* distribution with a finite number K of possible realizations $\xi_k = (q_k, h_k, T_k)$, $k = 1, ..., K$, (sometimes called *scenarios*), with the corresponding probabilities p_k. In that case $I\!E[Q(x,\xi)] = \sum_{k=1}^K p_k Q(x, \xi_k)$, where

$$Q(x, \xi_k) \;=\; \min\left\{q_k^T y_k : T_k x + W y_k = h_k,\ y_k \geq 0\right\}$$

Therefore, the above two-stage problem can be formulated as one large linear program:

$$\begin{array}{ll} \min & c^T x + \sum_{k=1}^K p_k q_k^T y_k \\ \text{s.t.} & Ax = b \\ & T_k x + W y_k = h_k \\ & x \geq 0,\ y_k \geq 0,\ k = 1, \ldots, K \end{array} \tag{3.6}$$

The linear program (3.6) has a certain block structure that makes it amenable to various decomposition methods. One such decomposition method is the popular L-shaped method developed by Van Slyke and Wets (1969). We refer the interested reader to the recent books by Kall and Wallace (1994) and Birge and Louveaux (1997) for a thorough discussion of stochastic programming with recourse.

The above numerical approach works reasonably well if the number K of scenarios is not too large. Suppose, however, that the random vector ξ has m independently distributed components each having just 3 possible realizations. Then the total number of different scenarios is $K = 3^m$. That is, the number of scenarios grows exponentially fast in the number m of random variables. In that case, even for a moderate number of random variables, say $m = 100$, the number of scenarios becomes so large that even modern computers cannot cope with the required calculations. It seems that the only way to deal with such exponential growth of the number of scenarios is to use sampling. Such approaches are discussed in Section 3.2.

It may also happen that some of the decision variables at the first or second stage are integers, such as binary variables representing "yes" or "no" decisions. Such integer (or discrete) stochastic programs are especially difficult to solve, and only very moderate progress has been reported so far. A discussion of two-stage stochastic integer programs with recourse can be found in Birge and Louveaux (1997). A branch and bound approach for solving stochastic discrete optimization problems was suggested by Norkin, Pflug and Ruszczyński (1998). Schultz, Stougie and Van der Vlerk (1998) suggested an algebraic approach for solving stochastic programs with integer recourse by using a framework of Gröbner basis reductions. For a recent survey of mainly theoretical results on stochastic integer programming see Klein Haneveld and Van der Vlerk (1999).

Conceptually the idea of two-stage programming with recourse can be readily extended to *multistage* programming with recourse. Such an approach tries to model the situation where decisions are made periodically (in stages) based on currently known realizations of some of the random variables. An H-stage stochastic linear program with fixed recourse can be written in the form

$$\begin{array}{ll} \min & c^1 x^1 + I\!E\left\{\min c^2(\omega) x^2(\omega) + \cdots + I\!E[\min c^H(\omega) x^H(\omega)]\right\} \\ \text{s.t.} & W^1 x^1 = h^1 \\ & T^1(\omega) x^1 + W^2 x^2(\omega) = h^2(\omega) \\ & \qquad \cdots\cdots \\ & T^{H-1}(\omega) x^{H-1}(\omega) + W^H x^H(\omega) = h^H(\omega) \\ & x^1 \geq 0, x^2(\omega) \geq 0, \ldots, x^H(\omega) \geq 0 \end{array} \tag{3.7}$$

The decision variables $x^2(\omega), \ldots, x^H(\omega)$ are allowed to depend on the random data ω. However, the decision $x^t(\omega)$ at time t can only depend on the part of the random data that is known at time t (these restrictions

are often called nonanticipativity constraints). The expectations are taken with respect to the distribution of the random variables whose realizations are not yet known.

Again, if the distribution of the random data is discrete with a finite number of possible realizations, then problem (3.7) can be written as one large linear program. However, it is clear that even for a small number of stages and a moderate number of random variables the total number of possible scenarios will be astronomical. Therefore, a current approach to such problems is to generate a "reasonable" number of scenarios and to solve the corresponding (deterministic) linear program, hoping to catch at least the flavor of the stochastic aspect of the problem. The argument is that the solution obtained in this way is more robust than the solution obtained by replacing the random variables with their means.

Often the same practical problem can be modeled in different ways. For instance, one can model a problem as a two-stage stochastic program with recourse, putting all random variables whose realizations are not yet known at the second stage of the problem. Then as realizations of some of the random variables become known, the solutions are periodically updated in a two-stage rolling horizon fashion, every time by solving an updated two-stage problem. Such an approach is different from a multistage program with recourse, where every time a decision is to be made, the modeler tries to take into account that decisions will be made at several stages in the future.

3.2 Sampling Methods

In this section we discuss a different approach that uses Monte Carlo sampling techniques to solve stochastic optimization problems.

Example 3.1 Consider a stochastic process I_t, $t = 1, 2, \ldots$, governed by the recursive equation

$$I_t \;=\; [I_{t-1} + x_t - D_t]^+ \tag{3.8}$$

with initial value I_0. Here D_t are random variables and x_t represent decision variables. (Note that $[a]^+ \equiv \max\{a, 0\}$.) The above process I_t can describe the waiting time of the tth customer in a $G/G/1$ queue, where D_t is the interarrival time between the $(t-1)$th and tth customers and x_t is the service time of $(t-1)$th customer. Alternatively, I_t may represent the inventory of a certain product at time t, with D_t and x_t representing the demand and production (or ordering) quantities, respectively, of the product at time t.

Suppose that the process is considered over a finite horizon with time periods $t = 1, \ldots, T$. Our goal is to minimize (or maximize) the expected value of an objective function involving $I_1, \ldots, I_T$. For instance, one may be interested in maximizing the expected value of a profit given by (Albritton, Shapiro and Spearman 1999)

$$\begin{array}{rcl} G(x, W) & \equiv & \sum_{t=1}^{T} \{\pi_t \min[I_{t-1} + x_t, D_t] - h_t I_t\} \\ & = & \sum_{t=1}^{T} \pi_t x_t + \sum_{t=1}^{T-1} (\pi_{t+1} - \pi_t - h_t) I_t + \pi_1 I_0 - (\pi_T + h_T) I_T \end{array} \tag{3.9}$$

Here $x = (x_1, \ldots, x_T)$ is a vector of decision variables, $W = (D_1, \ldots, D_T)$ is a random vector of the demands at periods $t = 1, \ldots, T$, and π_t and h_t are nonnegative parameters representing the marginal profit and the holding cost, respectively, of the product at period t.

If the initial value I_0 is sufficiently large, then with probability close to one, variables $I_1, \ldots, I_T$ stay above zero. In that case $I_1, \ldots, I_T$ become linear functions of the random data vector W, and hence components of the random vector W can be replaced by their means. However, in many practical situations the process I_t hits zero with high probability over the considered horizon T. In such cases the corresponding expected value function $g(x) \equiv I\!\!E[G(x, W)]$ cannot be written in a closed form. One can use a Monte Carlo simulation procedure to evaluate $g(x)$. Note that for any given realization of D_t, the corresponding values of I_t, and hence the value of $G(x, W)$, can be easily calculated using the iterative formula (3.8).

That is, let $W^i = (D_1^i, \ldots, D_T^i)$, $i = 1, \ldots, N$, be a random (or pseudorandom) sample of N independent realizations of the random vector W generated by computer, i.e., there are N generated realizations of the demand process D_t, $t = 1, 2, \ldots, T$, over the horizon T. Then for any given x the corresponding expected

value $g(x)$ can be approximated (estimated) by the sample average

$$\hat{g}_N(x) \equiv \frac{1}{N}\sum_{i=1}^{N} G(x, W^i) \tag{3.10}$$

We have that $I\!E[\hat{g}_N(x)] = g(x)$, and by the Law of Large Numbers, that $\hat{g}_N(x)$ converges to $g(x)$ with probability one (w.p.1) as $N \to \infty$. That is, $\hat{g}_N(x)$ is an *unbiased* and *consistent* estimator of $g(x)$.

Any reasonably efficient method for optimizing the expected value function $g(x)$, say by using its sample average approximations, is based on estimation of its first (and maybe second) order derivatives. This has an independent interest and is called *sensitivity* or *perturbation* analysis. We will discuss that in Section 3.3. Recall that $\nabla g(x) \equiv (\partial g(x)/\partial x_1, \ldots, \partial g(x)/\partial x_T)$ is called the gradient vector of $g(\cdot)$ at x.

It is possible to consider a stationary distribution of the process I_t (if it exists), and to optimize the expected value of an objective function with respect to the stationary distribution. Typically, such a stationary distribution cannot be written in a closed form, and is difficult to compute accurately. This introduces additional technical difficulties into the problem. Also, in some situations the probability distribution of the random variables D_t is given in a parametric form whose parameters are decision variables. We will discuss dealing with such cases later.

3.3 Perturbation Analysis

Consider the expected value function $g(x) \equiv I\!E[G(x, \omega)]$. An important question is under which conditions the first order derivatives of $g(x)$ can be taken inside the expected value, that is, under which conditions the equation

$$\nabla g(x) \equiv \nabla I\!E[G(x, \omega)] = I\!E[\nabla_x G(x, \omega)] \tag{3.11}$$

is correct. One reason why this question is important is the following. Let $\omega^1, \ldots, \omega^N$ denote a random sample of N independent realizations of the random variable with common probability distribution P, and let

$$\hat{g}_N(x) \equiv \frac{1}{N}\sum_{i=1}^{N} G(x, \omega^i) \tag{3.12}$$

be the corresponding sample average function. If the interchangeability equation (3.11) holds, then

$$I\!E[\nabla \hat{g}_N(x)] = \frac{1}{N}\sum_{i=1}^{N} I\!E\left[\nabla_x G(x, \omega^i)\right] = \frac{1}{N}\sum_{i=1}^{N} \nabla I\!E[G(x, \omega^i)] = \nabla g(x) \tag{3.13}$$

and hence $\nabla \hat{g}_N(x)$ is an unbiased and consistent estimator of $\nabla g(x)$.

Let us observe that in both examples 2.1 and 3.1 the function $G(\cdot, \omega)$ is piecewise linear for any realization of ω, and hence is not everywhere differentiable. The same holds for the optimal value function $Q(\cdot, \xi)$ of the second stage problem (3.5). If the distribution of the corresponding random variables is discrete, then the resulting expected value function is also piecewise linear, and hence is not everywhere differentiable.

On the other hand expectation with respect to a continuous distribution typically smoothes the corresponding function and in such cases equation (3.11) often is applicable. It is possible to show that if the following two conditions hold at a point x, then $g(\cdot)$ is differentiable at x and equation (3.11) holds:

(i) The function $G(\cdot, \omega)$ is differentiable at x w.p.1.

(ii) There exists a positive valued random variable $K(\omega)$ such that $I\!E[K(\omega)]$ is finite and the inequality

$$|G(x_1, \omega) - G(x_2, \omega)| \leq K(\omega)\|x_1 - x_2\| \tag{3.14}$$

holds w.p.1 for all x_1, x_2 in a neighborhood of x.

If the function $G(\cdot,\omega)$ is not differentiable at x w.p.1 (i.e., for P-almost every $\omega \in \Omega$), then the right hand side of equation (3.11) does not make sense. Therefore, clearly the above condition (i) is necessary for (3.11) to hold. Note that condition (i) requires $G(\cdot,\omega)$ to be differentiable w.p.1 at the given (fixed) point x and does not require differentiability of $G(\cdot,\omega)$ everywhere. The second condition (ii) requires $G(\cdot,\omega)$ to be continuous (in fact Lipschitz continuous) w.p.1 in a neighborhood of x.

Consider, for instance, function $G(x,D)$ of example 2.1 defined in (2.1). For any given D the function $G(\cdot,D)$ is piecewise linear and differentiable at every point x except at $x = D$. If the cdf $F(\cdot)$ of D is continuous at x, then the probability of the event $\{D = x\}$ is zero, and hence the interchangeability equation (3.11) holds. Then $\partial G(x,D)/\partial x$ is equal to $s - c$ if $x > D$, and is equal to $r - c$ if $x < D$. Therefore, if $F(\cdot)$ is continuous at x, then $G(\cdot,D)$ is differentiable at x and

$$g'(x) \;=\; (s-c)\mathbb{P}(D < x) + (r-c)\mathbb{P}(D > x)$$

which gives the same equation as (2.3). Note that the function $\partial G(\cdot,D)/\partial x$ is discontinuous at $x = D$. Therefore, the second order derivative of $\mathbb{E}[G(\cdot,D)]$ cannot be taken inside the expected value. Indeed, the second order derivative of $G(\cdot,D)$ is zero whenever it exists. Such behavior is typical in many interesting applications.

Let us calculate the derivatives of the process I_t, defined by the recursive equation (3.8), for a particular realization of the random variables D_t. Let τ_1 denote the first time that the process I_t hits zero, i.e., $\tau_1 \geq 1$ is the first time $I_{\tau_1 - 1} + x_{\tau_1} - D_{\tau_1}$ becomes less than or equal to zero, and hence $I_{\tau_1} = 0$. Let $\tau_2 > \tau_1$ be the second time that I_t hits zero, etc. Note that if $I_{\tau_1+1} = 0$, then $\tau_2 = \tau_1 + 1$, etc. Let $1 \leq \tau_1 < \cdots < \tau_n \leq T$ be the sequence of hitting times. (In queueing terminology, τ_i represents the starting time of a new busy cycle of the corresponding queue.) For a given time $t \in \{1,...,T\}$, let $\tau_{i-1} \leq t < \tau_i$. Suppose that the events $\{I_{\tau-1} + x_\tau - D_\tau = 0\}$, $\tau = 1,...,T$, occur with probability zero. Then, for almost every W, the gradient of I_s with respect to the components of vector x_t can be written as follows

$$\nabla_{x_t} I_s \;=\; \begin{cases} 1 & \text{if } t \leq s < \tau_i \text{ and } t \neq \tau_{i-1} \\ 0 & \text{otherwise} \end{cases} \tag{3.15}$$

Thus, by using equations (3.9) and (3.15), one can calculate the gradient of the sample average function $\hat{g}_N(\cdot)$ of example 3.1, and hence one can consistently estimate the gradient of the expected value function $g(\cdot)$.

Consider the process I_t defined by the recursive equation (3.8) again. Suppose now that variables x_t do not depend on t, and let x denote their common value. Suppose further that D_t, $t = 1,...,$ are independently and identically distributed with mean $\mu > 0$. Then for $x < \mu$ the process I_t is stable and has a stationary (steady state) distribution. Let $g(x)$ be the steady state mean (the expected value with respect to the stationary distribution) of the process $I_t = I_t(x)$. By the theory of regenerative processes it follows that for every $x \in (0,\mu)$ and any realization (called sample path) of the process D_t, $t = 1,...,$ the long run average $\hat{g}_T(x) \equiv \sum_{t=1}^{T} I_t(x)/T$ converges w.p.1 to $g(x)$ as $T \to \infty$. It is possible to show that $\nabla \hat{g}_T(x)$ also converges w.p.1 to $\nabla g(x)$ as $T \to \infty$. That is, by differentiating the long run average of a sample path of the process I_t we obtain a consistent estimate of the corresponding derivative of the steady state mean $g(x)$. Note that $\nabla I_t(x) = t - \tau_{i-1}$ for $\tau_{i-1} \leq t < \tau_i$, and hence the derivative of the long run average of a sample path of the process I_t can be easily calculated.

The idea of differentiation of a sample path of a process in order to estimate the corresponding derivative of the steady state mean function by a single simulation run is at the heart of the so-called *infinitesimal perturbation analysis*. We refer the interested reader to Glasserman (1991) and Ho and Cao (1991) for a thorough discussion of that topic.

3.4 Likelihood Ratio Method

The Monte Carlo sampling approach to derivative estimation introduced in Section 3.3 does not work if the function $G(\cdot,\omega)$ is discontinuous or if the corresponding probability distribution also depends on decision variables. In this section we discuss an alternative approach to derivative estimation known as the *likelihood ratio* (or *score function*) method.

Suppose that the expected value function is given in the form $g(\theta) \equiv \mathbb{E}_\theta[G(W)]$, where W is a random vector whose distribution depends on the parameter vector θ. Suppose further that the distribution of W has a probability density function (pdf) $f(\theta, w)$. Then for a chosen pdf $\phi(w)$ we can write

$$\mathbb{E}_\theta[G(W)] \;=\; \int G(w) f(\theta, w)\, dw \;=\; \int G(w) \frac{f(\theta, w)}{\phi(w)} \phi(w)\, dw$$

and hence

$$g(\theta) \;=\; \mathbb{E}_\phi[G(Z) L(\theta, Z)] \tag{3.16}$$

where $L(\theta, z) \equiv f(\theta, z)/\phi(z)$ is the so-called likelihood ratio function, $Z \sim \phi(\cdot)$ and $\mathbb{E}_\phi[\cdot]$ means that the expectation is taken with respect to the pdf ϕ. We assume in the definition of the likelihood ratio function that $0/0 = 0$ and that the pdf ϕ is such that if $\phi(w)$ is zero for some w, then $f(\theta, w)$ is also zero, i.e., we do not divide a positive number by zero.

The expected value in the right hand side of (3.16) is taken with respect to the distribution ϕ which does not depend on the vector θ. Therefore, under appropriate conditions ensuring interchangeability of the differentiation and integration operators, we can write

$$\nabla g(\theta) \;=\; \mathbb{E}_\phi[G(Z) \nabla_\theta L(\theta, Z)] \tag{3.17}$$

In particular, if for a given θ_0 we choose $\phi(\cdot) \equiv f(\theta_0, \cdot)$, then $\nabla_\theta L(\theta, z) = \nabla_\theta f(\theta, z)/f(\theta_0, z)$, and hence $\nabla_\theta L(\theta_0, z) = \nabla_\theta \ln[f(\theta_0, z)]$. The function $\nabla_\theta \ln[f(\theta, z)]$ is called the score function, which motivates the name of this technique.

Now by generating a random sample $Z^1, \ldots, Z^N$ from the pdf $\phi(\cdot)$, one can estimate $g(\theta)$ and $\nabla g(\theta)$ by the respective sample averages

$$\tilde{g}_N(\theta) \;\equiv\; \frac{1}{N} \sum_{i=1}^{N} G(Z^i) L(\theta, Z^i) \tag{3.18}$$

$$\nabla \tilde{g}_N(\theta) \;\equiv\; \frac{1}{N} \sum_{i=1}^{N} G(Z^i) \nabla_\theta L(\theta, Z^i) \tag{3.19}$$

This can be readily extended to situations where function $G(x, W)$ also depends on decision variables.

Typically, the density functions used in applications depend on the decision variables in a smooth and even analytic way. Therefore, usually there is no problem in taking derivatives inside the expected value in the right hand side of (3.16). When applicable, the likelihood ratio method often also allows estimation of second and higher order derivatives. However, note that the likelihood ratio method is notoriously unstable and a bad choice of the pdf ϕ may result in huge variances of the corresponding estimators. This should not be surprising since the likelihood ratio function may involve divisions by very small numbers, which of course is a very unstable procedure. We refer to Glynn (1990) and Rubinstein and Shapiro (1993) for a further discussion of the likelihood ratio method.

As an example consider the optimal value function of the second stage problem (3.5). Suppose that only the right hand side vector $h = h(\omega)$ of the second stage problem is random. Then $Q(x, h) = G(h - Tx)$, where $G(\chi) \equiv \min\{q^T y : Wy = \chi,\ y \geq 0\}$. Suppose that the random vector h has a pdf $f(\cdot)$. By using the transformation $z = h - Tx$ we obtain

$$\mathbb{E}_f[Q(x, h)] \;=\; \int G(\eta - Tx) f(\eta)\, d\eta \;=\; \int G(z) f(z + Tx)\, dz \;=\; \mathbb{E}_\phi[G(Z) L(x, Z)] \tag{3.20}$$

Here ϕ is a chosen pdf, Z is a random vector having pdf ϕ, and $L(x, z) \equiv f(z + Tx)/\phi(z)$ is the corresponding likelihood ratio function. It can be shown by duality arguments of linear programming that $G(\cdot)$ is a piecewise linear convex function. Therefore, $\nabla_x Q(x, h)$ is piecewise constant and discontinuous, and hence second order derivatives of $\mathbb{E}_f[Q(x, h)]$ cannot be taken inside the expected value. On the other hand, the likelihood ratio function is as smooth as the pdf $f(\cdot)$. Therefore, if $f(\cdot)$ is twice differentiable, then the second order derivatives can be taken inside the expected value in the right hand side of (3.20), and consequently the second order derivatives of $\mathbb{E}_f[Q(x, h)]$ can be consistently estimated by a sample average.

3.5 Simulation Based Optimization Methods

There are basically two approaches to the numerical solution of stochastic optimization problems by using Monte Carlo sampling techniques. One approach is known as the *stochastic approximation* method and originated in Robbins and Monro (1951). The other method was discovered and rediscovered by different researchers and is known under various names.

Suppose that the feasible set $\mathcal{X}$ is convex and that at any point $x \in \mathcal{X}$ an estimate $\hat{\gamma}(x)$ of the gradient $\nabla g(x)$ can be computed, say by a Monte Carlo simulation method. The stochastic approximation method generates the iterates by the recursive equation

$$x_{\nu+1} \;=\; \Pi_{\mathcal{X}}\left(x_\nu - \alpha_\nu \hat{\gamma}(x_\nu)\right) \tag{3.21}$$

where $\alpha_\nu > 0$ are chosen step sizes and $\Pi_{\mathcal{X}}$ denotes the projection onto the set $\mathcal{X}$, i.e., $\Pi_{\mathcal{X}}(x)$ is the point in $\mathcal{X}$ closest to x. Under certain regularity conditions the iterates x_ν converge to a locally optimal solution of the corresponding stochastic optimization problem, i.e., to a local minimizer x^* of $g(x)$ over $\mathcal{X}$. Typically, in order to guarantee this convergence the following two conditions are imposed on the step sizes: (i) $\sum_{\nu=1}^{\infty} \alpha_\nu = \infty$, and (ii) $\sum_{\nu=1}^{\infty} \alpha_\nu^2 < \infty$. For example, one can take $\alpha_\nu \equiv c/\nu$ for some $c > 0$.

If the exact value $\gamma_\nu \equiv \nabla g(x_\nu)$ of the gradient is known, then $-\gamma_\nu$ gives the direction of steepest descent at the point x_ν. This guarantees that if $\gamma_\nu \neq 0$, then moving along the direction $-\gamma_\nu$ the value of the objective function decreases, i.e., $g(x_\nu - \alpha\gamma_\nu) < g(x_\nu)$ for $\alpha > 0$ small enough. The iterative procedure (3.21) tries to mimic that idea by using the estimates $\hat{\gamma}(x_\nu)$ of the corresponding "true" gradients. The projection $\Pi_{\mathcal{X}}$ is needed in order to enforce feasibility of the generated iterates. If the problem is unconstrained, i.e., the feasible set $\mathcal{X}$ coincides with the whole space, then this projection is the identity mapping and can be omitted from (3.21). Note that $\hat{\gamma}(x_\nu)$ does not need to be an accurate estimator of $\nabla g(x_\nu)$.

Kushner and Clark (1978) and Benveniste, Métivier and Priouret (1990) contain expositions of the theory of stochastic approximation. Applications of the stochastic approximation method, combined with the infinitesimal perturbation analysis technique for gradient estimation, to the optimization of the steady state means of single server queues were studied by Chong and Ramadge (1992) and L'Ecuyer and Glynn (1994).

An attractive feature of the stochastic approximation method is its simplicity and ease of implementation in those cases in which the projection $\Pi_{\mathcal{X}}(\cdot)$ can be easily computed. However, it also has severe shortcomings. The crucial question in implementations is the choice of the step sizes α_ν. Small step sizes result in very slow progress towards the optimum while large step sizes make the iterates zigzag. Also, a few wrong steps in the beginning of the procedure may require many iterations to correct. For instance, the algorithm is extremely sensitive to the choice of the constant c in the step size rule $\alpha_\nu = c/\nu$. Therefore, various step size rules were suggested in which the step sizes are chosen adaptively (see Ruppert (1991) for a discussion of that topic).

Another drawback of the stochastic approximation method is that it lacks good stopping criteria and often has difficulties with handling even relatively simple linear constraints.

Another simulation based approach to stochastic optimization is based on the following idea. Let $\hat{g}_N(x)$ be the sample average function defined in (3.12), based on a sample of size N. Consider the optimization problem

$$\min_{x \in \mathcal{X}} \hat{g}_N(x) \tag{3.22}$$

We can view the above problem as the sample average approximation of the "true" (or expected value) problem (3.1). The function $\hat{g}_N(x)$ is random in the sense that it depends on the corresponding sample. However, note that once the sample is generated, $\hat{g}_N(x)$ becomes a deterministic function whose values and derivatives can be computed for a given value of the argument x. Consequently, problem (3.22) becomes a deterministic optimization problem and one can solve it with an appropriate deterministic optimization algorithm.

Let $\hat{v}_N$ and $\hat{x}_N$ denote the optimal objective value and an optimal solution of the sample average problem (3.22), respectively. By the Law of Large Numbers we have that $\hat{g}_N(x)$ converges to $g(x)$ w.p.1 as $N \to \infty$. It is possible to show that under mild additional conditions, $\hat{v}_N$ and $\hat{x}_N$ converge w.p.1 to the

optimal objective value and an optimal solution of the true problem (3.1), respectively. That is, $\hat{v}_N$ and $\hat{x}_N$ are consistent estimators of their "true" counterparts.

This approach to the numerical solution of stochastic optimization problems is a natural outgrowth of the Monte Carlo method of estimation of the expected value of a random function. The method is known by various names and it is difficult to point out who was the first to suggest this approach. In the recent literature a variant of this method, based on the likelihood ratio estimator $\tilde{g}_N(x)$, was suggested in Rubinstein and Shapiro (1990) under the name *stochastic counterpart method* (also see Rubinstein and Shapiro (1993) for a thorough discussion of such a likelihood ratio-sample approximation approach). In Robinson (1996) such an approach is called the *sample path method*. This idea can also be applied to cases in which the set $\mathcal{X}$ is finite, i.e., to stochastic discrete optimization problems (Kleywegt and Shapiro 1999).

Of course, in a practical implementation of such a method one has to choose a specific algorithm for solving the sample average approximation problem (3.22). For example, in the unconstrained case one can use the steepest descent method. That is, iterates are computed by the procedure

$$x_{\nu+1} \;=\; x_\nu - \alpha_\nu \nabla \hat{g}_N(x_\nu) \tag{3.23}$$

where the step size α_ν is obtained by a line search, e.g., $\alpha_\nu \equiv \arg\min_\alpha \hat{g}_N\left(x_\nu - \alpha \nabla \hat{g}_N(x_\nu)\right)$. Note that this procedure is different from the stochastic approximation method (3.21) in two respects. Typically a reasonably large sample size N is used in this procedure, and, more importantly, the step sizes are calculated by a line search instead of being defined a priori. In many interesting cases $\hat{g}_N(x)$ is a piecewise smooth (and even piecewise linear) function and the feasible set is defined by linear constraints. In such cases bundle type optimization algorithms are quite efficient (see Hiriart-Urruty and Lemarechal (1993) for a discussion of the bundle method).

A well developed statistical inference of the estimators $\hat{v}_N$ and $\hat{x}_N$ exists (Rubinstein and Shapiro 1993). That inference aids in the construction of stopping rules, validation analysis and error bounds for obtained solutions, and, furthermore, suggests variance reduction methods that may substantially enhance the rate of convergence of the numerical procedure. For a discussion of this topic and an application to two-stage stochastic programming with recourse we refer to Shapiro and Homem-de-Mello (1998).

If the function $g(x)$ is twice differentiable, then the above sample path method produces estimators that converge to an optimal solution of the true problem at the same asymptotic rate as the stochastic approximation method provided that the stochastic approximation method is applied with the *asymptotically optimal* step sizes (Shapiro 1996). On the other hand, if the underlying probability distribution is discrete and $g(x)$ is piecewise linear and convex, then w.p.1 the sample path method provides an exact optimal solution of the true problem for N large enough, and moreover the probability of that event approaches one exponentially fast as $N \to \infty$ (Shapiro and Homem-de-Mello 1999).

4 Dynamic Programming

Dynamic programming (DP) is an approach for the modeling of dynamic and stochastic decision problems, the analysis of the structural properties of these problems, as well as for the solution of these problems. Dynamic programs are also referred to as Markov decision processes (MDP). Slight distinctions can be made between DP and MDP, such as that in the case of some deterministic problems the term dynamic programming is used rather than Markov decision processes. The term stochastic optimal control is also often used for these types of problems. We shall use these terms synonymously.

Dynamic programs and multistage stochastic programs deal with essentially the same types of problems, namely dynamic and stochastic decision problems. The major distinction between dynamic programming and stochastic programming is in the structures that are used to formulate the models. For example, in DP, the so-called state of the process, as well as the value function, that depends on the state, are two structures that play a central role, but these concepts are usually not used in stochastic programs. Section 4.2 provides an introduction to concepts that are important in dynamic programming.

Much has been written about dynamic programming. Some books in this area are Bellman (1957), Bellman (1961), Bellman and Dreyfus (1962), Nemhauser (1966), Hinderer (1970), Bertsekas and Shreve (1978), Denardo (1982), Ross (1983), Puterman (1994), Bertsekas (1995), and Sennott (1999).

The dynamic programming modeling concepts presented in this article are illustrated with an example, which is both a multiperiod extension of the single period newsvendor example of Sections 2.1 and 2.5, as well as an example of a dynamic pricing problem. The example is called a revenue management problem, and is described in Section 4.1.

4.1 Revenue Management Example

Example 4.1 Managers often have to make decisions repeatedly over time regarding how much inventory to obtain for future sales, as well as how to determine the selling prices. This may involve inventory of one or more products, and the inventory may be located at one or more locations, such as warehouses and retail stores. The inventory may be obtained from a production operation that is part of the same company as the decision maker, and such a production operation may be a manufacturing operation or a service operation, such as an airline, hotel, or car rental company, or the inventory may be purchased from independent suppliers. The decision maker may also have the option to move inventory between locations, such as from warehouses to retail stores. Often the prices of the products can be varied over time to attempt to find the most favorable balance between the supply of the products and the dynamically evolving demand for the products. Such a decision maker can have several objectives, such as to maximize the expected profit over the long run. The profit involves both revenue, which is affected by the pricing decisions, as well as cost, which is affected by the inventory replenishment decisions.

In Section 4.2 examples are given of the formulation of such a revenue management problem with a single product at a single location as a dynamic program.

4.2 Basic Concepts in Dynamic Programming

In this section the basic concepts used in dynamic programming models are introduced.

4.2.1 Decision Times

Decisions can be made at different points in time, and a dynamic programming model should distinguish between the decisions made at different points in time. The major reason why it is important to distinguish between the decisions made at different points in time, is that the information available to the decision maker is different at different points in time—typically more information is available at later points in time (in fact, many people hold this to be the definition of time).

A second reason why distinguishing decision points is useful, is that for many types of DP models it facilitates the computation of solutions. This seems to be the major reason why dynamic programming is used for deterministic decision problems. In this context, the time parameter in the model does not need to correspond to the notion of time in the application. The important feature is that a solution is decomposed into a sequence of distinct decisions. This facilitates computation of the solution if it is easier to compute the individual decisions and then put them together to form a solution, than it is to compute a solution in a more direct way.

The following are examples of ways in which the decision points can be determined in a DP model.

- Decisions can be made at predetermined discrete points in time. In the revenue management example, the decision maker may make a decision once per day regarding what prices to set during the day, as well as how much to order on that day.
- Decisions can be made continuously in time. In the revenue management example, the decision maker may change prices continuously in time (which is likely to require a sophisticated way of communicating the continuously changing prices).
- Decisions can be made at random points in time when specific events take place. In the revenue management example, the decision maker may decide on prices at the random points in time when customer requests are received, and may decide whether to order and how much to order at the random points in time when the inventory changes.

A well-formulated DP model specifies the way in which the decision points in time are determined.

Most of the results presented in this article are for DP models where decisions are made at predetermined discrete points in time, denoted by $t = 0, 1, \ldots, T$, where T denotes the length of the time horizon. DP models with infinite time horizons are also considered. DP models such as these are often called discrete time DP models.

4.2.2 States

A fundamental concept in DP is that of a state, denoted by s. The set $\mathcal{S}$ of all possible states is called the state space. The decision problem is often described as a controlled stochastic process that occupies a state $S(t)$ at each point in time t.

Describing the stochastic process for a given decision problem is an exercise in modeling. The modeler has to determine an appropriate choice of state description for the problem. The basic idea is that the state should be a sufficient, and efficient, summary of the available information that affect the future of the stochastic process. For example, for the revenue management problem, choosing the state to be the amount of the product in inventory may be an appropriate choice. If there is a cost involved in changing the price, then the previous price should also form part of the state. Also, if competitors' prices affect the demand for the product, then additional information about competitors' prices and behavior should be included in the state.

Several considerations should be taken into account when choosing the state description, some of which are described in more detail in later sections. A brief overview is as follows. The state should be a sufficient summary of the available information that affect the future of the stochastic process in the following sense. The state at a point in time should not contain information that is not available to the decision maker at that time, because the decision is based on the state at that point in time. (There are also problems, called partially observed Markov decision processes, in which what is also called the state contains information that is not available to the decision maker. These problems are often handled by converting them to Markov decision processes with observable states. This topic is discussed in Bertsekas (1995).) The set of feasible decisions at a point in time should depend only on the state at that point in time, and maybe on the time itself, and not on any additional information. Also, the costs and transition probabilities at a point in time should depend only on the state at that point in time, the decision made at that point in time, and maybe on the time itself, and not on any additional information. Another consideration is that often one would like to choose the number of states to be as small as possible, since the computational effort of many algorithms increase with the size of the state space. However, the number of states is not the only factor that affect the computational effort. Sometimes it may be more efficient to choose a state description that leads to a larger state space. In this sense the state should be an efficient summary of the available information.

The state space $\mathcal{S}$ can be a finite, countably infinite, or uncountable set. This article addresses mostly dynamic programs with finite or countably infinite, also called discrete, state spaces $\mathcal{S}$.

4.2.3 Decisions

At each decision point in time, the decision maker has to choose a decision, also called an action or control. At any point in time t, the state s at time t, and the time t, should be sufficient to determine the set $\mathcal{A}(s,t)$ of feasible decisions, that is, no additional information is needed to determine the admissible decisions. (Note that the definition of the state of the process should be chosen in such a way that this holds for the decision problem under consideration.) Sometimes the set of feasible decisions depends only on the current state s, in which case the set of feasible decisions is denoted by $\mathcal{A}(s)$. Although most examples have finite sets $\mathcal{A}(s,t)$ or $\mathcal{A}(s)$, these sets may also be countably or uncountably infinite.

In the revenue management example, the decisions involve how much of the product to order, as well as how to set the price. Thus decision $a = (q, r)$ denotes that quantity q is ordered, and that the price is set at r. Suppose the supplier requires that an integer amount between a and b be ordered at a time. Also suppose that the state s denotes the current inventory, and that the inventory may not exceed capacity Q at any time. Then the order quantity may be no more than $Q - s$. Also suppose that the price can be set to be any real number between r_1 and r_2. Then the set of feasible decisions is $\mathcal{A}(s) = \{a, a+1, a+2, \ldots, \min\{Q-s, b\}\} \times [r_1, r_2]$.

The decision maker may randomly select a decision. For example, the decision maker may roll a die and base the decision on the outcome of the die roll. This type of decision is called a randomized decision, as opposed to a nonrandomized, or deterministic, decision. A randomized decision for state s at time t can be represented by a probability distribution on $\mathcal{A}(s,t)$ or $\mathcal{A}(s)$. The decision at time t is denoted by $A(t)$.

4.2.4 Transition Probabilities

The dynamic process changes from state to state over time. The transitions between states may be deterministic or random. The presentation here is for a dynamic program with discrete time parameter $t = 0, 1, \ldots$, and with random transitions.

The transitions have a memoryless, or Markovian, property, in the following sense. Given the history $H(t) \equiv (S(0), A(0), S(1), A(1), \ldots, S(t))$ of the process up to time t, as well as the decision $A(t) \in \mathcal{A}(S(t), t)$ at time t, the probability distribution of the state that the process is in at time $t+1$ depends only on $S(t)$, $A(t)$, and t, that is, the additional information in the history $H(t)$ of the process up to time t provides no additional information for the probability distribution of the state at time $t+1$. (Note that the definition of the state of the process should be chosen in such a way that the probability distribution has this memoryless property.)

Such memoryless random transitions can be represented in several ways. One representation is by transition probabilities. For problems with discrete state spaces, the transition probabilities are denoted by $p[s'|s,a,t] \equiv \mathbb{P}[S(t+1) = s'|H(t), S(t) = s, A(t) = a]$. For problems with uncountable state spaces, the transition probabilities are denoted by $p[B|s,a,t] \equiv \mathbb{P}[S(t+1) \in B|H(t), S(t) = s, A(t) = a]$, where B is a subset of states. Another representation is by a transition function f, such that given $H(t)$, $S(t) = s$, and $A(t) = a$, the state at time $t+1$ is $S(t+1) = f(s,a,t,\omega)$, where ω is a random variable with a known probability distribution. The two representations are equivalent, and in this article we use mostly transition probabilities. When the transition probabilities do not depend on the time t beside depending on the state s and decision a at time t, they are denoted by $p[s'|s,a]$.

In the revenue management example, suppose the demand has probability mass function $\tilde{p}(r,d) \equiv \mathbb{P}[D = d \,|\, \text{price } = r]$ with $d \in \mathbb{Z}_+$. Also suppose that a quantity q that is ordered at time t is received before time $t+1$, and that unsatisfied demand is backordered. Then $\mathcal{S} = \mathbb{Z}$, and the transition probabilities are as follows.

$$p[s'|s,(q,r)] \;=\; \begin{cases} \tilde{p}(r, s+q-s') & \text{if} \quad s' \le s+q \\ 0 & \text{if} \quad s' > s+q \end{cases}$$

If a quantity q that is ordered at time t is received after the demand at time t, and unsatisfied demand is lost, then $\mathcal{S} = \mathbb{Z}_+$, and the transition probabilities are as follows.

$$p[s'|s,(q,r)] \;=\; \begin{cases} \tilde{p}(r, s+q-s') & \text{if} \quad q < s' \le s+q \\ \sum_{d=s}^{\infty} \tilde{p}(r,d) & \text{if} \quad s' = q \\ 0 & \text{if} \quad s' < q \text{ or } s' > s+q \end{cases}$$

4.2.5 Rewards and Costs

Dynamic decision problems often have as objective to maximize the sum of the rewards obtained in each time period, or equivalently, to minimize the sum of the costs incurred in each time period. Other types of objectives sometimes encountered are to maximize or minimize the product of a sequence of numbers resulting from a sequence of decisions, or to maximize or minimize the maximum or minimum of a sequence of resulting numbers.

In this article we focus mainly on the objective of maximizing the expected sum of the rewards obtained in each time period. At any point in time t, the state s at time t, the decision $a \in \mathcal{A}(s,t)$ at time t, and the time t, should be sufficient to determine the expected reward $r(s,a,t)$ at time t. (Again, the definition of the state should be chosen so that this holds for the decision problem under consideration.) When the rewards do not depend on the time t beside depending on the state s and decision a at time t, they are denoted by $r(s,a)$.

Note that, even if in the application the reward $\tilde{r}(s,a,t,s')$ at time t depends on the state s' at time $t+1$, in addition to the state s and decision a at time t, and the time t, the expected reward at time t can still be found as a function of only s, a, and t, because

$$r(s,a,t) \quad = \quad I\!E\left[\tilde{r}(s,a,t,s')\right] \quad = \quad \begin{cases} \sum_{s' \in \mathcal{S}} \tilde{r}(s,a,t,s') p[s'|s,a,t] & \text{if } \mathcal{S} \text{ discrete} \\ \int_{\mathcal{S}} \tilde{r}(s,a,t,s') p[ds'|s,a,t] & \text{if } \mathcal{S} \text{ uncountable} \end{cases}$$

In the revenue management example, suppose unsatisfied demand is backordered, and that an inventory cost/shortage penalty of $h(s)$ is incurred when the inventory level is s at the beginning of the time period. Then $\tilde{r}(s,(q,r'),s') = r'(s+q-s') - h(s)$ with $s' \leq s+q$. Thus

$$r(s,(q,r')) \quad = \quad \sum_{d=0}^{\infty} \tilde{p}(r',d) r' d - h(s)$$

If unsatisfied demand is lost, then $\tilde{r}(s,(q,r'),s') = r'(s+q-s') - h(s)$ with $q \leq s' \leq s+q$. Thus

$$r(s,(q,r')) \quad = \quad \sum_{d=0}^{s-1} \tilde{p}(r',d) r' d + \sum_{d=s}^{\infty} \tilde{p}(r',d) r' s - h(s)$$

In finite horizon problems, there may be a salvage value $v(s)$ if the process terminates in state s at the end of the time horizon T. Such a feature can be incorporated in the previous notation, by letting $\mathcal{A}(s,T) = \{0\}$, and $r(s,0,T) = v(s)$ for all $s \in \mathcal{S}$.

Often the rewards are discounted with a discount factor $\alpha \in [0,1]$, so that the discounted expected value of the reward at time t is $\alpha^t r(s,a,t)$. Such a feature can again be incorporated in the previous notation, by letting $r(s,a,t) = \alpha^t \bar{r}(s,a,t)$ for all s, a, and t, where $\bar{r}$ denotes the undiscounted reward function. When the undiscounted reward does not depend on time, it is convenient to explicitly denote the discounted reward by $\alpha^t r(s,a)$.

4.2.6 Policies

A policy, sometimes called a strategy, prescribes the way a decision is to be made at each point in time, given the information available to the decision maker at the point in time. Therefore, a policy is a solution for a dynamic program.

There are different classes of policies of interest, depending on which of the available information the decisions are based on. A policy can base decisions on all the information in the history of the process up to the time the decision is to be made. Such policies are called history dependent policies. Given the memoryless nature of the transition probabilities, as well as the fact that the sets of feasible decisions and the expected rewards depend on the history of the process only through the current state, it seems intuitive that it should be sufficient to consider policies that base decisions only on the current state and time, and not on any additional information in the history of the process. Such policies are called memoryless, or Markovian, policies. If the transition probabilities, sets of feasible decisions, and rewards do not depend on the current time, then it also seems intuitive that it should be sufficient to consider policies that base decisions only on the current state, and not on any additional information in the history of the process or on the current time. (However, this intuition may be wrong, as shown by counterexamples in Section 4.2.7). Under such policies decisions are made in the same way each time the process is in the same state. Such policies are called stationary policies.

The decision maker may also choose to use some irrelevant information to make a decision. For example, the decision maker may roll a die, or draw a card from a deck of cards, and then base the decision on the outcome of the die roll or the drawn card. In other words, the decision maker may randomly select a decision. Policies that allow such randomized decisions are called randomized policies, and policies that do not allow randomized decisions are called nonrandomized or deterministic policies.

Combining the above types of information that policies can base decisions on, the following types of policies are obtained: the class Π^{HR} of history dependent randomized policies, the class Π^{HD} of history

dependent deterministic policies, the class Π^{MR} of memoryless randomized policies, the class Π^{MD} of memoryless deterministic policies, the class Π^{SR} of stationary randomized policies, and the class Π^{SD} of stationary deterministic policies. The classes of policies are related as follows: $\Pi^{SD} \subset \Pi^{MD} \subset \Pi^{HD} \subset \Pi^{HR}$, $\Pi^{SD} \subset \Pi^{MD} \subset \Pi^{MR} \subset \Pi^{HR}$, $\Pi^{SD} \subset \Pi^{SR} \subset \Pi^{MR} \subset \Pi^{HR}$.

For the revenue management problem, an example of a stationary deterministic policy is to order quantity $q = s_2 - s$ if the inventory level $s < s_1$, for chosen constants $s_1 \leq s_2$, and to set the price at level $r = \check{r}(s)$ for a chosen function $\check{r}(s)$ of the current state s. An example of a stationary randomized policy is to set the price at level $r = \check{r}_1(s)$ with probability $p_1(s)$ and at level $r = \check{r}_2(s)$ with probability $1 - p_1(s)$ for chosen functions $\check{r}_1(s)$, $\check{r}_2(s)$, and $p_1(s)$ of the current state s. An example of a memoryless deterministic policy is to order quantity $q = s_2(t) - s$ if the inventory level $s < s_1(t)$, for chosen functions $s_1(t) \leq s_2(t)$ of the current time t, and to set the price at level $r = \check{r}(s,t)$ for a chosen function $\check{r}(s,t)$ of the current state s and time t.

Policies are functions, defined as follows. Let $\mathcal{H}(t) \equiv \{(S(0), A(0), S(1), A(1), \ldots, S(t))\}$ denote the set of all histories up to time t, and let $\mathcal{H} \equiv \cup_{t=0}^{\infty} \mathcal{H}(t)$ denote the set of all histories. Let $\mathcal{A} \equiv \cup_{s \in \mathcal{S}} \cup_{t=0}^{\infty} \mathcal{A}(s,t)$ denote the set of all feasible decisions. Let $\mathcal{P}(s,t)$ denote the set of probability distributions on $\mathcal{A}(s,t)$ (satisfying regularity conditions), and let $\mathcal{P} \equiv \cup_{s \in \mathcal{S}} \cup_{t=0}^{\infty} \mathcal{P}(s,t)$ denote the set of all such probability distributions. Then Π^{HR} is the set of functions $\pi : \mathcal{H} \mapsto \mathcal{P}$, such that for any t, and any history $H(t)$, $\pi(H(t)) \in \mathcal{P}(S(t), t)$ (again regularity conditions may be required). Π^{HD} is the set of functions $\pi : \mathcal{H} \mapsto \mathcal{A}$, such that for any t, and any history $H(t)$, $\pi(H(t)) \in \mathcal{A}(S(t), t)$. Π^{MR} is the set of functions $\pi : \mathcal{S} \times \mathbb{Z}_+ \mapsto \mathcal{P}$, such that for any state $s \in \mathcal{S}$, and any time $t \in \mathbb{Z}_+$, $\pi(s,t) \in \mathcal{P}(s,t)$. Π^{MD} is the set of functions $\pi : \mathcal{S} \times \mathbb{Z}_+ \mapsto \mathcal{A}$, such that for any state $s \in \mathcal{S}$, and any time $t \in \mathbb{Z}_+$, $\pi(s,t) \in \mathcal{A}(s,t)$. Π^{SR} is the set of functions $\pi : \mathcal{S} \mapsto \mathcal{P}$, such that for any state $s \in \mathcal{S}$, $\pi(s) \in \mathcal{P}(s)$. Π^{SD} is the set of functions $\pi : \mathcal{S} \mapsto \mathcal{A}$, such that for any state $s \in \mathcal{S}$, $\pi(s) \in \mathcal{A}(s)$.

4.2.7 Examples

In this section a number of examples are presented that illustrate why it is sometimes desirable to consider more general classes of policies, such as memoryless and/or randomized policies instead of stationary deterministic policies, even if the sets of feasible solutions, transition probabilities, and rewards are stationary. The examples may also be found in Ross (1970), Ross (1983), Puterman (1994), and Sennott (1999).

The examples are for dynamic programs with stationary input data and objective to minimize the long-run average cost per unit time, $\limsup_{T \to \infty} I\!E \left[\sum_{t=0}^{T-1} r(S(t), A(t)) \mid S(0) \right] / T$. For any policy π, let

$$V^{\pi}(s) \equiv \limsup_{T \to \infty} \frac{1}{T} I\!E^{\pi} \left[\sum_{t=0}^{T-1} r(S(t), A(t)) \middle| S(0) = s \right]$$

denote the long-run average cost per unit time under policy π if the process starts in state s, where $I\!E^{\pi}[\cdot]$ denotes the expected value if policy π is followed.

A policy π^* is called optimal if $V^{\pi^*}(s) = \inf_{\pi \in \Pi^{HR}} V^{\pi}(s)$ for all states s.

Example 4.2 It is clear that if some feasible sets $\mathcal{A}(s)$ are infinite, for example $\mathcal{A}(s) = (0,1)$, then an optimal policy may not exist. The first example shows that an optimal policy may not exist, even if $\mathcal{A}(s)$ is finite for all states s.

The state space $\mathcal{S} = \{1, 1', 2, 2', 3, 3', \ldots\}$. Feasible decision sets are $\mathcal{A}(i) = \{a, b\}$, and $\mathcal{A}(i') = \{a\}$, for each $i \in \{1, 2, 3, \ldots\}$. Transitions are deterministic; from a state i we either go up to state $i+1$ if we make decision a, or we go across to state i' if we make decision b. Once in a state i', we remain in state i'. That is, the transition function is $f(i,a) = i+1$, $f(i,b) = i'$, and $f(i',a) = i'$. In a state i a cost of 1 is incurred, and in a state i' a cost of $1/i$ is incurred. That is, the costs are $r(i,a) = r(i,b) = 1$, and $r(i',a) = 1/i$.

Suppose the process starts in state 1. The idea is simple: we would like to go to a high state i, before moving over to state i'. However, a policy π that chooses decision a for each state i, has long-run average cost per unit time of $V^{\pi}(1) = 1$, which is as bad as can be. The only other possibility is that there exists a state j such that policy π chooses decision b with positive probability p_j when the process reaches state j. In that case $V^{\pi}(1) \geq p_j / j > 0$. Thus $V^{\pi}(1) > 0$ for all policies π.

The stationary deterministic policy π_j that chooses decision a for states $i = 1, 2, \ldots, j-1$, and chooses decision b for state j, has long-run average cost per unit time of $V^{\pi_j}(1) = 1/j$. By choosing j arbitrarily large, $V^{\pi_j}(1)$ can be made arbitrarily close to zero, but no policy π has long-run average cost per unit time $V^{\pi}(1)$ less than or equal to zero. Thus an optimal policy π^*, with $V^{\pi^*}(1) = 0$, does not exist.

However, for any policy π, there exists a stationary deterministic policy π_j such that $V^{\pi_j}(1) < V^{\pi}(1)$.

Example 4.3 The second example shows that it is not always the case that for any policy π, there exists a stationary deterministic policy π' that is at least as good as π.

The state space $\mathcal{S} = \{1, 2, 3, \ldots\}$. Feasible decision sets are $\mathcal{A}(i) = \{a, b\}$ for each $i \in \mathcal{S}$. Transitions are deterministic; from a state i we either remain in state i if we make decision a, or we go up to state $i+1$ if we make decision b. That is, the transition function is $f(i, a) = i$, and $f(i, b) = i + 1$. When decision a is made in a state i, a cost of $1/i$ is incurred, and when decision b is made, a cost of 1 is incurred. That is, the costs are $r(i, a) = 1/i$, and $r(i, b) = 1$.

Suppose the process starts in state 1. Again, the idea is simple: we would like to go to a high state i, and then make decision a. However, a stationary deterministic policy π that chooses decision b for each state i, has long-run average cost per unit time of $V^{\pi}(1) = 1$, which is as bad as can be. The only other possibility for a stationary deterministic policy π is to choose decision a for the first time in state j. In that case $V^{\pi}(1) = 1/j > 0$. Thus $V^{\pi}(1) > 0$ for all stationary deterministic policies π.

Consider the memoryless deterministic policy π^* that chooses decision a the first i times that the process is in state i, and then chooses decision b. Thus the sequence of states under policy π^* is $1, 1, 2, 2, 2, 3, 3, 3, 3, \ldots$. The sequence of decisions under policy π^* is $a, b, a, a, b, a, a, a, b, \ldots$. The sequence of costs under policy π^* is $1, 1, 1/2, 1/2, 1, 1/3, 1/3, 1/3, 1, \ldots$. Note that the total cost incurred while the process is in state i is 2 for each i, so that the total cost incurred from the start of the process until the process leaves state i is $2i$. The total time until the process leaves state i is $2 + 3 + \cdots + i + 1 = i(i+3)/2$. Thus the average cost per unit time if the process is currently in state $i+1$ is less than $2(i+1)/(i(i+3)/2)$, which becomes arbitrarily small as i becomes large. Thus $V^{\pi^*}(1) = 0$, and the memoryless deterministic policy π^* is better than any stationary deterministic policy.

However, there exists a stationary randomized policy $\tilde{\pi}$ with the same expected long-run average cost per unit time as policy π^*. When in state i, $\tilde{\pi}$ chooses decision a with probability $i/(i+1)$ and decision b with probability $1/(i+1)$. The expected amount of time that the process under $\tilde{\pi}$ spends in state i is $i+1$, and the expected cost incurred while the process under $\tilde{\pi}$ is in state i is 2. Thus the cost incurred under $\tilde{\pi}$ is similar to the cost incurred under π^*, and it can be shown that $V^{\tilde{\pi}}(1) = 0$.

Example 4.4 The third example shows that it is not always the case that for any policy π, there exists a stationary randomized policy $\tilde{\pi}$ that is at least as good as π.

The state space $\mathcal{S} = \{0, 1, 1', 2, 2', 3, 3', \ldots\}$. Feasible decision sets are $\mathcal{A}(0) = \{a\}$, $\mathcal{A}(i) = \{a, b\}$, and $\mathcal{A}(i') = \{a\}$ for each $i \in \{1, 2, 3, \ldots\}$. When in state 0, a cost of 1 is incurred, otherwise there is no cost. That is, the costs are $r(0, a) = 1$, and $r(i, a) = r(i, b) = r(i', a) = 0$. In this example transitions are random, with transition probabilities

$$\begin{aligned}
p[i|0,a] &= p[i'|0,a] &&= \frac{3}{2}\left(\frac{1}{4}\right)^i \\
p[0|i,a] &= p[i+1|i,a] &&= \frac{1}{2} \\
p[0|i,b] &= 1 - p[i'|i,b] &&= \left(\frac{1}{2}\right)^i \\
p[0|i',a] &= 1 - p[i'|i',a] &&= \left(\frac{1}{2}\right)^i
\end{aligned}$$

Again, the idea is simple: we would like to visit state 0 as infrequently as possible. Thus we would like the process to move to a high state i or i', where the probability of making a transition to state 0 can be made small. However, to move to a high state requires decision a to be made, which involves a high risk of moving to state 0. The policy that always makes decision a is as bad as possible.

Let M_{i0}^{π} denote the mean time for the process to move from state i to state 0 under stationary policy π. Thus $V^{\pi}(0) = 1/M_{00}^{\pi}$.

First consider the stationary deterministic policy π_j that chooses decision a for states $i = 1, 2, \ldots, j-1$, and chooses decision b for states $i = j, j+1, \ldots$. Then for $i = 1, 2, \ldots, j-1$, $M_{i0}^{\pi_j} = 2 + 2^i - (1/2)^{j-i-1}$. From this it follows that $M_{00}^{\pi_j} < 5$ and thus $V^{\pi_j}(0) > 1/5$ for all j.

Next consider any stationary randomized policy π, and let $\pi(i,a)$ denote the probability that decision a is made in state i. Then, given that the process is in state i, the probability is $\pi(i,a)\pi(i+1,a)\cdots\pi(j-1,a)\pi(j,b)$ that the process under policy π behaves the same until state 0 is reached as under policy π_j. Thus

$$\begin{aligned} M_{i0}^{\pi} &= \sum_{j=i}^{\infty} \left[\pi(j,b) \prod_{k=i}^{j-1} \pi(k,a) \right] M_{i0}^{\pi_j} + 2 \prod_{k=i}^{\infty} \pi(k,a) \\ &< \left(2 + 2^i\right) \left[\sum_{j=i}^{\infty} \pi(j,b) \prod_{k=i}^{j-1} \pi(k,a) + 2 \prod_{k=i}^{\infty} \pi(k,a) \right] \\ &= 2 + 2^i \end{aligned}$$

From this it follows that $M_{00}^{\pi} < 5$ and thus $V^{\pi}(0) > 1/5$ for any stationary randomized policy π.

Consider the memoryless deterministic policy π^* that uses the decisions of π_1 for $t = 1, 2, \ldots, T_1$, π_2 for $t = T_1 + 1, T_1 + 2, \ldots, T_2, \ldots$, π_j for $t = T_{j-1} + 1, T_{j-1} + 2, \ldots, T_j, \ldots$. For appropriate choice of T_j's it follows that $V^{\pi^*}(0) = 1/5$, and thus the memoryless deterministic policy π^* is better than any stationary randomized policy.

Example 4.5 In all the examples presented so far, it is the case that for any policy $\tilde{\pi}$, and any $\varepsilon > 0$, there exists a stationary deterministic policy π that has value function V^{π} within ε of the value function $V^{\tilde{\pi}}$. The fourth example shows that this does not always hold.

The state space $\mathcal{S} = \{0, 1, 1', 2, 2', 3, 3', \ldots\}$. Feasible decision sets are $\mathcal{A}(0) = \{a\}$, $\mathcal{A}(i) = \{a, b\}$, and $\mathcal{A}(i') = \{a\}$ for each $i \in \{1, 2, 3, \ldots\}$. When in state $i \in \{0, 1, 2, \ldots\}$, a cost of 2 is incurred, otherwise there is no cost. That is, the costs are $r(0,a) = 2$, $r(i,a) = r(i,b) = 2$, and $r(i',a) = 0$. The transition probabilities are as follows.

$$\begin{aligned} p[0|0,a] &= 1 \\ p[i+1|i,a] &= 1 \\ p[i'|i,b] &= 1 - p[0|i,b] = p_i \\ p[(i-1)'|i',a] &= 1 \qquad \text{for all } i \geq 2 \\ p[1|1',a] &= 1 \end{aligned}$$

The values p_i can be chosen to satisfy

$$\begin{aligned} p_i &< 1 \qquad \text{for all } i \\ \prod_{i=1}^{\infty} p_i &= \frac{3}{4} \end{aligned}$$

Suppose the process starts in state 1. Again, the idea is simple: we would like to go down the chain $i', (i-1)', \ldots, 1'$ as much as possible. To do that, we also need to go up the chain $1, 2, \ldots, i$, and then go from state i to state i' by making decision b. When we make decision b in state i, there is a risk $1 - p_i > 0$ of making a transition to state 0, which is very bad.

A stationary deterministic policy π that chooses decision a for each state i, has long-run average cost per unit time of $V^{\pi}(1) = 2$, which is as bad as can be. The only other possibility for a stationary deterministic policy π is to choose decision b for the first time in state j. In that case, each time state j is visited, there is a positive probability $1 - p_j > 0$ of making a transition to state 0. It follows that the mean time until a transition to state 0 is made is less than $2j/(1 - p_j) < \infty$, and the long-run average cost per unit time is $V^{\pi}(1) = 2$. Thus $V^{\pi}(1) = 2$ for all stationary deterministic policies π.

Consider the memoryless deterministic policy $\tilde{\pi}$ that on its jth visit to state 1, chooses decision a, $j-1$ times and then chooses decision b. With probability $\prod_{i=1}^{\infty} p_i = 3/4$ the process never makes a transition to state 0, and the long-run average cost per unit time is 1. Otherwise, with probability $1 - \prod_{i=1}^{\infty} p_i = 1/4$, the process makes a transition to state 0, and the long-run average cost per unit time is 2. Hence, the expected long-run average cost per unit time is $V^{\tilde{\pi}}(1) = 3/4 \times 1 + 1/4 \times 2 = 5/4$. Thus, there is no ε-optimal stationary deterministic policy for $\varepsilon \in (0, 3/4)$. In fact, by considering memoryless deterministic policies $\tilde{\pi}_k$ that on their jth visit to state 1, choose decision a, $j+k$ times and then choose decision b, one obtains policies with expected long-run average cost per unit time $V^{\tilde{\pi}_k}(1)$ arbitrarily close to 1 for sufficiently large values of k. It is clear that $V^{\pi}(1) \geq 1$ for all policies π, and thus $V^*(1) = 1$, and there is no ε-optimal stationary deterministic policy for $\varepsilon \in (0, 1)$.

4.3 Finite Horizon Dynamic Programs

In this section we investigate dynamic programming models for optimization problems with the form

$$\max_{(A(0), A(1), \ldots, A(T))} I\!E \left[\sum_{t=0}^{T} r(S(t), A(t), t) \right] \tag{4.1}$$

where $T < \infty$ is the known finite horizon length, and decisions $A(t), t = 0, 1, \ldots, T$, have to be feasible and may depend only on the information available to the decision maker at each time t, that is the history $H(t)$ of the process up to time t, and possibly some randomization. For the presentation we assume that $\mathcal{S}$ is countable and r is bounded. Similar results hold in more general cases, subject to regularity conditions.

4.3.1 Optimality Results

For any policy $\pi \in \Pi^{HR}$, and any history $h(t) \in \mathcal{H}(t)$, let

$$U^{\pi}(h(t)) \equiv I\!E^{\pi} \left[\sum_{\tau=t}^{T} r(S(\tau), A(\tau), \tau) \middle| H(t) = h(t) \right] \tag{4.2}$$

denote the expected value under policy π from time t onwards, given the history $h(t)$ of the process up to time t; U^{π} is called the value function under policy π. The optimal value function U^* is given by

$$U^*(h(t)) \equiv \sup_{\pi \in \Pi^{HR}} U^{\pi}(h(t)) \tag{4.3}$$

It follows from r being bounded that U^{π} and U^* are bounded. A policy $\pi^* \in \Pi^{HR}$ is called optimal if $U^{\pi^*}(h(t)) = U^*(h(t))$ for all $h(t) \in \mathcal{H}(t)$ and all $t \in \{0, 1, \ldots, T\}$. Also, a policy $\pi^*_{\varepsilon} \in \Pi^{HR}$ is called ε-optimal if $U^{\pi^*_{\varepsilon}}(h(t)) + \varepsilon > U^*(h(t))$ for all $h(t) \in \mathcal{H}(t)$ and all $t \in \{0, 1, \ldots, T\}$.

It is easy to see that the value function U^{π} satisfies the following inductive equation for any $\pi \in \Pi^{HR}$ and any history $h(t) = (h(t-1), a(t-1), s)$.

$$U^{\pi}(h(t)) = I\!E^{\pi} \left[r(s, \pi(h(t)), t) + U^{\pi}(H(t+1)) \mid H(t) = h(t) \right] \tag{4.4}$$

Using (4.4), U^{π} can be computed inductively; this is called the finite horizon policy evaluation algorithm. This result is also used to establish the result that U^* satisfy the following optimality equation for all histories $h(t) = (h(t-1), a(t-1), s)$.

$$U^*(h(t)) = \sup_{a \in \mathcal{A}(s,t)} \left\{ r(s, a, t) + I\!E \left[U^*(H(t+1)) \mid H(t) = h(t), A(t) = a \right] \right\} \tag{4.5}$$

From the memoryless properties of the feasible sets, transition probabilities, and rewards, it is intuitive that $U^*(h(t))$ should depend on $h(t) = (h(t-1), a(t-1), s)$ only through the state s at time t and the time

t, and that it should be sufficient to consider memoryless policies. To establish these results, inductively define the memoryless function V^* along the lines of the optimality equation (4.5) for U^*.

$$\begin{aligned} V^*(s,T+1) &\equiv 0 \\ V^*(s,t) &\equiv \sup_{a\in\mathcal{A}(s,t)} \left\{ r(s,a,t) + I\!E\left[V^*(S(t+1),t+1) \mid S(t)=s, A(t)=a\right] \right\} \\ & \qquad\qquad t = T, T-1, \ldots, 1, 0 \end{aligned} \tag{4.6}$$

Then it is easy to show, again by induction, that for any history $h(t) = (h(t-1), a(t-1), s)$, $U^*(h(t)) = V^*(s,t)$.

For any memoryless policy $\pi \in \Pi^{MR}$, inductively define the function

$$\begin{aligned} V^\pi(s,T+1) &\equiv 0 \\ V^\pi(s,t) &\equiv I\!E^\pi\left[r(s,\pi(s,t),t) + V^\pi(S(t+1),t+1) \mid S(t)=s\right] \\ & \qquad\qquad t = T, T-1, \ldots, 1, 0 \end{aligned} \tag{4.7}$$

Then, for any history $h(t) = (h(t-1), a(t-1), s)$, $U^\pi(h(t)) = V^\pi(s,t)$, that is, V^π is the (simpler) value function of policy $\pi \in \Pi^{MR}$.

In a similar way it can be shown that it is sufficient to consider only memoryless deterministic policies, in the following sense. First suppose that for each $s \in \mathcal{S}$ and each $t \in \{0, 1, \ldots, T\}$, there exists a decision $a^*(s,t)$ such that

$$\begin{aligned} & r(s,a^*(s,t),t) + I\!E\left[V^*(S(t+1),t+1) \mid S(t)=s, A(t)=a^*(s,t)\right] \\ = & \sup_{a\in\mathcal{A}(s,t)} \left\{ r(s,a,t) + I\!E\left[V^*(S(t+1),t+1) \mid S(t)=s, A(t)=a\right] \right\} \end{aligned} \tag{4.8}$$

Then the memoryless deterministic policy π^* with $\pi^*(s,t) = a^*(s,t)$ is optimal, that is, for any history $h(t) = (h(t-1), a(t-1), s)$, $U^{\pi^*}(h(t)) = V^{\pi^*}(s,t) = V^*(s,t) = U^*(h(t))$. If, for some s and t, there does not exist such an optimal decision $a^*(s,t)$, then there also does not exist an optimal history dependent randomized policy. In such a case it still holds that for any $\varepsilon > 0$, there exists an ε-optimal memoryless deterministic policy π^*_ε, obtained by choosing decisions $\pi^*_\varepsilon(s,t)$ such that

$$\begin{aligned} & r(s,\pi^*_\varepsilon(s,t),t) + I\!E\left[V^*(S(t+1),t+1) \mid S(t)=s, A(t)=\pi^*_\varepsilon(s,t)\right] + \frac{\varepsilon}{T+1} \\ > & \sup_{a\in\mathcal{A}(s,t)} \left\{ r(s,a,t) + I\!E\left[V^*(S(t+1),t+1) \mid S(t)=s, A(t)=a\right] \right\} \end{aligned} \tag{4.9}$$

Solving a finite horizon dynamic program usually involves computing V^* with a backward induction algorithm using (4.6). An optimal policy $\pi^* \in \Pi^{MD}$ is then obtained using (4.8), or an ε-optimal policy $\pi^*_\varepsilon \in \Pi^{MD}$ is obtained using (4.9).

Finite Horizon Backward Induction Algorithm

0. Set $V^*(s,T+1) = 0$ for all $s \in \mathcal{S}$.
1. For $t = T, \ldots, 1$, repeat steps 2 and 3.
2. For each $s \in \mathcal{S}$, compute

$$V^*(s,t) = \sup_{a\in\mathcal{A}(s,t)} \left\{ r(s,a,t) + I\!E\left[V^*(S(t+1),t+1) \mid S(t)=s, A(t)=a\right] \right\} \tag{4.10}$$

3. For each $s \in \mathcal{S}$, choose a decision

$$\pi^*(s,t) \in \arg\max_{a\in\mathcal{A}(s,t)} \left\{ r(s,a,t) + I\!E\left[V^*(S(t+1),t+1) \mid S(t)=s, A(t)=a\right] \right\}$$

if the maximum on the right hand side is attained. Otherwise, for any chosen $\varepsilon > 0$, choose a decision $\pi^*_\varepsilon(s,t)$ such that

$$\begin{aligned} & r(s,\pi^*_\varepsilon(s,t),t) + I\!E\left[V^*(S(t+1),t+1) \mid S(t)=s, A(t)=\pi^*_\varepsilon(s,t)\right] + \frac{\varepsilon}{T+1} \\ > & \sup_{a\in\mathcal{A}(s,t)} \left\{ r(s,a,t) + I\!E\left[V^*(S(t+1),t+1) \mid S(t)=s, A(t)=a\right] \right\} \end{aligned}$$

4.3.2 Structural Properties

Dynamic programming is useful not only for the computation of optimal policies and optimal expected values, but also for determining insightful structural characteristics of optimal policies. In fact, for many interesting applications the state space is too big to compute optimal policies and optimal expected values exactly, but dynamic programming can still be used to establish qualitative characteristics of optimal quantities. Some such structural properties are illustrated with examples.

Example 4.6 The Secretary Problem. Suppose a decision maker has to choose one out of N candidates. The decision maker observes the candidates one at a time, and after a candidate has been observed, the decision maker either has to choose that candidate, and the process terminates, or the decision maker has to reject the candidate and observe the next candidate. Rejected candidates cannot be recalled. The number N of candidates is known, but the decision maker knows nothing else about the candidates beforehand. The decision maker can rank any candidates that have been observed. That is, for any two candidates i and j that have been observed, either i is preferred to j, denoted by $j \prec i$, or j is preferred to i, denoted by $i \prec j$. The preferences are transitive, that is, if $i \prec j$ and $j \prec k$, then $i \prec k$. The candidates are observed in random sequence, that is, the $N!$ permutations of candidates are equally likely. The decision maker wants to maximize the probability of selecting the best candidate. This problem can be formulated as a dynamic program. The discrete time parameter corresponds to the number of candidates that have been observed so far, and the current state is an indicator whether the current candidate is the best candidate observed so far or not. If the current candidate is selected, then the expected reward is the probability that the current candidate is the best candidate overall. If the current candidate is rejected, then the current reward is zero, and the process makes a transition to the next stage. Dynamic programming can be used to show that the following policy is optimal. Let

$$\tau(N) \quad \equiv \quad \max\left\{ n \in \{1,\ldots,N\} \ : \ \frac{1}{n} + \frac{1}{n+1} + \cdots + \frac{1}{N-1} > 1 \right\}$$

The optimal policy is then to observe the first $\tau(N)$ candidates without selecting any candidate, and then to select the first candidate thereafter that is preferred to all the previously observed candidates. It can be shown that $\tau(N)/N$ converges to $1/e$ quite rapidly. Thus for a reasonably large number N of candidates (say $N > 15$), a good policy is to observe the first N/e candidates without selecting any candidate, and then to select the first candidate thereafter that is preferred to all the previous candidates. It is also interesting that the optimal probability of selecting the best candidate decreases in N, but it never decreases below $1/e \approx 37\%$, no matter how large the number of candidates.

Example 4.7 Inventory Replenishment. A business purchases and sells a particular product. A decision maker has to decide regularly, say once every day, how much of the product to buy. The business does not have to wait to receive the purchased product. Unlike the newsvendor problem, here product that is not sold on a particular day can be kept in inventory for the future. The business pays a fixed cost K plus a variable cost c per unit of product each time product is purchased. Thus, if a units of product is purchased, then the purchasing cost is $K + ca$ if $a > 0$, and it is 0 if $a = 0$. In addition, if the inventory level at the beginning of the day is s, and a units of product is purchased, then an inventory cost of $h(s+a)$ is incurred, where h is a convex function. The demand for the product on different days are independent

and identically distributed. If the demand D is greater than the available inventory $s + a$, then the excess demand is backlogged until additional inventory is obtained, at which time the backlogged demand is filled immediately. Inventory remaining at the end of the time horizon has no value. The objective is to minimize the expected total cost over the time horizon. This problem can be formulated as a discrete time dynamic program. The state $S(t)$ is the inventory at the beginning of day t. The decision $A(t)$ is the quantity purchased on day t, and the single stage cost $r(s, a) = (K + ca)I_{\{a>0\}} + h(s + a)$. The transitions are given by $S(t + 1) = S(t) + A(t) - D(t)$. Dynamic programming can be used to show that the following policy is optimal. If the inventory level $S(t) < \sigma^*(t)$, where $\sigma^*(t)$ is called the optimal reorder point at time t, then it is optimal to purchase $\Sigma^*(t) - S(t)$ units of product at time t, where $\Sigma^*(t)$ is called the optimal order-up-to point at time t. If the inventory level $S(t) \geq \sigma^*(t)$, then it is optimal not to purchase any product. Such a policy is often called an (s, S)-policy, or a (σ, Σ)-policy. Similar results hold in the infinite horizon case, except that σ^* and Σ^* do not depend on time t anymore.

Example 4.8 Resource Allocation. A decision maker has an amount of resource that can be allocated over some time horizon. At each discrete point in time, a request for some amount of resource is received. If the request is for more resource than the decision maker has available, then the request has to be rejected. Otherwise, the request can be accepted or rejected. A request must be accepted or rejected as a whole—the decision maker cannot allocate a fraction of the amount of resource requested. Rejected requests cannot be recalled later. If the request is accepted, the amount of resource available to the decision maker is reduced by the amount of resource requested, and the decision maker receives an associated reward in return. The amounts of resource and the rewards of future requests are unknown to the decision maker, but the decision maker knows the probability distribution of these. At the end of the time horizon, the decision maker receives a salvage reward for the remaining amount of resource. The objective is to maximize the expected total reward over the time horizon. Problems of this type are encountered in revenue management and the selling of assets such as real estate and vehicles. This resource allocation problem can be formulated as a dynamic program. The state $S(t)$ is the amount of resource available to the decision maker at the beginning of time period t. The decision $A(t)$ is the rule that will be used for accepting or rejecting requests during time period t. If a request for amount Q of resource with an associated reward R is accepted in time period t, then the single stage reward is R and the next state is $S(t + 1) = S(t) - Q$. If the request is rejected, then the next state is $S(t + 1) = S(t)$. It is easy to see that the optimal value function $V^*(s, t)$ is increasing in s and decreasing in t. The following threshold policy, with reward threshold function $x^*(q, s, t) = V^*(s, t + 1) - V^*(s - q, t + 1)$, is optimal. Accept a request for amount Q of resource with an associated reward R if $Q \leq S(t)$ and $R \geq x^*(Q, S(t), t)$, and reject the request otherwise. If each request is for the same amount of resource (say 1 unit of resource), and the salvage reward is concave in the remaining amount of resource, then the optimal value function $V^*(s, t)$ is concave in s and t, and the optimal reward threshold $x^*(1, s, t) = V^*(s, t + 1) - V^*(s - 1, t + 1)$ is decreasing in s and t. These intuitive properties do not hold in general if the requests are for random amounts of resource.

Structural properties of the optimal value functions and optimal policies of dynamic programs have been investigated for many different applications. Some general structural results are given in Serfozo (1976), Topkis (1978), and Heyman and Sobel (1984).

4.4 Infinite Horizon Dynamic Programs

In this section we present dynamic programming models with an infinite time horizon. Although an infinite time horizon is a figment of the imagination, these models often are useful for decision problems with many decision points. Many infinite horizon models also have the desirable feature that there exist stationary deterministic optimal policies. Thus optimal decisions depend only on the current state of the process, and not on the sometimes artificial notion of time, as in finite horizon problems. This characteristic makes optimal policies easier to understand, compute, and implement, which is desirable in applications.

We again assume that $\mathcal{S}$ is countable and r is bounded. Similar results hold in more general cases, subject to regularity conditions. We also assume that the sets $\mathcal{A}(s)$ of feasible decisions depend only on the states s, the transition probabilities $p[s'|s, a]$ depend only on the states s, s', and decisions a, and the rewards $r(s, a)$ depend only on the states s and decisions a, and not on time, as in the finite horizon case.

In this article we focus on dynamic programs with total discounted reward objectives. As illustrated in the examples of Section 4.2.7, infinite horizon dynamic programs with other types of objectives, such as long-run average reward objectives, may exhibit undesirable behavior. A proper treatment of dynamic programs with these types of objectives requires more space than we have available here, and therefore we refer the interested reader to the references. Besides, in most practical applications, rewards and costs in the near future are valued more than rewards and costs in the more distant future, and hence total discounted reward objectives are preferred for applications.

4.5 Infinite Horizon Discounted Dynamic Programs

In this section we investigate dynamic programming models for optimization problems with the form

$$\max_{(A(0),A(1),\dots)} I\!E \left[\sum_{t=0}^{\infty} \alpha^t r(S(t), A(t)) \right] \tag{4.11}$$

where $\alpha \in (0,1)$ is a known discount factor. Again, decisions $A(t), t = 0, 1, \dots$, have to be feasible and may depend only on the information available to the decision maker at each time t, that is the history $H(t)$ of the process up to time t, and possibly some randomization.

4.5.1 Optimality Results

The establishment of optimality results for infinite horizon discounted dynamic programs is quite similar to that for finite horizon dynamic programs. An important difference though is that backward induction cannot be used in the infinite horizon case.

We again start by defining the value function U^π for a policy $\pi \in \Pi^{HR}$,

$$U^\pi(h(t)) \quad \equiv \quad I\!E^\pi \left[\sum_{\tau=t}^{\infty} \alpha^{\tau-t} r(S(\tau), A(\tau)) \middle| H(t) = h(t) \right] \tag{4.12}$$

The optimal value function U^* is then defined as in (4.3). It follows from r being bounded and $\alpha \in (0,1)$ that U^π and U^* are bounded. Again, a policy $\pi^* \in \Pi^{HR}$ is called optimal if $U^{\pi^*}(h(t)) = U^*(h(t))$ for all $h(t) \in \mathcal{H}(t)$ and all $t \in \{0, 1, \dots\}$, and a policy $\pi^*_\varepsilon \in \Pi^{HR}$ is called ε-optimal if $U^{\pi^*_\varepsilon}(h(t)) + \varepsilon > U^*(h(t))$ for all $h(t) \in \mathcal{H}(t)$ and all $t \in \{0, 1, \dots\}$.

The value function U^π satisfies an inductive equation similar to (4.4) for the finite horizon case, for any $\pi \in \Pi^{HR}$ and any history $h(t) = (h(t-1), a(t-1), s)$.

$$U^\pi(h(t)) \quad = \quad I\!E^\pi \left[r(s, \pi(h(t))) + \alpha U^\pi(H(t+1)) \mid H(t) = h(t) \right] \tag{4.13}$$

However, unlike the finite horizon case, U^π cannot in general be computed inductively using (4.13). We also do not proceed in the infinite horizon case by establishing an optimality equation similar to (4.5). However, we do proceed by considering an optimality equation similar to (4.6).

From the stationary properties of the feasible sets, transition probabilities, and rewards, it is intuitive that $U^*(h(t))$ should depend on $h(t) = (h(t-1), a(t-1), s)$ only through the most recent state s, and that it should be sufficient to consider stationary policies. However, it is convenient to show, as an intermediate step, that it is sufficient to consider memoryless policies. For any $\pi \in \Pi^{HR}$ and any history $h(t)$, define the memoryless randomized policy $\tilde{\pi} \in \Pi^{MR}$ as follows.

$$\tilde{\pi}(s, t+\tau)(A) \quad \equiv \quad I\!P^\pi[A(t+\tau) \in A | S(t+\tau) = s, H(t) = h(t)]$$

for any $s \in \mathcal{S}$, any $\tau \in \{0, 1, 2, \dots\}$, and any $A \subseteq \mathcal{A}(s)$. (Recall that $\tilde{\pi}(s,t)(A)$ denotes the probability, given state s at time t, that a decision in $A \subseteq \mathcal{A}(s)$ is chosen under policy $\tilde{\pi}$.) Then it is easy to show that for any $s \in \mathcal{S}$, any $\tau \in \{0, 1, 2, \dots\}$, and any $A \subseteq \mathcal{A}(s)$, $I\!P^{\tilde{\pi}}[S(t+\tau) = s, A(t+\tau) \in A | H(t) = h(t)] = I\!P^\pi[S(t+\tau) = s, A(t+\tau) \in A | H(t) = h(t)]$. Thus, for any $\pi \in \Pi^{HR}$ and any history $h(t)$, there exists a memoryless randomized policy $\tilde{\pi}$ that behaves exactly like π from time t onwards, and hence

$U^{\tilde{\pi}}(h(t+\tau)) = U^{\pi}(h(t+\tau))$ for any history $h(t+\tau)$ that starts with $h(t)$. It follows that $U^*(h(t)) \equiv \sup_{\pi \in \Pi^{HR}} U^{\pi}(h(t)) = \sup_{\pi \in \Pi^{MR}} U^{\pi}(h(t))$, that is, it is sufficient to consider memoryless randomized policies.

For any memoryless randomized policy π and any history $h(t) = (h(t-1), a(t-1), s)$, $U^{\pi}(h(t))$ depends on $h(t)$ only through the most recent state s and the time t. Instead of exploring this result in more detail as for the finite horizon case, we use another property of memoryless randomized policies. Using the stationary properties of the problem parameters, it follows that, for any memoryless randomized policy π and any time t, π behaves in the same way from time t onwards as another memoryless randomized policy $\tilde{\pi}$ behaves from time 0 onwards, where $\tilde{\pi}$ is obtained from π by shifting π backwards through t, as follows. Define the shift function $\theta : \Pi^{MR} \mapsto \Pi^{MR}$ by $\theta(\pi)(s,t) \equiv \pi(s,t+1)$ for all $s \in \mathcal{S}$ and all $t \in \{0, 1, \dots\}$. That is, policy $\theta(\pi) \in \Pi^{MR}$ makes the same decisions at time t as policy $\pi \in \Pi^{MR}$ makes at time $t+1$. Also, inductively define the convolution $\theta^{t+1}(\pi) \equiv \theta(\theta^t(\pi))$. Thus the shifted policy $\tilde{\pi}$ described above is given by $\tilde{\pi} = \theta^t(\pi)$. Also note that for a stationary policy π, $\theta(\pi) = \pi$.

Now it is useful to focus on the value function V^{π} for a policy $\pi \in \Pi^{HR}$ from time 0 onwards,

$$V^{\pi}(s) \equiv I\!E^{\pi}\left[\sum_{t=0}^{\infty} \alpha^t r(S(t), A(t)) \middle| S(0) = s\right] \tag{4.14}$$

That is, $V^{\pi}(s) = U^{\pi}(h(0))$, where $h(0) = (s)$. Then it follows that, for any memoryless randomized policy π and any history $h(t) = (h(t-1), a(t-1), s)$,

$$U^{\pi}(h(t)) = V^{\theta^t(\pi)}(s) \tag{4.15}$$

Thus we obtain the further simplification that $U^*(h(t)) \equiv \sup_{\pi \in \Pi^{HR}} U^{\pi}(h(t)) = \sup_{\pi \in \Pi^{MR}} U^{\pi}(h(t)) = \sup_{\pi \in \Pi^{MR}} V^{\pi}(s)$. Define the optimal value function V^* by

$$V^*(s) \equiv \sup_{\pi \in \Pi^{MR}} V^{\pi}(s) \tag{4.16}$$

Thus, for any history $h(t) = (h(t-1), a(t-1), s)$,

$$U^*(h(t)) = V^*(s) \tag{4.17}$$

and hence $U^*(h(t))$ depends only on the most recent state s, as expected.

It also follows from (4.13) and (4.15) that for any $\pi \in \Pi^{MR}$,

$$\begin{aligned} V^{\pi}(s) &= U^{\pi}(h(0)) = I\!E^{\pi}\left[r(s, \pi(s,0)) + \alpha U^{\pi}(H(1)) \mid H(0) = h(0) = (s)\right] \\ &= I\!E^{\pi}\left[r(s, \pi(s,0)) + \alpha V^{\theta(\pi)}(S(1)) \mid S(0) = s\right] \end{aligned} \tag{4.18}$$

As a special case, for a stationary policy π,

$$V^{\pi}(s) = I\!E^{\pi}\left[r(s, \pi(s)) + \alpha V^{\pi}(S(1)) \mid S(0) = s\right] \tag{4.19}$$

Motivated by the finite horizon optimality equation (4.6), as well as by (4.18), we expect V^* to satisfy the following optimality equation.

$$V^*(s) = \sup_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \alpha I\!E\left[V^*(S(1)) \mid S(0) = s, A(0) = a\right] \right\} \tag{4.20}$$

Unlike the finite horizon case, we cannot use induction to establish the validity of (4.20). Instead we use the following approach. Let $\mathcal{V}$ denote the set of bounded functions $V : \mathcal{S} \mapsto I\!R$. Define the function $L^* : \mathcal{V} \mapsto \mathcal{V}$ by

$$L^*(V)(s) \equiv \sup_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \alpha I\!E\left[V(S(1)) \mid S(0) = s, A(0) = a\right] \right\}$$

Let $\|\cdot\|_\infty$ denote the supremum-norm on $\mathcal{V}$, that is, for any $V \in \mathcal{V}$, $\|V\|_\infty \equiv \sup_{s\in\mathcal{S}} |V(s)|$. For any $V_1, V_2 \in \mathcal{V}$, L^* satisfies $\|L^*(V_1) - L^*(V_2)\|_\infty \leq \alpha\|V_1 - V_2\|_\infty$. Then because $\alpha \in [0,1)$, L^* is a contraction mapping on $\mathcal{V}$, and it follows from the Banach fixed point theorem that L^* has a unique fixed point $v^* \in \mathcal{V}$, that is, there exists a unique function $v^* \in \mathcal{V}$ that satisfies $V = L^*(V)$. Thus optimality equation (4.20) has a unique solution v^*, and it remains to be shown that v^* is equal to V^* as defined in (4.16). Similarly, for any stationary policy π, define the function $L^\pi : \mathcal{V} \mapsto \mathcal{V}$ by

$$L^\pi(V)(s) \quad \equiv \quad I\!E^\pi \left[r(s, \pi(s)) + \alpha V(S(1)) \mid S(0) = s \right]$$

It follows in the same way as for L^* that L^π has a unique fixed point, and it follows from (4.19) that V^π is the fixed point of L^π.

Consider any $V \in \mathcal{V}$ such that $V \geq L^*(V)$. Then for any $\pi \in \Pi^{MR}$, it follows by induction that $V \geq V^\pi$, and thus $V \geq \sup_{\pi\in\Pi^{MR}} V^\pi \equiv V^*$. Similarly, consider any $V \in \mathcal{V}$ such that $V \leq L^*(V)$. Then for any $\varepsilon > 0$ there exists a stationary deterministic policy π_ε such that $V \leq V^{\pi_\varepsilon} + \varepsilon \leq V^* + \varepsilon$, and thus $V \leq V^*$. Combining these results, it follows that for any $V \in \mathcal{V}$ such that $V = L^*(V)$, it holds that $V = V^*$, and thus $v^* = V^*$, that is, V^* is the unique fixed point of L^*, and the validity of the optimality equation (4.20) has been established.

It can now be shown that it is sufficient to consider only stationary deterministic policies, in the following sense. First suppose that for each $s \in \mathcal{S}$, there exists a decision $a^*(s)$ such that

$$\begin{aligned} & r(s, a^*(s)) + \alpha I\!E \left[V^*(S(1)) \mid S(0) = s, A(0) = a^*(s) \right] \\ = \quad & \sup_{a\in\mathcal{A}(s)} \left\{ r(s,a) + \alpha I\!E \left[V^*(S(1)) \mid S(0) = s, A(0) = a \right] \right\} \end{aligned} \tag{4.21}$$

Let the stationary deterministic policy π^* be given by $\pi^*(s) = a^*(s)$. Then (4.21) implies that $L^{\pi^*}(V^*) = L^*(V^*) = V^*$, that is, V^* is a fixed point of L^{π^*}, and thus $V^{\pi^*} = V^*$. That is, for any history $h(t) = (h(t-1), a(t-1), s)$, $U^{\pi^*}(h(t)) = V^{\pi^*}(s) = V^*(s) = U^*(h(t))$, and thus π^* is an optimal policy. If, for some s, there does not exist such an optimal decision $a^*(s)$, then there also does not exist an optimal history dependent randomized policy. In such a case it still holds that for any $\varepsilon > 0$, there exists an ε-optimal stationary deterministic policy π^*_ε, obtained by choosing decisions $\pi^*_\varepsilon(s)$ such that

$$\begin{aligned} & r(s, \pi^*_\varepsilon(s)) + \alpha I\!E \left[V^*(S(1)) \mid S(0) = s, A(0) = \pi^*_\varepsilon(s) \right] \; + \; (1-\alpha)\varepsilon \\ > \quad & \sup_{a\in\mathcal{A}(s)} \left\{ r(s,a) + \alpha I\!E \left[V^*(S(1)) \mid S(0) = s, A(0) = a \right] \right\} \end{aligned} \tag{4.22}$$

4.5.2 Algorithms

Solving an infinite horizon discounted dynamic program usually involves computing V^*. An optimal policy $\pi^* \in \Pi^{SD}$ is then obtained using (4.21), or an ε-optimal policy $\pi^*_\varepsilon \in \Pi^{SD}$ is obtained using (4.22).

Unlike the finite horizon case, V^* is not computed directly using backward induction. An approach that is often used is to compute a sequence of approximating functions $V_i, i = 0, 1, 2, \ldots$, such that $V_i \to V^*$ as $i \to \infty$.

Approximating functions provide good policies, as shown by the following result. Suppose V^* is approximated by $\hat{V}$ such that $\left\| V^* - \hat{V} \right\|_\infty \leq \varepsilon$. Consider any policy $\hat{\pi} \in \Pi^{SD}$ such that

$$r(s, \hat{\pi}(s)) + \alpha \sum_{s'\in\mathcal{S}} p[s'|s, \hat{\pi}(s)] \hat{V}(s') \; + \; \delta \quad \geq \quad \sup_{a\in\mathcal{A}(s)} \left\{ r(s,a) + \alpha \sum_{s'\in\mathcal{S}} p[s'|s,a] \hat{V}(s') \right\}$$

for all $s \in \mathcal{S}$, that is, decision $\hat{\pi}(s)$ is within δ of the optimal decision using approximating function $\hat{V}$ on the right hand side of the optimality equation (4.20). Then

$$V^{\hat{\pi}}(s) \quad \geq \quad V^*(s) - \frac{2\alpha\varepsilon + \delta}{1-\alpha} \tag{4.23}$$

for all $s \in \mathcal{S}$, that is, policy $\hat{\pi}$ has value function within $(2\alpha\varepsilon + \delta)/(1-\alpha)$ of the optimal value function.

Value Iteration One algorithm based on a sequence of approximating functions V_i is called value iteration, or successive approximation. The iterates V_i of value iteration correspond to the value function $V^*(s, T+1-i)$ of the finite horizon dynamic program with the same problem parameters. Specifically, starting with initial approximation $V_0(s) = 0 = V^*(s, T+1)$ for all s, the ith approximating function $V_i(s)$ is the same as the value function $V^*(s, T+1-i)$ of the corresponding finite horizon dynamic program, that is, the value function for time $T+1-i$ that is obtained after i steps of the backward induction algorithm.

Value Iteration Algorithm
0. Choose initial approximation $V_0 \in \mathcal{V}$ and stopping tolerance ε. Set $i \leftarrow 0$.
1. For each $s \in \mathcal{S}$, compute

$$V_{i+1}(s) \quad = \quad \sup_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s,a] V_i(s') \right\} \tag{4.24}$$

2. If $\|V_{i+1} - V_i\|_\infty < (1-\alpha)\varepsilon/2\alpha$, then go to step 3. Otherwise, set $i \leftarrow i+1$ and go to step 1.
3. For each $s \in \mathcal{S}$, choose a decision

$$\pi^*_\varepsilon(s) \quad \in \quad \arg\max_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s,a] V_{i+1}(s') \right\}$$

if the maximum on the right hand side is attained. Otherwise, for any chosen $\delta > 0$, choose a decision $\pi^*_\delta(s)$ such that

$$\begin{aligned} & r(s, \pi^*_\delta(s)) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s, \pi^*_\delta(s)] V_{i+1}(s') + (1-\alpha)\delta \\ > \quad & \sup_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s,a] V_{i+1}(s') \right\} \end{aligned}$$

It can be shown, using the contraction property of L^*, that $V_i \to V^*$ as $i \to \infty$ for any initial approximation $V_0 \in \mathcal{V}$. Also, the convergence is geometric with rate α. Specifically, for any $V_0 \in \mathcal{V}$, $\|V_i - V^*\|_\infty \leq \alpha^i \|V_0 - V^*\|_\infty$. That implies that the convergence rate is faster if the discount factor α is smaller.

When the value iteration algorithm stops, the final approximation V_{i+1} satisfies $\|V_{i+1} - V^*\|_\infty < \varepsilon/2$. Furthermore, the chosen policy π^*_ε is an ε-optimal policy, and the chosen policy π^*_δ is an $(\varepsilon+\delta)$-optimal policy.

There are several versions of the value iteration algorithm. One example is Gauss-Seidel value iteration, which uses the most up-to-date approximation $V_{i+1}(s')$ on the right hand side of (4.24) as soon as it becomes available, instead of using the previous approximation $V_i(s')$ as shown in (4.24). Gauss-Seidel value iteration has the same convergence properties and performance guarantees given above, but in practice it usually converges faster.

Policy Iteration Policy iteration is an algorithm based on a sequence of policies π_i, and their value functions V^{π_i}.

Policy Iteration Algorithm
0. Choose initial policy $\pi_0 \in \Pi^{SD}$ and stopping tolerance ε. Set $i \leftarrow 0$.
1. Compute the value function V^{π_i} of policy π_i by solving the system of linear equations

$$V^{\pi_i}(s) \quad = \quad r(s, \pi_i(s)) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s, \pi_i(s)] V^{\pi_i}(s') \tag{4.25}$$

for each $s \in \mathcal{S}$.
2. For each $s \in \mathcal{S}$, choose a decision

$$\pi_{i+1}(s) \quad \in \quad \arg\max_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s,a] V^{\pi_i}(s') \right\}$$

if the maximum on the right hand side is attained. Otherwise, for any chosen $\delta > 0$, choose a decision $\pi_{i+1}(s)$ such that

$$\begin{aligned} & r(s, \pi_{i+1}(s)) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s, \pi_{i+1}(s)] V^{\pi_i}(s') + (1-\alpha)\delta \\ > & \sup_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s,a] V^{\pi_i}(s') \right\} \end{aligned}$$

3. If $\pi_{i+1} = \pi_i$ or $i > 0$ and $\|V^{\pi_i} - V^{\pi_{i-1}}\|_\infty < (1-\alpha)\varepsilon/2\alpha$, then stop with chosen policy π_{i+1}. Otherwise, set $i \leftarrow i+1$ and go to step 1.

It can be shown that policy iteration converges at least as fast as value iteration as $i \to \infty$. However, the amount of work involved in each iteration of the policy iteration algorithm is usually more than the amount of work involved in each iteration of the value iteration algorithm, because of the computational effort required to solve (4.25) for V^{π_i}. The total computational effort to satisfy the stopping criterion with policy iteration is usually more than the total computational effort with value iteration.

A desirable property of the iterates V^{π_i} is that if each π_i attains the maximum on the right hand side, then they are monotonically improving, that is $V^{\pi_0} \leq V^{\pi_1} \leq \cdots \leq V^*$. Thus, each iteration produces a better policy than before. If one starts with a reasonably good heuristic policy π_0, then even if one performs only one iteration of the policy iteration algorithm, one obtains the benefit of an even better policy.

Suppose the policy iteration algorithm stops with $\pi_{i+1} = \pi_i$. If π_{i+1} attains the maximum on the right hand side, then $V^{\pi_i} = V^*$ and the chosen policy π_{i+1} is optimal. If π_{i+1} chooses a decision within $(1-\alpha)\delta$ of the maximum on the right hand side, then $\|V^{\pi_i} - V^*\|_\infty < \delta$ and the chosen policy π_{i+1} is δ-optimal. Otherwise, suppose the policy iteration algorithm stops with $\|V^{\pi_i} - V^{\pi_{i-1}}\|_\infty < (1-\alpha)\varepsilon/2\alpha$. If π_{i+1} attains the maximum on the right hand side, then $\|V^{\pi_{i+1}} - V^*\|_\infty < \varepsilon$ and the chosen policy π_{i+1} is ε-optimal. If π_{i+1} chooses a decision within $(1-\alpha)\delta$ of the maximum on the right hand side, then $\|V^{\pi_{i+1}} - V^*\|_\infty < \varepsilon + \delta$ and the chosen policy π_{i+1} is $(\varepsilon+\delta)$-optimal.

Modified Policy Iteration It was mentioned that one of the drawbacks of the policy iteration algorithm is the computational effort required to solve (4.25) for V^{π_i}. An iterative algorithm called the Gauss-Seidel method can be used to solve (4.25). For any stationary policy π, L^π is a contraction mapping. It follows that for any $V_0 \in \mathcal{V}$, the sequence of functions $V_j, j = 0, 1, 2, \ldots$, inductively computed by

$$V_{j+1}(s) \quad = \quad r(s, \pi(s)) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s, \pi(s)] V_j(s')$$

for each $s \in \mathcal{S}$, converges to V^π as $j \to \infty$. Modified policy iteration uses this Gauss-Seidel method to compute V^π, but it performs only a few iterations to compute an approximation to V^π, and then moves on to the next policy, instead of letting $j \to \infty$ to compute V^π exactly. This compensates for the drawbacks of policy iteration. Modified policy iteration is usually more efficient than value iteration and policy iteration.

Modified Policy Iteration Algorithm
0. Choose initial approximation $V_{1,0} \in \mathcal{V}$, a method to generate a sequence $N_i, i = 1, 2, \ldots$, of positive integers, and stopping tolerance ε. Set $i \leftarrow 1$.
1. For each $s \in \mathcal{S}$, choose a decision

$$\pi_i(s) \quad \in \quad \arg\max_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s,a] V_{i,0}(s') \right\}$$

if the maximum on the right hand side is attained. Otherwise, for any chosen $\delta > 0$, choose a decision $\pi_i(s)$

such that

$$r(s,\pi_i(s)) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s,\pi_i(s)] V_{i,0}(s') + (1-\alpha)\delta$$
$$> \sup_{a \in \mathcal{A}(s)} \left\{ r(s,a) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s,a] V_{i,0}(s') \right\}$$

2. For $j = 1, 2, \ldots, N_i$, compute

$$V_{i,j}(s) \quad = \quad r(s,\pi_i(s)) + \alpha \sum_{s' \in \mathcal{S}} p[s'|s,\pi_i(s)] V_{i,j-1}(s')$$

for each $s \in \mathcal{S}$.
3. If $\|V_{i,1} - V_{i,0}\|_\infty < (1-\alpha)\varepsilon/2\alpha$, then stop with chosen policy π_i. Otherwise, set $V_{i+1,0} = V_{i,N_i}$, set $i \leftarrow i+1$ and go to step 1.

It can be shown that modified policy iteration converges at least as fast as value iteration as $i \to \infty$. The special case of modified policy iteration with $N_i = 1$ for all i is the same as value iteration (as long as $V_{i,1}$ is set equal to the maximum of the right hand side). When $N_i \to \infty$ for all i, modified policy iteration is the same as policy iteration.

The sequence $N_i, i = 1, 2, \ldots$, can be chosen in many ways. Some alternatives are to choose $N_i = N$ for some chosen fixed N for all i, or to choose N_i to be the first minor iteration j such that $\|V_{i,j} - V_{i,j-1}\|_\infty < \eta_i$ for some chosen sequence (typically decreasing) η_i. The idea is to choose N_i to obtain the best trade-off between the computational requirements of step 1, in which an optimization problem is solved to obtain a new policy, and that of step 2, in which a more accurate approximation of the value function of the current policy is computed. If the optimization problem in step 1 requires a lot of computational effort, then it is better to obtain more accurate approximations of the value functions between successive executions of step 1, that is, it is better to choose N_i larger, and vice versa. Also, if the policy does not change much from one major iteration to the next, that is, if policies π_{i-1} and π_i are very similar, then it is also better to obtain a more accurate approximation of the value function V^{π_i} by choosing N_i larger. It is typical that the policies do not change much later in the algorithm, and hence it is typical to choose N_i to be increasing in i.

When the modified policy iteration algorithm stops, the approximation $V^{i,1}$ satisfies $\|V^{i,1} - V^*\|_\infty < \varepsilon/2$. If the chosen policy π_i attains the maximum on the right hand side, then π_i is ε-optimal. If π_i chooses a decision within $(1-\alpha)\delta$ of the maximum on the right hand side, then π_i is $(\varepsilon+\delta)$-optimal. Furthermore, if the initial approximation $V_{1,0}$ satisfies $L^*(V_{1,0}) \geq V_{1,0}$, such as if $V_{1,0} = V^{\pi_0}$ for some initial policy π_0, and if each π_i attains the maximum on the right hand side, then the sequence of policies π_i are monotonically improving, and $V_{i,j-1} \leq V_{i,j}$ for each i and j, from which it also follows that $V_{i-1,0} \leq V_{i,0}$ for each i.

4.6 Approximation Methods

For many interesting applications the state space $\mathcal{S}$ is too big for any of the algorithms discussed so far to be used. This is usually due to the "curse of dimensionality"—the phenomenon that the number of states grows exponentially in the number of dimensions of the state space. When the state space is too large, not only is the computational effort required by these algorithms excessive, but storing the value function and policy values for each state is impossible with current technology.

Recall that solving a dynamic program usually involves using (4.6) in the finite horizon case or (4.20) in the infinite horizon case to compute the optimal value function V^*, and an optimal policy π^*. To accomplish this, the following major computational tasks are performed.

1. Estimation of the optimal value function V^* on the right hand side of (4.6) or (4.20).

2. Estimation of the expected value on the right hand side of (4.6) or (4.20). For many applications, this is a high dimensional integral that requires a lot of computational effort to compute accurately.

3. The maximization problem on the right hand side of (4.6) or (4.20) has to be solved to determine the optimal decision for each state. This maximization problem may be easy or hard, depending on the application. The first part of this article discusses several methods for solving such stochastic optimization problems.

Approximation methods usually involve approaches to perform one or more of these computational tasks efficiently, sometimes by sacrificing optimality.

For many applications the state space is uncountable and the transition and cost functions are too complex for closed form solutions to be obtained. To compute solutions for such problems, the state space is often discretized. Discretization methods and convergence results are discussed in Wong (1970a), Fox (1973), Bertsekas (1975), Kushner (1990), Chow and Tsitsiklis (1991), and Kushner and Dupuis (1992).

For many other applications, such as queueing systems, the state space is countably infinite. Computing solutions for such problems usually involves solving smaller dynamic programs with finite state spaces, often obtained by truncating the state space of the original DP, and then using the solutions of the smaller DPs to obtain good solutions for the original DP. Such approaches and their convergence are discussed in Fox (1971), White (1980a), White (1980b), White (1982), Thomas and Stengos (1985), Cavazos-Cadena (1986), Van Dijk (1991b), Van Dijk (1991a), Sennott (1997a), and Sennott (1997b).

Even if the state space is not infinite, the number of states may be very large. A natural approach is to aggregate states, usually by collecting similar states into subsets, and then to solve a related DP with the aggregated state space. Aggregation and aggregation/disaggregation methods are discussed in Simon and Ando (1961), Mendelssohn (1982), Stewart (1983), Chatelin (1984), Schweitzer (1984), Schweitzer, Puterman and Kindle (1985), Schweitzer (1986), Schweitzer and Kindle (1986), Bean, Birge and Smith (1987), Feinberg and Chiu (1987), and Bertsekas and Castanon (1989).

Another natural approach for dealing with a large-scale DP is to decompose the DP into smaller related DPs, which are easier to solve, and then to use the solutions of the smaller DPs to obtain a good solution for the original DP. Decomposition methods are discussed in Wong (1970b), Collins and Lew (1970), Collins (1970), Collins and Angel (1971), Courtois (1977), Courtois and Semal (1984), Stewart (1984), and Kleywegt, Nori and Savelsbergh (1999).

Some general state space reduction methods that include many of the methods mentioned above are analyzed in Whitt (1978), Whitt (1979b), Whitt (1979a), Hinderer (1976), Hinderer and Hübner (1977), Hinderer (1978), and Haurie and L'Ecuyer (1986). Surveys are given in Morin (1978), and Rogers et al. (1991).

Another natural and quite different approach for dealing with DPs with large state spaces, is to approximate the optimal value function V^* with an approximating function $\hat{V}$. It was shown in Section 4.5.2 that good approximations $\hat{V}$ to the optimal value function V^* lead to good policies $\hat{\pi}$. Polynomial approximations, often using orthogonal polynomials such as Legendre and Chebychev polynomials, have been suggested by Bellman and Dreyfus (1959), Chang (1966), Bellman, Kalaba and Kotkin (1963), and Schweitzer and Seidman (1985). Approximations using splines have been suggested by Daniel (1976), and approximations using regression splines by Chen, Ruppert and Shoemaker (1999). Estimation of the parameters of approximating functions for infinite horizon discounted DPs has been studied in Tsitsiklis and Van Roy (1996), Van Roy and Tsitsiklis (1996), and Bertsekas and Tsitsiklis (1996). Some of this work was motivated by approaches proposed for reinforcement learning; see Sutton and Barto (1998) for an overview.

References

Albritton, M., Shapiro, A. and Spearman, M. L. 1999. Finite Capacity Production Planning with Random Demand and Limited Information. preprint.

Beale, E. M. L. 1955. On Minimizing a Convex Function Subject to Linear Inequalities. *Journal of the Royal Statistical Society, Series B*, **17**, 173–184.

Bean, J. C., Birge, J. R. and Smith, R. L. 1987. Aggregation in Dynamic Programming. *Operations Research*, **35**, 215–220.

Bellman, R. and Dreyfus, S. 1959. Functional Approximations and Dynamic Programming. *Mathematical Tables and Other Aids to Computation*, **13**, 247–251.

Bellman, R. E. 1957. *Dynamic Programming*. Princeton University Press, Princeton, NJ.

Bellman, R. E. 1961. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, Princeton, NJ.

Bellman, R. E. and Dreyfus, S. 1962. *Applied Dynamic Programming*. Princeton University Press, Princeton, NJ.

Bellman, R. E., Kalaba, R. and Kotkin, B. 1963. Polynomial Approximation—A New Computational Technique in Dynamic Programming: Allocation Processes. *Mathematics of Computation*, **17**, 155–161.

Benveniste, A., Métivier, M. and Priouret, P. 1990. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, Berlin, Germany.

Berger, J. O. 1985. *Statistical Decision Theory and Bayesian Analysis*. 2nd edn, Springer-Verlag, New York, NY.

Bertsekas, D. P. 1975. Convergence of Discretization Procedures in Dynamic Programming. *IEEE Transactions on Automatic Control*, **AC-20**, 415–419.

Bertsekas, D. P. 1995. *Dynamic Programming and Optimal Control*. Athena Scientific, Belmont, MA.

Bertsekas, D. P. and Castanon, D. A. 1989. Adaptive Aggregation Methods for Infinite Horizon Dynamic Programming. *IEEE Transactions on Automatic Control*, **AC-34**, 589–598.

Bertsekas, D. P. and Shreve, S. E. 1978. *Stochastic Optimal Control: The Discrete Time Case*. Academic Press, New York, NY.

Bertsekas, D. P. and Tsitsiklis, J. N. 1996. *Neuro-Dynamic Programming*. Athena Scientific, New York, NY.

Birge, J. R. and Louveaux, F. 1997. *Introduction to Stochastic Programming*. Springer Series in Operations Research, Springer-Verlag, New York, NY.

Cavazos-Cadena, R. 1986. Finite-State Approximations for Denumerable State Discounted Markov Decision Processes. *Applied Mathematics and Optimization*, **14**, 1–26.

Chang, C. S. 1966. Discrete-Sample Curve Fitting Using Chebyshev Polynomials and the Approximate Determination of Optimal Trajectories via Dynamic Programming. *IEEE Transactions on Automatic Control*, **AC-11**, 116–118.

Chatelin, F. 1984. Iterative Aggregation/Disaggregation Methods. In *Mathematical Computer Performance and Reliability*. G. Iazeolla, P. J. Courtois and A. Hordijk (editors). Elsevier Science Publishers B.V., Amsterdam, Netherlands, chapter 2.1, 199–207.

Chen, V. C. P., Ruppert, D. and Shoemaker, C. A. 1999. Applying Experimental Design and Regression Splines to High-Dimensional Continuous-State Stochastic Dynamic Programming. *Operations Research*, **47**, 38–53.

Chong, E. K. P. and Ramadge, P. J. 1992. Convergence of Recursive Optimization Algorithms Using Infinitesimal Perturbation Analysis Estimates. *Discrete Event Dynamic Systems: Theory and Applications*, **1**, 339–372.

Chow, C. S. and Tsitsiklis, J. N. 1991. An Optimal One-Way Multigrid Algorithm for Discrete-Time Stochastic Control. *IEEE Transactions on Automatic Control*, **AC-36**, 898–914.

Collins, D. C. 1970. Reduction of Dimensionality in Dynamic Programming via the Method of Diagonal Decomposition. *Journal of Mathematical Analysis and Applications*, **31**, 223–234.

Collins, D. C. and Angel, E. S. 1971. The Diagonal Decomposition Technique Applied to the Dynamic Programming Solution of Elliptic Partial Differential Equations. *Journal of Mathematical Analysis and Applications*, **33**, 467–481.

Collins, D. C. and Lew, A. 1970. A Dimensional Approximation in Dynamic Programming by Structural Decomposition. *Journal of Mathematical Analysis and Applications*, **30**, 375–384.

Courtois, P. J. 1977. *Decomposability: Queueing and Computer System Applications*. Academic Press, New York, NY.

Courtois, P. J. and Semal, P. 1984. Error Bounds for the Analysis by Decomposition of Non-Negative Matrices. In *Mathematical Computer Performance and Reliability*. G. Iazeolla, P. J. Courtois and A. Hordijk (editors). Elsevier Science Publishers B.V., Amsterdam, Netherlands, chapter 2.2, 209–224.

Daniel, J. W. 1976. Splines and Efficiency in Dynamic Programming. *Journal of Mathematical Analysis and Applications*, **54**, 402–407.

Dantzig, G. B. 1955. Linear Programming under Uncertainty. *Management Science*, **1**, 197–206.

Denardo, E. V. 1982. *Dynamic Programming Models and Applications*. Prentice-Hall, Englewood Cliffs, NJ.

Feinberg, B. N. and Chiu, S. S. 1987. A Method to Calculate Steady-State Distributions of Large Markov Chains by Aggregating States. *Operations Research*, **35**, 282–290.

Fox, B. L. 1971. Finite-State Approximations to Denumerable-State Dynamic Programs. *Journal of Mathematical Analysis and Applications*, **34**, 665–670.

Fox, B. L. 1973. Discretizing Dynamic Programs. *Journal of Optimization Theory and Applications*, **11**, 228–234.

Glasserman, P. 1991. *Gradient Estimation via Perturbation Analysis*. Kluwer Academic Publishers, Norwell, MA.

Glynn, P. W. 1990. Likelihood Ratio Gradient Estimation for Stochastic Systems. *Communications of the ACM*, **33**, 75–84.

Haurie, A. and L'Ecuyer, P. 1986. Approximation and Bounds in Discrete Event Dynamic Programming. *IEEE Transactions on Automatic Control*, **AC-31**, 227–235.

Heyman, D. P. and Sobel, M. J. 1984. *Stochastic Models in Operations Research*. Vol. II, McGraw-Hill, New York, NY.

Hinderer, K. 1970. *Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter*. Springer-Verlag, Berlin.

Hinderer, K. 1976. Estimates for Finite-Stage Dynamic Programs. *Journal of Mathematical Analysis and Applications*, **55**, 207–238.

Hinderer, K. 1978. On Approximate Solutions of Finite-Stage Dynamic Programs. In *Dynamic Programmming and its Applications*. M. L. Puterman (editor). Academic Press, New York, NY, 289–317.

Hinderer, K. and Hübner, G. 1977. On Exact and Approximate Solutions of Unstructured Finite-Stage Dynamic Programs. In *Markov Decision Theory : Proceedings of the Advanced Seminar on Markov Decision Theory held at Amsterdam, The Netherlands, September 13–17, 1976*. H. C. Tijms and J. Wessels (editors). Mathematisch Centrum, Amsterdam, The Netherlands, 57–76.

Hiriart-Urruty, J. B. and Lemarechal, C. 1993. *Convex Analysis and Minimization Algorithms*. Springer-Verlag, Berlin, Germany.

Ho, Y. C. and Cao, X. R. 1991. *Perturbation Analysis of Discrete Event Dynamic Systems*. Kluwer Academic Publishers, Norwell, MA.

Kall, P. and Wallace, S. W. 1994. *Stochastic Programming*. John Wiley & Sons, Chichester, England.

Klein Haneveld, W. K. and Van der Vlerk, M. H. 1999. Stochastic Integer Programming: General Models and Algorithms. *Annals of Operations Research*, **85**, 39–57.

Kleywegt, A. J., Nori, V. S. and Savelsbergh, M. W. P. 1999. The Stochastic Inventory Routing Problem with Direct Deliveries, *Technical Report TLI99-01*, The Logistics Institute, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332-0205.

Kleywegt, A. J. and Shapiro, A. 1999. The Sample Average Approximation Method for Stochastic Discrete Optimization. Preprint, available at: Stochastic Programming E-Print Series, `http://dochost.rz.hu-berlin.de/speps/`.

Kushner, H. J. 1990. Numerical Methods for Continuous Control Problems in Continuous Time. *SIAM Journal on Control and Optimization*, **28**, 999–1048.

Kushner, H. J. and Clark, D. S. 1978. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, Berlin, Germany.

Kushner, H. J. and Dupuis, P. 1992. *Numerical Methods for Stochastic Control Problems in Continuous Time*. Springer-Verlag, New York, NY.

L'Ecuyer, P. and Glynn, P. W. 1994. Stochastic Optimization by Simulation: Convergence Proofs for the GI/G/1 Queue in Steady-State. *Management Science*, **11**, 1562–1578.

Mendelssohn, R. 1982. An Iterative Aggregation Procedure for Markov Decision Processes. *Operations Research*, **30**, 62–73.

Morin, T. 1978. Computational Advances in Dynamic Programming. In *Dynamic Programmming and its Applications*. M. L. Puterman (editor). Academic Press, New York, NY, 53–90.

Nemhauser, G. L. 1966. *Introduction to Dynamic Programming*. Wiley, New York, NY.

Norkin, V. I., Pflug, G. C. and Ruszczyński, A. 1998. A Branch and Bound Method for Stochastic Global Optimization. *Mathematical Programming*, **83**, 425–450.

Puterman, M. L. 1994. *Markov Decision Processes*. John Wiley & Sons, Inc., New York, NY.

Robbins, H. and Monro, S. 1951. On a Stochastic Approximation Method. *Annals of Mathematical Statistics*, **22**, 400–407.

Robinson, S. M. 1996. Analysis of Sample-Path Optimization. *Mathematics of Operations Research*, **21**, 513–528.

Rogers, D. F., Plante, R. D., Wong, R. T. and Evans, J. R. 1991. Aggregation and Disaggregation Techniques and Methodology in Optimization. *Operations Research*, **39**, 553–582.

Ross, S. M. 1970. *Applied Probability Models with Optimization Applications*. Dover, New York, NY.

Ross, S. M. 1983. *Introduction to Stochastic Dynamic Programming*. Academic Press, New York, NY.

Rubinstein, R. Y. and Shapiro, A. 1990. Optimization of Simulation Models by the Score Function Method. *Mathematics and Computers in Simulation*, **32**, 373–392.

Rubinstein, R. Y. and Shapiro, A. 1993. *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. John Wiley & Sons, Chichester, England.

Ruppert, D. 1991. Stochastic Approximation. In *Handbook of Sequential Analysis*. B. K. Ghosh and P. K. Sen (editors). Marcel Dekker, New York, NY, 503–529.

Schultz, R., Stougie, L. and Van der Vlerk, M. H. 1998. Solving Stochastic Programs with Integer Recourse by Enumeration: a Framework Using Gröbner Basis Reductions. *Mathematical Programming*, **83**, 229–252.

Schweitzer, P. J. 1984. Aggregation Methods for Large Markov Chains. In *Mathematical Computer Performance and Reliability*. G. Iazeolla, P. J. Courtois and A. Hordijk (editors). Elsevier Science Publishers, Amsterdam, Netherlands, 275–286.

Schweitzer, P. J. 1986. An Iterative Aggregation-Disaggregation Algorithm for Solving Undiscounted Semi-Markovian Reward Processes. *Stochastic Models*, **2**, 1–41.

Schweitzer, P. J. and Kindle, K. W. 1986. Iterative Aggregation for Solving Undiscounted Semi-Markovian Reward Processes. *Communications in Statistics. Stochastic Models*, **2**, 1–41.

Schweitzer, P. J., Puterman, M. L. and Kindle, K. W. 1985. Iterative Aggregation-Disaggregation Procedures for Discounted Semi-Markov Reward Processes. *Operations Research*, **33**, 589–605.

Schweitzer, P. J. and Seidman, A. 1985. Generalized Polynomial Approximations in Markovian Decision Processes. *Journal of Mathematical Analysis and Applications*, **110**, 568–582.

Sennott, L. I. 1997a. The Computation of Average Optimal Policies in Denumerable State Markov Decision Chains. *Advances in Applied Probability*, **29**, 114–137.

Sennott, L. I. 1997b. On Computing Average Cost Optimal Policies with Application to Routing to Parallel Queues. *Zeitschrift für Operations Research*, **45**, 45–62.

Sennott, L. I. 1999. *Stochastic Dynamic Programming and the Control of Queueing Systems*. John Wiley & Sons, New York, NY.

Serfozo, R. F. 1976. Monotone Optimal Policies for Markov Decision Processes. *Mathematical Programming Study*, **6**, 202–215.

Shapiro, A. 1996. Simulation-based Optimization: Convergence Analysis and Statistical Inference. *Stochastic Models*, **12**, 425–454.

Shapiro, A. and Homem-de-Mello, T. 1998. A Simulation-Based Approach to Two-Stage Stochastic Programming with Recourse. *Mathematical Programming*, **81**, 301–325.

Shapiro, A. and Homem-de-Mello, T. 1999. On Rate of Convergence of Monte Carlo Approximations of Stochastic Programs. Preprint, available at: Stochastic Programming E-Print Series, `http://dochost.rz.hu-berlin.de/speps/`.

Simon, H. A. and Ando, A. 1961. Aggregation of Variables in Dynamic Systems. *Econometrica*, **29**, 111–138.

Stewart, G. W. 1983. Computable Error Bounds for Aggregated Markov Chains. *Journal of the Association for Computing Machinery*, **30**, 271–285.

Stewart, G. W. 1984. On the Structure of Nearly Uncoupled Markov Chains. In *Mathematical Computer Performance and Reliability*. G. Iazeolla, P. J. Courtois and A. Hordijk (editors). Elsevier Science Publishers B.V., Amsterdam, Netherlands, chapter 2.7, 287–302.

Sutton, R. S. and Barto, A. G. 1998. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Thomas, L. C. and Stengos, D. 1985. Finite State Approximation Algorithms for Average Cost Denumerable State Markov Decision Processes. *OR Spektrum*, **7**, 27–37.

Topkis, D. M. 1978. Minimizing a Submodular Function on a Lattice. *Operations Research*, **26**, 305–321.

Tsitsiklis, J. N. and Van Roy, B. 1996. Feature-Based Methods for Large-Scale Dynamic Programming. *Machine Learning*, **22**, 59–94.

Van Dijk, N. 1991a. On Truncations and Perturbations of Markov Decision Problems with an Application to Queueing Network Overflow Control. *Annals of Operations Research*, **29**, 515–536.

Van Dijk, N. 1991b. Truncation of Markov Chains with Applications to Queueing. *Operations Research*, **39**, 1018–1026.

Van Roy, B. and Tsitsiklis, J. N. 1996. Stable Linear Approximations to Dynamic Programming for Stochastic Control Problems with Local Transitions. *Advances in Neural Information Processing Systems 8*. MIT Press, Cambridge, MA, 1045–1051.

Van Slyke, R. and Wets, R. J. B. 1969. L-Shaped Linear Programs with Application to Optimal Control and Stochastic Programming. *SIAM Journal on Applied Mathematics*, **17**, 638–663.

White, D. J. 1980a. Finite-State Approximations for Denumerable-State Infinite-Horizon Discounted Markov Decision Processes: The Method of Successive Approximations. In *Recent Developments in Markov Decision Processes*. R. Hartley, L. C. Thomas and D. J. White (editors). Academic Press, New York, NY, 57–72.

White, D. J. 1980b. Finite-State Approximations for Denumerable-State Infinite-Horizon Discounted Markov Decision Processes. *Journal of Mathematical Analysis and Applications*, **74**, 292–295.

White, D. J. 1982. Finite-State Approximations for Denumerable-State Infinite Horizon Discounted Markov Decision Processes with Unbounded Rewards. *Journal of Mathematical Analysis and Applications*, **86**, 292–306.

Whitt, W. 1978. Approximations of Dynamic Programs, I. *Mathematics of Operations Research*, **3**, 231–243.

Whitt, W. 1979a. A-Priori Bounds for Approximations of Markov Programs. *Journal of Mathematical Analysis and Applications*, **71**, 297–302.

Whitt, W. 1979b. Approximations of Dynamic Programs, II. *Mathematics of Operations Research*, **4**, 179–185.

Wong, P. J. 1970a. An Approach to Reducing the Computing Time for Dynamic Programming. *Operations Research*, **18**, 181–185.

Wong, P. J. 1970b. A New Decomposition Procedure for Dynamic Programming. *Operations Research*, **18**, 119–131.