C

Calculus of Variations

Stephen G. Nash
George Mason University, Fairfax, VA, USA

Introduction

The calculus of variations is the grandparent of mathematical programming. From it came such concepts as duality and Lagrange multipliers. Many central ideas in optimization were first developed for the calculus of variations, then specialized to nonlinear programming, all of this happening years before linear programming came along.

The calculus of variations solves optimization problems whose parameters are not simple variables, but rather functions. For example, how should the shape of an automobile hood be chosen so as to minimize air resistance? Or, what path does a ray of light follow in an irregular medium? The calculus of variations is closely related to optimal control theory, where a set of controls are used to achieve a certain goal in an optimal way. For example, the pilot of an aircraft might wish to use the throttle and flaps to achieve a particular cruising altitude and velocity in a minimum amount of time or using a minimum amount of fuel. The modern world is full of devices designed using optimal control — in cars, elevators, heating systems, stereos, etc.

Brachistochrone Problem

The calculus of variations was inspired by problems in mechanics, especially the study of three-dimensional motion. It was used in the 18th and 19th centuries to derive many important laws of physics. This was done using the Principle of Least Action. Action is defined to be the integral of the product of mass, velocity, and distance. The Principle of Least Action asserts that nature acts so as to minimize this integral. To apply the principle, the formula for the action integral would be specialized to the setting under study, and then the calculus of variations would be used to optimize the integral. This general approach was used to derive important equations in mechanics, fluid dynamics, and other fields.

The most famous problem in the calculus of variations was posed in 1696 by John Bernoulli. It is called the Brachistochrone ("least time") problem, and asks what path a pellet should follow to drop between two points in the shortest amount of time, with gravity the only force acting on the pellet. The solution to the Brachistochrone problem can be found by solving

$$\underset{y(t)}{\text{minimize}}\ \frac{1}{\sqrt{2g}}\int_{t_1}^{t_2}\sqrt{\frac{1+y'(t)^2}{y(t)}}dt$$

where g is the gravitational constant. If this were a finite-dimensional problem then it could be solved by setting the derivative of the objective function equal to zero, but seventeenth-century mathematics did not know how to take a derivative with respect to a function.

The Brachistochrone problem was solved at the time by Newton and others, but the general techniques that inspired the name calculus of variations were not developed until several decades

S.I. Gass, M.C. Fu (eds.), *Encyclopedia of Operations Research and Management Science*,
DOI 10.1007/978-1-4419-1153-7,

later. The first major results were obtained by Euler in the 1740s. He considered various problems of the general form

$$\underset{y(t)}{\text{minimize}} \int_{t_1}^{t_2} f(t, y(t), y'(t))dt.$$

The Brachistochrone problem is of this form. Euler solved these problems by discretizing the solution $y(t)$ — approximating the solution by its values at finitely many points. This gave a finite-dimensional problem that could be solved using the techniques of calculus. Euler then took the limit of the approximate solutions as the number of discretization points tended to infinity. This approach was difficult and restrictive, because it had to be adapted to the specifics of the problem being solved, and because there were restrictions on the types of problems for which it was successful.

Far more influential was the approach of Lagrange. He suggested that the solution be perturbed or varied from $y(t)$ to $y(t) + \varepsilon\, z(t)$, where ε is a small number and $z(t)$ is some arbitrary function that satisfies $z(t_1) = z(t_2) = 0$. For the Brachistochrone problem this latter condition ensures that the perturbed function still represents a path between the two points.

If $y(t)$ is a solution to the problem

$$\underset{y(t)}{\text{minimize}} \int_{t_1}^{t_2} f(t, y(t), y'(t))dt,$$

then $\varepsilon = 0$ will be a solution to

$$\underset{\varepsilon}{\text{minimize}} \int_{t_1}^{t_2} f(t, y(t) + \varepsilon z(t), y'(t) + \varepsilon z'(t))dt$$

This observation allowed Lagrange to convert the original infinite-dimensional problem to a one-dimensional problem that could be analyzed using ordinary calculus. Setting the derivative of the integral with respect to $\in$ equal to zero at the point $\varepsilon = 0$ leads to the equation

$$\frac{d}{dt}\frac{\partial f}{\partial y'} - \frac{\partial f}{\partial y} = 0.$$

This final condition is a first-order optimality condition for an unconstrained calculus-of-variations problem. It was first discovered by Euler, but the derivation here is due to Lagrange.

The name "calculus of variations was chosen by Euler and was inspired by Lagrange's approach in varying the function $y(t)$. The optimality condition is stated as the first variation must equal zero by analogy with the condition $f'(x) = 0$ for a one-variable optimization problem. Euler was so impressed with Lagrange's work that he held back his own papers on the topic so that Lagrange could publish first, a magnanimous gesture by the renowned Euler to the then young and unknown Lagrange.

There are additional first-order optimality conditions for calculus of variations problems. The theory is more complicated than for finite-dimensional optimization, and the necessary and sufficient conditions for an optimal solution were not fully understood until the 1870s, when Weierstrass studied this topic. A discussion of this theory can be found in Gregory and Lin (1992).

Multipliers

Constraints can be added to problems in the calculus of variations just as in other optimization problems. A constraint might represent the principle of conservation of energy, or perhaps that the motion was restricted in some way, for example that a planet was traveling in a particular orbit around the sun.

Both Euler and Lagrange considered problems of this type, and both were led to the concept of a multiplier. In the calculus of variations the multiplier might be a scalar (as it is in finite-dimensional problems) or, depending on the particular form of the constraint, it might be a function of the independent variable t. They have come to be called Lagrange multipliers; but, as with the optimality condition, Euler discovered them first.

In his book *Mécanique Analytique*, Lagrange includes an interpretation of the multiplier terms. He writes that they can be considered as representing the moments of forces acting on the moving particle, and serving to keep the constraints satisfied. This point of view is the basis for duality theory, although Lagrange does not seem to have followed up on this idea.

Duality

Duality theory did not become fully developed until early in this century, with many of the important steps

coming from the calculus of variations. At first there were only isolated examples of duality. That is, someone would notice that a pair of problems — one a maximization problem, one a minimization problem — would have optimal solutions that were related to each other. An early example of this type was published in 1755, and is described in Kuhn (1991). In the nineteenth century various other examples were noticed, such as the relationship between currents and voltages in an electrical circuit. Gradually it was understood that duality was not an accidental phenomenon peculiar to these examples but rather a general principle that applied to wide classes of optimization problems. By the 1920s techniques had been developed for obtaining upper and lower bounds on the solutions to optimization problems by finding approximate solutions to the primal and dual problems. Duality as a general idea is described in the book by Courant and Hilbert (1953).

Euler and Lagrange only considered problems with equality constraints, but later authors allowed inequality constraints as well. When specialized to finite-dimensional problems, the optimality condition is referred to as the Karush-Kuhn-Tucker condition. Kuhn and Tucker derived this result in a 1951 paper. It was later discovered that Karush had proven the same result in his master's thesis (1939) at the University of Chicago under the supervision of Bliss. There are two aspects to the result: its treatment of inequality constraints, and the assumption or constraint qualification that was used to prove it. The first idea can be traced to Weierstrass and the second to Mayer (1886), and both are outgrowths of the calculus of variations.

In the 1870s Weierstrass studied the calculus of variations and presented the results of his investigations in lectures. Weierstrass did not publish his work and it only became widely known years later through the writings of those in attendance. According to Bolza (1904), Weierstrass converted the inequality constraint

$$g(y) \leq 0$$

to an equivalent equality constraint

$$g(y) + s^2 = 0$$

using a squared slack variable s. This technique is described in many sources from 1900 onward. Bolza later became a professor at the University of Chicago, establishing a connection from Weierstrass to Bliss to Karush. Karush used this technique in his thesis.

The constraint qualification used by Karush, Kuhn and Tucker relates feasible arcs (paths of feasible points leading to the solution) and the gradients of the constraints at the solution. This same condition was used by Mayer (1886), although applied to a calculus of variations problem with equality constraints, and then in a chain of papers by various authors (including Bliss) leading to Karush's thesis. In these papers it is called a normality condition, and it is equivalent to requiring that the matrix of constraint gradients at the solution be of full rank. The implicit function theorem can be used to relate this to the condition on feasible arcs, an observation that is explicit in Mayer's work.

Concluding Remarks

The calculus of variations has influenced many areas of applied mathematics. It is a technical tool for solving optimization problems whose parameters are functions, and in this way it continues to be used in optimal control. It was the setting for the development of the most important concepts in optimization, such as duality and the treatment of constraints. And, when coupled with the Principle of Least Action, it was the vehicle for deriving the fundamental laws of physics.

See

- ▶ Control Theory
- ▶ Lagrange Multipliers
- ▶ Linear Programming
- ▶ Nonlinear Programming

References

Bliss, G. A. (1925). *Calculus of variations*. Chicago: Open Court.

Bolza, O. (1904). *Lectures on the calculus of variations*. Chicago: University of Chicago Press.

Courant, R., & Hilbert, D. (1953). *Methods of mathematical physics* (Vol. I). New York: Interscience.

Dacorogna, B. (2004). *Introduction to the calculus of variations*. London: Imperial College Press.

Goldstine, H. H. (1980). *A history of the calculus of variations from the 17th through the 19th century*. New York: Springer-Verlag.

Gregory, J., & Lin, C. (1992). *Constrained optimization in the calculus of variations and optimal control theory*. New York: Van Nostrand Reinhold.
Hestenes, M. R. (1966). *Calculus of variations and optimal control theory*. New York: John Wiley.
Kuhn, H. W. (1991). Nonlinear programming: A historical note. In J. K. Lenstra, A. H. G. Rinnooy Kan, & A. Schrijver (Eds.), *History of mathematical programming* (pp. 82–96). Amsterdam: North-Holland.
Lagrange, J. L. (1888–1889). *Oeuvres de Lagrange* (Vols. XI and XII). Paris: Gauthier-Villars.
Mayer, A. (1886). Begründung der Lagrange'schen Multiplicatorenmethode in der Variationsrechnung. *Mathematische Annalen, 26*, 74–82.

Call and Contact Centers

Vijay Mehrotra[1], Thomas A. Grossman[1] and Douglas A. Samuelson[2]
[1]University of San Francisco, San Francisco, CA, USA
[2]Infologix, Inc., Annandale, VA, USA

Introduction

All companies have direct and indirect means of contacting customers, potential customers, or other clients. The basic ways include postal mail, email, and, of course, the telephone. Of special importance and interest is the ability of a company's representatives (agents) to talk with call-in clients or called parties on a large scale, that is, via call centers. Call centers are an important channel for businesses to interact with customers and stakeholders. Such centers generate large transaction volumes and can have a significant impact on client attitudes towards a company and its products. Examples include commercial software support, outbound sales prospecting, customer service, internal company help desk services, municipal information dissemination, emergency services dispatch, and financial transaction processing.

Many call centers have expanded to become contact centers that communicate with clients and called parties through a variety of means such as voice calls, planned callbacks (sometimes through virtual queueing), voice mail, cellular text messaging, and email. Like call centers, these types of contact centers are used by organizations to provide a wide variety of services.

Historically, A. K. Erlang, by his paper, "On the rational determination of the number of circuits," written in 1924 and first published in Brockmeyer et al. (1948), is considered the founder of call center analysis. Call and contact centers (hereafter collectively referred to as centers) are a large global industry. In 2008, the U.S. had an estimated 47,000 centers and 2.7 million agents; Europe, the Middle East, and Africa had 45,000 centers and 2.1 million agents; and Canada and Latin America had 35,000 centers and 730,000 agents. Since then, the industry continued to grow rapidly worldwide.

Centers can be inbound, outbound, or blended. Inbound centers receive calls and other contacts from clients. Outbound centers, which normally rely on voice, generate calls that are usually for telemarketing or collections. Blended centers do both and typically deploy agents who perform outbound work when inbound arrival rates are low. Inbound centers provide staff based on advance predictions of call rates and duration; poor predictions can cause serious degradation in performance. Thus, inbound centers need high-quality forecasts of arrival rates and service times that are random and non-stationary.

Outbound center managers have the luxury of choosing when to initiate contact and closely map their actions to the number of agents on duty. Computers are used to generate outbound calls and are programmed to pace calls such that a called party picks up the phone just as an agent ends a call. Hence, outbound centers need predictions of the expected length of contact, and the time interval between a computer-placed call and when the called party answers the phone. Otherwise, the system generates a nuisance call by abandoning the call when the called party answers, or the called party hangs up because there is no one on the line. U.S. law prescribes penalties for generating large numbers of calls abandoned by the system.

Ongoing improvements in information technology and reductions in telecommunications costs allow multiple physical locations to be managed as a single very large virtual center, thus enhancing pooling effects. Contacts can be given complex routings depending on the client's identity, product, need, or service history. Information can be obtained from clients via interactive voice response. Centers maybe off-shored to locations with lower labor costs; they can be run in-house or outsourced to a contractor.

Contact center business and operations issues are discussed in the following surveys and related studies: Aksin et al. (2007a), reviews opportunities for OR to improve practice; Gans et al. (2003) cites 164 papers associated with call centers; an expanded Web-based bibliography by Mandelbaum includes over 450 papers, as well as case studies; and Koole and Mandelbaum (2002) survey queuing models, while L'Ecuyer (2006) surveys optimization models; multi-skill centers are reviewed by Koole and Pot (2006) and Aksin et al. (2007a, b). OR/MS models applied within this domain tend to not consider human resources issues, although these issues are reviewed by Holman (2005) and Aksin et al. (2007b).

Much of the OR/MS knowledge and research related to centers is proprietary and unpublished. Trade magazines and patent filings can be important sources of information.

Inbound Systems

Inbound centers serve clients who initiate contact with an organization to receive service. Such centers need to react to calls that arrive randomly. Typically, after initiating contact, the client is connected to an agent or placed into a queue for later connection. Upon connection, the client receives service for some random amount of time. There can be other outcomes, such as when all incoming phone lines are in use and the call is blocked (the client is given a busy signal) or the client may decide to abandon the call. In some centers, a client may be connected to another agent with different skills, or wait in that agent's queue. The client might call back or otherwise reinitiate contact if the issue was not satisfactorily resolved. See Cleveland (2006) for further details on how inbound call centers operate.

Managers of inbound call centers seek to provide high-quality client service while keeping costs under control. Cost is straightforward, with the largest expense being labor. Cost performance is generally measured using labor cost, for which agent utilization (percent of time an agent is engaged with clients) serving as a proxy. Labor costs typically constitute 60% to 80% of a call center's operating expense. Telephone (or for contact centers, the Internet) costs have been a concern when queues or service times were high. These pressures are diminishing in most developed countries due to rapidly declining telephone and Internet rates. An issue of labor costs is that centers with excessive agent utilization or low pay may experience high employee turnover with concomitant expenses for recruitment and training.

In contrast to labor cost, service quality is a more complex measure. Service quality can include issues such as agent training and professionalism, and the ability to resolve client problems on the first call. Operationally, service quality is often measured by some function of the amount of time a client waits prior to talking to an agent. As the waiting time experienced by an individual client is a random variable, performance measures are typically some function of the waiting time distribution. The two most common measures are the average client waiting time (ASA) and percentage of calls answered within a designated time, the service level (SL).

When clients hang up before talking to an agent, they are said to abandon the queue, an example of queueing's concept of reneging. An important measure of interest is the client abandonment rate (CAR). A client who abandons the queue is presumably dissatisfied, an undesirable event, but this reduces the waiting time for subsequent clients in the queue, which enhances the center's waiting time measures (Mandelbaum and Zeltyn 2007).

Many call centers are able to track whether an issue is successfully resolved by the first phone call or requires one or more follow-up calls. The metric associated with this data is known as the first call resolution (FCR) rate.

Creating Agent Schedules

Managers schedule agents into time blocks that are typically 15 minutes to one hour in length. A 24-hour center with 15-minute time blocks has 96 time blocks each day. They try to keep staff costs low while having enough agents on duty to meet quality targets. The first OR/MS application of this tradeoff was for toll booth staffing, Edie (1954).

The process for scheduling agents is typically performed in five steps, as follows:

Step 1: Forecast call arrivals. Centers use standard statistical and forecasting techniques such as regression, exponential smoothing and its variants, and the time series models of ARIMA. Difficulties in making accurate forecasts are caused by noisy data due to small time blocks. In centers that have

complex routings with multiple queues, each queue requires a forecast. Further, call patterns can be complex in that those that are blocked or abandoned can affect future arrivals, as can call backs caused by inadequate problem resolution. Gans et al. (2003) discusses opportunities to improve center forecasts.

Step 2: Develop an estimate of operational performance measures. To plan effectively, managers must be able to estimate the impact of their decisions on operational performance measures to trade-off cost and client experience. Cost measures typically include total labor cost and average agent utilization. Client experience measures typically include ASA, SL and CAR. These measures can be estimated using analytic models or discrete-event simulation. The most common method is to apply the Erlang C formula (for determining the waiting probability in a queue) to produce estimates for ASA and SL. Arrival rates are assumed to be homogeneous and come from the Step 1 forecast.

Advanced call centers are characterized by complex routing arrangements that shunt clients among multiple queues. Skill-based routing sends calls initially to a queue that processes the most basic client inquiries and routes more challenging calls to better trained and more highly paid agents in a different queue. The task of estimating operational performance measures is thus complicated because arrivals in later queues depend upon performance, including other agent pools. Discrete-event simulation is the tool of choice in these circumstances; see Mehrotra and Fama (2003).

Step 3: Determine the number the number of agents to assign. The manager must set (or staff) the number of agents to be on duty in each time block. This is an aggregate decision, and does not consider the identities or work schedules of individual agents, which are addressed in Steps 4 and 5. Typically, the manager assigns agents to a time block to minimize total agents, while meeting a target performance measure, usually ASA or SL.

Step 4: Develop multi-time block shifts. The manager must take the number of agents assigned to each time block in Step 3 and back out a set of individual, multi-time block shifts that, in aggregate, sum up to the number of assigned agents in each time block, while honoring work rules, contract requirements, and labor laws. This can produce an infeasible or a difficult-to-apply solution, and approximations with high cost are often required.

Step 5: Assign individual agents to each shift. The manager makes final shift schedules, that is, rosters of named agents. This creates challenges of managing total hours worked per day and week to conform to labor laws, as well as managing personal preferences for work schedules and days off.

Integration of the Five Steps. The agent scheduling process is often executed step-by-step. There are obvious interactions across steps with opportunities to integrate the steps (Aksin et al., 2007a, b). Avramidis et al. (2009) show that integrating the staffing and scheduling steps in a center with skill-based routing can lead to better results. Cezik and L'Ecuyer (2008) used linear programming combined with simulation for a center with skills-based routing.

Outbound Systems

In outbound systems, a computer automatically calls designated parties from a given list. The computer recognizes and processes busy signals, no-answers, and telephone company messages. Answered calls are routed to call center agents. Typically, the computer predicts when agents will become free and dials in anticipation of agent availability, thereby reducing the time agents wait between connections.

A key analytical challenge is to determine the pacing, that is, when to dial the next call. If the pacing is too slow, agent time is wasted. If the pacing is too fast, a called party answers when no agent is available, creating a nuisance for the called party (who usually hangs up), a wasted expense for the system. Research in this area is mostly proprietary; there is scant research literature. The solution resulted in the first U.S. patent based on queueing theory (Samuelson 1989). This method estimates service durations, times from dialing to answer, and proportions of dial attempts that result in answers, and uses these statistics, updated frequently, to synchronize dialing attempts to finish shortly after predicted agent service completions (Samuelson 1999). Other patents, such as David (1997), expand and extend this method.

Other proprietary procedures establish call centers based on cloud computing. This approach drastically reduces facility costs, as agents can work from home. It also presents a new version of the predictive dialing problem: the situation is more complex and more subject to quick changes, but the huge computational resources readily available can be employed to do massive parallel simulations in real time to compute the required predictive parameters (Kaiser-Nyman et al. 2011).

Blended Systems

Blended call centers allow agents to be switched in real time between inbound and outbound calls. Bhulai and Koole (2003) discuss a queueing model which yields a threshold policy for assigning agents to outbound calls. Deslauriers et al. (2007) provide a set of Markov chain models for a call center where outbound agents can be diverted to serve inbound calls. Call center managers believe that frequent switching between inbound and outbound calls degrades agent performance for both types of calls, and common practice is to make reassignments for blocks of time rather than call by call. This added constraint makes the performance modeling and scheduling problem substantially more difficult.

Operational Trends and Research Opportunities

Forecasting and Workload Requirements

Some traditional call center assumptions have been questioned by OR/MS researchers, e.g. (Aksin et al. 2007a). One area concerns replacing the standard point-forecast of arrival rates for a short-time block with a stochastic forecast. It is possible to relax the assumption of independent time block call arrivals and model correlation of arrivals across time blocks. More general assumptions on arrival rates can affect the scheduling and rostering problems, see Steckley et al. (2009), Robbins and Harrison (2010), and Gans et al. (2009). Bassamboo et al. (2009) proposed a methodology for capacity planning and dynamic system control in the presence of random arrival rates and multiple inbound call types.

Also, significant research attention has been paid to developing and applying advanced statistical techniques to call center arrival forecasts, with many of these approaches being used to generate not only a point estimate, but also distributional forecasts. Channouf et al. (2007) tested different forecasting models for an emergency medical system. Weinberg et al. (2007) provides a model for forecasting for the short-time blocks commonly found in practice. Avramidis et al. (2004) examined how call volumes correlate across time blocks within a day, and suggested that call arrival data from early in the day can be used to update forecasts for later in the day. Shen and Huang (2008) developed a singular value decomposition model for updating same-day forecasts based on early data, and show it to be superior to benchmarks commonly used in practice. Soyer and Tarimcilar (2008) applied a Bayesian approach for modeling and analyzing call center arrival data.. Aldor-Noiman et al. (2009) developed a Gaussian mixed-model framework that allows for exogenous variables to model the contribution of specific events to forecasted call volumes.

Scheduling

Call center workforce scheduling is more complex than shift scheduling for many other service delivery organizations, such as hospitals (nurses) or transportation (bus drivers), because of the possible need to shift workload quickly to match skills required by the incoming customer to skills of the available agents. That is, call center workforce scheduling decisions are dynamic. As updates on call arrivals and agent availability become available, short-term forecasts and agent schedules can be adjusted. Mehrotra et al. (2010) developed a methodology for intra-day forecast and schedule updating, while Gans et al. (2009) suggest a stochastic-programming model with recourse to account for both random arrival rates and intra-day schedule updates.

Resource Acquisition

Call center resource acquisition is an important and continuing area of interest. Additional research is needed on long-term forecasting, personnel planning for general multi-skill call centers in the presence of both learning and attrition (Ryder et al. 2008), and for complex networks of service providers (Aksin et al. 2007a, Section 2.2). Companies routinely outsource

call center operations to third party service providers. Ren and Zhou (2008), and Milner and Olson (2008) explore issues associated with establishing and managing these relationships.

Use of Real-Time Data

Call center models generally have had to assume that agents' service times are identically distributed for a given class of client. This was due to limited computational resources that necessitated updating parameters at intervals (ten minutes is common) much longer than the typical call duration. Outbound models have similar limitations with respect to the proportion of called parties who answer. Actual call center data, however, indicate persistent differences among agent service times, even for probabilistically identical clients, and runs of high or low proportions of good contacts and of live answers. Therefore, using real-time data to adjust call center operations could produce improvement in performance, although call center managers are quick to point out that efficiency must be balanced against robustness. Kaiser-Nyman et al. (2011) report significant performance improvement from methods that do use the real-time capabilities of cloud computing.

Performance of Outbound Systems

Despite the amount of time that has passed since the solution of the basic predictive dialing problem, many interesting unsolved problems remain, as research has largely concentrated on inbound systems. For systems running multiple simultaneous outbound campaigns and applying multiple predictive dialing systems in parallel, a common tactic is to switch agents from one campaign to another as answer rates change This impairs productivity if agents are switched too abruptly or too often, hence pacing that takes human factors into account would be valuable.

Balancing the utilization of agents in blended systems, where agents could serve both inbound and outbound parties, is generally done with heuristics that tend to under-optimize productivity to ensure that high-priority, high-value inbound calls always get handled quickly. Again, the available heuristics under-optimize and overlook significant human factors.

In some outbound systems, the protocol is to have the first conversation introduce playing a recorded message to a called party who agrees to listen to it, then to have an agent (not necessarily the same one who had the first conversation) conduct a second live conversation. If agents can be switched between first and second conversations quickly, there is an opportunity for greater productivity with a predictive dialing method, but the synchronization problem is quite complex. Also, again, human factors considerations may add additional constraints.

Research Data

Available operational data tend to be aggregated into time-based averages, which is problematic from a queueing science perspective. Fortunately, the Web-based DataMOCCA Project provides a clean source of high-granularity, call-based client call data from several sources that can be used to test proposed center advances in a research environment.

Call Routing

Skill-based routing, in which different agents are capable of handling different subsets of calls in an environment with multiple call types, is a major trend in the call center industry (L'Ecuyer 2006). These systems route clients to different agents depending on their needs and support the creation of a hierarchy of agents with highly skilled personnel handling only the most challenging calls. There is an opportunity for research regarding design and appropriate performance measures in such systems, and in the dependency and interaction among staffing, scheduling, and routing. When there are multiple types of calls and multiple types of agents, performance modeling, staffing, scheduling, and rostering problems all become significantly more complex, which leads to many interesting and important research problems, see Fukunaga et al. (2002) and Avramidis et al. (2009, 2010).

Concluding Remarks

Call and contact center managers do not view models and algorithms as intrinsically appealing. Successful OR solutions need to integrate tightly with a center's existing software systems for data collection, analysis, decision support, and schedule creation. The OR value proposition can extend beyond just cost savings. Managers value OR professionals who can reduce future call volumes using process management

techniques, such as call content analysis, that collects structured data on caller issues and performs Pareto analysis to direct the organization to improve product quality, user manuals, and agent training (Mehrotra and Grossman 2009). Managers also value reduced service time, reduced labor headaches from improved scheduling, and, in some centers, sales made or clients retained. In addition to the technical aspects of the subject, there is room for more study of how to assess and address the business needs.

See

- ▶ Communications Networks
- ▶ Forecasting
- ▶ Manpower Planning
- ▶ Networks of Queues
- ▶ Queueing Theory
- ▶ Simulation of Stochastic Discrete-Event Systems

References

Aksin, Z., Armony, M., & Mehrotra, V. (2007). The modern call center: A multi-disciplinary perspective on operations management research. *Prodution and Operations Management, 16*(6), 665–688.

Aksin, Z., Karaesmen, F., & Ormeci, E. (2007). A review of workforce cross-training in call centers from an operations management perspective. In D. Nembhard (Ed.), *Workforce cross training handbook*. Boca Raton, FL: CRC Press.

Aldor-Noiman, S., Feigin, P., & Mandelbaum, A. (2009). Workload forecasting for a call center: Methodology and a case study. *The Annals of Applied Statistics, 3*, 1403–1447.

Avramidis, A., Chan, W., Gendreau, M., L'Ecuyer, P., & Pisacane, O. (2010). Optimizing daily agent scheduling in a multiskill call center. *European Journal of Operational Research, 200*, 822–832.

Avramidis, A., Chan, W., & L'Ecuyer, P. (2009). Staffing multi-skill call centers via search methods and a performance approximation. *IIE Transactions, 41*, 483–497.

Avramidis, A., Deslauriers, A., & L'Ecuyer, P. (2004). Modeling daily arrivals to a telephone call center. *Management Science, 50*(7), 896–908.

Bassamboo, A., Harrison, J. M., & Zeevi, A. (2009). Pointwise stationary fluid models for stochastic processing networks. *Manufacturing and Service Operations Management, 11*(1), 70–89.

Bhulai, S., & Koole, G. (2003). A queueing model for call blending in call centers. *IEEE Transactions on Automatic Control, 48*(8), 1434–1438.

Brockmeyer, E., Halstrom, H., & Jensen, A. (Eds.). (1948). *The life and works of A. K. Erlang*. Copenhagen: The Copenhagen Telephone Company.

Cezik, M., & L'Ecuyer, P. (2008). Staffing multiskill call centers via linear programming and simulation. *Management Science, 54*(2), 310–323.

Channouf, N., L'Ecuyer, P., Ingolfsson, A., et al. (2007). The application of forecasting techniques to modeling emergency medical system calls in Calgary, Alberta. *Health Care Management Science, 10*(1), 25–45.

Cleveland, B. (2006). *Call center management on fast forward: Succeeding in today's dynamic inbound environment*. Colorado Springs (CO): ICMI Press.

David, J. (1997). Outbound call pacing method which statistically matches the number of calls dialed to the number of available operators. *U.S. Patent 5,640,445*. Washington, DC: U.S. Patent Office.

Deslauriers, A., L'Ecuyer, P., Pichitlamken, J., et al. (2007). Markov chain models of a telephone call center with call blending. *Computers and Operations Research, 34*(6), 1616–1645.

Edie, L. (1954). Traffic delays at toll booths. *Journal of the Operations Research Society of America, 2*(2), 107–138.

Fukunaga, A., Hamilton, E., Fama, J., Andre, D., Matan, O., & Nourbakhsh, I. (2002). Staff scheduling for inbound call centers and customer contact centers. In R. Dechter., M. Kearns, & R. Sutton (Eds.). 18th National Conference on Artificial Intelligence; July 28 – August 1, 2002, Edmonton, Alberta. Menlo Park (CA): American Association for Artificial Intelligence, 822–829.

Gans, N., Koole, G., & Mandelbaum, A. (2003). Telephone call centers: Tutorial, review and research prospects. *Manufacturing and Service Operations Management, 5*(2), 79–141.

Gans, N., Shen, H., Zhou, Y.-P., et al. (2009). *Parametric stochastic programming for call-center workforce scheduling*. Philadelphia: Wharton School, University of Pennsylvania.

Holman, D. (2005). Call centers. In D. Holman, T. D. Wall, C. W. Clegg, et al. (Eds.), *The essential of the new workplace: A guide to the human impact of modern work practices*. New York: John Wiley & Sons.

Ingolfsson, A., Akhmetshina, E., Budge, S., et al. (2007). A survey and experimental comparison of service level approximation methods for non-stationary M/M/s queueing systems. *INFORMS Journal on Computing, 19*, 201–214.

Kaiser-Nyman, M., Samuelson, D. A., & Swieskowski, B. (2011). Predictive dialing system. *U.S. Provisional Patent No. 61/564,756*. Washington, DC: U.S. Patent Office.

Koole, G., & Mandelbaum, A. (2002). Queueing models of call centers: An introduction. *Annals of Operations Research, 113*(1–4), 41.

Koole, G., & Pot, A. (2006). *An overview of routing and staffing algorithms in multi-skill customer contact centers*. Working paper, Amsterdam: Department of Mathematics, Vrije Universiteit Amsterdam.

L'Ecuyer, P. (2006). Modeling and optimization problems in contact centers. *Proceedings of the 3rd International Conference on the Quantitative Evaluation of Systems (QEST 2006)*. September 11–14, 2006. University of California, Riverside. Washington, DC: IEEE Computing Society, 145–154.

Mandelbaum, A., & Zeltyn, S. (2007). Service engineering in action: The Palm/Erlang-A queue, with applications to call centers. In D. Spath & K.-P. Fähnrich (Eds.), *Advances in services innovations* (pp. 17–48). Berlin-Heidelberg: Springer.
Mehrotra, V., & Fama, J. (2003). Call center simulation modeling: Methods, challenges, and opportunities. In S. Chick., P. J. Sánchez., D. Ferrin, & D. J. Morrice (Eds.), *Proceedings of the 2003 Winter Simulation Conference*, pp. 135–143.
Mehrotra, V., & Grossman, T. A. (2009). OR process skills transform an out of control call center into a strategic asset. *Interfaces, 39*(4), 346–352.
Mehrotra, V., Ozluk, O., & Saltzman, R. (2010). Intelligent procedures for intra-day updating of call center agent schedules. *Production and Operations Management, 19*(3), 353–367.
Milner, J., & Olson, T. (2008). Service-level agreements in call centers: Perils and prescriptions. *Management Science, 54*(2), 238–252.
Ren, Z., & Zhou, Y. (2008). Call center outsourcing: Coordinating staffing level and service quality. *Management Science, 254*(2), 369–383.
Robbins, T., & Harrison, T. (2010). A stochastic programming model for scheduling call centers with global service level agreements. *European Journal of Operational Research, 207*(3), 1608–1619.
Ryder, G., Ross, K., & Musacchio, J. (2008). Optimal service policies under learning effects. *International Journal of Services and Operations Management, 4*(6), 631–651.
Samuelson, D. (1989). System for regulating arrivals of customers to servers. *U.S. Patent 4,858,120*. Washington, DC: U.S. Patent Office.
Samuelson, D. (1999). Call attempt pacing for outbound telephone dialing systems. *Interfaces, 29*(5), 66–81.
Shen, H., & Huang, J. (2008). Interday forecasting and intraday updating of call center arrivals. *Manufacturing and Service Operations Management, 10*(3), 391–410.
Soyer, R., & Tarimcilar, M. (2008). Modeling and analysis of call center arrival data: A Bayesian approach. *Management Science, 54*, 266–278.
Steckley, S., Henderson, S., & Mehrotra, V. (2009). Forecast errors in service systems. *Probability in the Engineering and Informational Sciences, 23*(2), 305–332.
Taylor, J. (2008). A comparison of univariate time series methods for forecasting intraday arrivals at a call center. *Management Science, 54*(2), 253–265.
Weinberg, J., Brown, L., & Stroud, J. (2007). Bayesian forecasting of an inhomogeneous Poisson process with applications to call center data. *Journal of the American Statistical Association, 102*, 1185–1199.

Call Priorities

A strategy for handling calls with varying degrees of urgency. Many emergency services have instituted formal procedures for responding differently (e.g., with and without flashing lights and sirens) to calls depending upon priority level.

See

► Emergency Services

Candidate Rules

A group of rules that the inference engine has determined to be of immediate relevance at the present juncture in a reasoning process. These rules will be considered according to a particular selection order and subject to a prescribed degree of rigor.

See

► Artificial Intelligence

Capacitated Transportation Problem

A version of the transportation problem in which upper bounds are imposed on some or all of the flows between origins and destinations.

See

► Transportation Problem

Capital Budgeting

Reuven R. Levary
Saint Louis University, St. Louis, MO, USA

Introduction

The desired end result of the capital budgeting process is the selection of an optimal portfolio of investments

from a set of alternative investment proposals. An optimal portfolio of investments is defined as the set of investments that makes the greatest possible contribution to the achievement of the organization's goals, given the organization's constraints. The constraints faced by a corporation in the capital budgeting process can include limited supplies of capital or other resources as well as dependencies between investment proposals. A dependency occurs if two projects are mutually exclusive, acceptance of one requires rejection of the other, or if one project can be accepted only if another is accepted. Assuming that the organizational goals and constraints can be formulated as linear functions, the optimal set of capital investments can be found using linear programming (LP).

Capital Budgeting Under Capital Rationing

Capital rationing is a constrained capital budgeting problem in which the amount of capital available for investment is limited.

Pure Capital Rationing, with No Lending or Borrowing Allowed — Consider a firm that has an opportunity to invest in several independent projects. It is assumed that both the future cash flows associated with each project and the firm's future cost of capital can be forecast. These forecasts enable calculation of the net present value for each project, assuming that the firm expects to be affiliated with the projects for a period of N years. It is also assumed that the firm has a given fixed budget for funding the projects for each of the N years, with both the budget and the cost of capital in future periods being unaffected by investments made in previous periods. Finally, it is assumed that any portion of the budget not used in one year cannot be carried over to future years.

The basic model for capital budgeting under pure capital rationing is as follows:

$$\text{maximize} \quad \sum_{i=1}^{M} P_i x_i \tag{1}$$

$$\text{subject to} - \sum_{i=1}^{M} f_{it} x_i \leq b_i \quad \text{for } t = 1, 2, \ldots, N \tag{2}$$

$$0 \leq x_i \leq 1 \quad \text{for i} = 1, 2, \ldots, M \tag{3}$$

where P_i is the net present value for the ith project (calculated based on forecasts of future cash flows), f_{it} is the expected cash flow for project i during year t (cash flow is defined to be positive if it is inflow and negative if it is outflow), b_t is the available budget for year t, M is the number of alternative projects and x_i is the fraction of project i to be funded.

The objective function (1) represents the total expected net present value of the investment proposals that should be funded. Constraints (2) represent restrictions on the available yearly budget. Constraints (3) ensure that no more than one project of a given type will be included in the optimal portfolio. By adding the constraint that x_i be integer for $i = 1, 2, \ldots, M$, the problem becomes an integer program. In this case, no fractional projects will be allowed; a project is either accepted or rejected. Constraints on scarce resources, mutually exclusive projects, and contingent projects can easily be added to the above model when necessary.

Capital Budgeting Where Borrowing and Lending are Allowed — In this model, the amount available for lending in a given year is the "left-over" money for that year. This amount can be carried over to the next year at a given rate of interest r. Consider the case when the interest rate for borrowing, or cost of funds, depends on the amounts borrowed. The cost of borrowing is assumed to have the shape of a step function; that is, the larger the amount borrowed, with limits, the higher the interest rate. Let r_j be the interest rate that applies to borrowing an amount greater than C_{j-1} and less than or equal to C_j. A firm will borrow at interest rate r_j if it exhausts the limits placed on its borrowing at lower interest rates.

If the firm expects to be affiliated with the proposed projects for N years, then the objective is to maximize the total related cash flows at the end of the Nth year, that is, the horizon. Let α_t and β_t be, respectively, the amount lent and the amount borrowed (at interest rate r_j) in year t. Also, let f_{it} be the cash flow in year t resulting from approval of project i. All flows in this model are current values, that is, not present values. Revenues and expenditures are defined, respectively, to be positive and negative cash flows. A given project can generate cash flows after the Nth year as well. Let $\hat{f}_i$ be the present value of total cash flows at the horizon (i.e., year N) that are expected to be generated by project i at years following year N. These flows are

discounted to year N, assuming an interest rate equivalent to the firm's weighted average cost of capital. The model is formulated as follows:

$$\text{maximize} \quad \sum_{i=l}^{M} \hat{f}_i x_i + \alpha N - \sum_{j=l}^{m} \beta_{jN} \tag{4}$$

$$\text{subject to} \quad -\sum_{i=l}^{M} f_{il} x_i + \alpha l - \sum_{j=l}^{m} \beta_{jl} \le b_l \tag{5}$$

$$-\sum_{i=1}^{M} f_{il} x_i - (1+r)\alpha_{l-1} + \alpha_l + \sum_{j=1}^{m} (1+r_j)\beta_{jl-1} - \sum_{j=1}^{m} \beta_{jl} \le b_l \quad \forall \; t = 2,\ 3,\ \ldots,\ N \tag{6}$$

$$\beta_{jt} \le C_{jt} \quad \forall \; t = 1,\ 2,\ \ldots,\ N;\ j = 1,\ 2,\ \ldots,\ m \tag{7}$$

$$0 \le x_i \le 1 \quad \forall \, i = 1,\ 2,\ \ldots,\ M \tag{8}$$

$$\alpha_t,\ \beta_{jt} \ge 0 \quad \forall \; t = 1,\ 2,\ \ldots,\ N;\ j = 1,\ 2,\ \ldots,\ m \tag{9}$$

where m represents the number of different interest rates in the supply of funds schedule. The limit on borrowing during year t, at interest rate r_j, is denoted by C_{jt}. Objective function (4) represents the total flows resulting from the proposed projects at the end of the Nth year. The first component $\sum_{i=1}^{M} \hat{f}_i x_i$ of the objective function represents the present value at the horizon of the cash flows expected to be generated by the projects in years following the horizon year N. The second component $\sum_{j=1}^{M} \beta_{jN}$ is the amount lent minus the amount borrowed during the horizon year N. Inequality (5) and inequalities (6) represent the constraints on the available budget for a given year. The limits on borrowing are represented by constraints (7). This model can be extended by adding constraints on scarce resources and by incorporating mutually exclusive and contingent projects when applicable.

Fractional Projects

All LP models can result in an optimal portfolio of projects composed of fractional projects. Weingartner (1967) showed that the number of fractional projects in the optimal solution set of the basic LP model [described by relations (1)–(3)] cannot exceed the number of time periods for which constraints are imposed. Additional constraints such as mutual exclusion, contingency, and scarce resources can increase the maximum number of fractional projects. Each additional constraint increases the maximum number of fractional projects by one. Weingartner (1967) also showed that the number of fractional projects in the optimal solution of the model where borrowing and lending are allowed is no larger than the number of time periods during which the firm does not lend or borrow money.

Because solutions to LP models can include fractional projects, these models are only an approximation of the exact solution. The exact solution can be obtained by applying integer programming solution procedures. The fractions of mutually exclusive projects, which can be the solution of an LP model, may have a useful interpretation. Fractional projects may suggest the possibility of a partnership. For example, one might interpret the decision to fund the expenses of building a fraction of a shopping center to mean that it would be beneficial for the company to engage in a partnership arrangement.

Dual Linear Programming and Capital Budgeting

Consider the basic model for capital budgeting under pure capital rationing formulated by relations (1)–(3). To evaluate the profitability of various projects, a discount factor must be incorporated into the capital budgeting analysis. Define d_t as the discount factor for period t: $d_t = (1 + r_t)^{-1}$ where r_t is the interest rate at period t. The net present value for project i is

$$P_i = \sum_{t=l}^{N} f_{it}\, d_t. \tag{10}$$

Substitution of Equation (10) into (1) results in the following formulation, called Problem **P**:

$$\text{maximize} \quad Z = \sum_{i=l}^{M} \sum_{t=l}^{N} f_{it}\, d_t\, X_i$$

$$\text{subject to (2) and (3).} \quad (\mathbf{P})$$

Let y_t be the dual variable associated with the budget constraint for year t. The value of y_t at the optimal solution, y_t^*, represents the increase in the total combined net present value of the projects that results from an addition of $1 to the budget for year t.

Assume that V dollars are added to the budget in period t. This results in an increase of the net present value (the objective function) by $v \times y_t^*$. The net present value of v is $v \times d_t$. This implies that the discount factor d_t should be equal to the dual variable y_t^* at the optimal solution (Baumol and Quandt 1965). Problem **P** is called consistent if its optimal solution has the property $d_t = y_t^* \quad \forall t$. A solution to a capital budgeting problem under pure capital rationing where dual variables do not equal the discount factor is not optimal. Therefore, such a problem is inconsistent.

Ananalysis of consistent solutions helps clarify the relationship between Capital budgeting discount factors in discount factors and dual variables, as well as the choice of an objective function. Several properties of consistent solutions were summarized by Freeland and Rosenblatt (1978) and are:

1. The value of the objective function of Problem **P** equals zero if there are no upper bounds on the decision variables (i.e., in the case when the X_i are not restricted to be less than one).
2. When the value of the objective function is zero, the only way to obtain a consistent solution is by having all discount factors equal zero. This is a meaningless situation.
3. To ensure a meaningful consistent solution, the decision variables must have upper bounds. Furthermore, some projects must be fully accepted.
4. For a consistent solution to be meaningful, the optimal value of the objective function must be positive and the budget vector must include both positive and negative components.
5. If unused funds cannot be carried forward, the discount factor in period t may exceed the discount factor in period $t + 1$.

Finding the "Right" Discount Factors

Because different optimal solutions to Problem **P** are obtained for various values of the discount factor, it is necessary to find the right discount factor for the pure capital rationing case before Problem **P** is solved. Freeland and Rosenblatt (1978) reported that most of the proposed iterative procedures for finding the right discount factors described in the literature do not work properly. Problems involved in finding the right discount factors are avoided by using horizon models, such as (4)–(9). Center for Naval Analyses (CNA) origin of The horizon value of the model where borrowing and lending are allowed is $\alpha_N - \sum_{j=1}^{M} \beta_{jN}$ [see relation (4)] when there are no cash flows beyond the horizon. In this case, no discount rate is used in maximizing the horizon value and therefore the problem of finding the right discount factor is irrelevant. In the case where there are cash flows beyond the horizon, management must estimate the respective discount rates using financial and economic forecasting. The calculation of these estimates is external to the LP models used in capital budgeting decisions, and therefore is not linked to the solution procedure of the LP model.

Alternative Capital Budgeting Models

Some capital budgeting problems have multiple objectives. Such problems can be formulated as goal programming problems. In many cases, the values of variables affecting the cash flows of the projects are not known with certainty. Such variables include future interest rates, length of useful economic lives, and salvage values. Computer simulation can be used to handle the uncertainty surrounding capital budgeting decisions (Levary and Seitz 1990). Simulation can also be used to analyze the risk consequences of various capital budgeting alternatives. Decision tree analysis is a widely used method for analyzing risk associated with a single investment alternative (Levary and Seitz 1990). Expected return on investments can be adjusted for risk using the capital asset pricing model (CAPM). CAPM was generalized by Richard (1979) to include environmental uncertainty.

Applications of chance-constrained programming to capital budgeting problems have been reported in the literature. Byrne et al. (1967, 1969) incorporated payback and liquidity constraints into chance-constrained programming models for capital budgeting. The payback is represented in these models in the form of chance-constraints that filter acceptable from unacceptable risks. The liquidity constraints handle risks related to situations such as unplanned demand for cash and unplanned technological breakthroughs. Hillier (1969) formulated the net cash flows in each time period of a capital budgeting model as probabilistic constraints.

The objective function in the model is to maximize the expected utility of the shareholders at the horizon period. Näslund (1966) extended the horizon model [relations (4)–(9)], by including risks. Näslund assumed that the yearly cash flows were independent, normally distributed random variables having known means and standard deviations. He also assumed that no other random variables existed in his model. The adjusted model is a chance-constrained programming model. Näslund developed a deterministic equivalent to his chance-constrained programming model.

Relationships among investments contribute to portfolio risk and can be measured by covariances. Quadratic programming models for capital budgeting can be used in situations where the covariances between returns of various projects can be estimated. Various characteristics of a specific capital budgeting problem, like tax consequences, can be modeled using mathematical programming.

See

- ▶ Chance-Constrained Programming
- ▶ Goal Programming
- ▶ Integer and Combinatorial Optimization
- ▶ Linear Programming
- ▶ Portfolio Theory: Mean-Variance Model
- ▶ Quadratic Programming

References

Baumol, W. J., & Quandt, R. E. (1965). Investment and discount rates under capital rationing — A programming approach. *Economic Journal, 75*(298), 317–329.

Byrne, R., Charnes, A., Cooper, W. W., & Kortanek, K. (1967). A chance-constrained approach to capital budgeting with portfolio type payback and liquidity constraints and horizon posture controls. *Journal of Financial and Quantitative Analysis, 2*(4), 339–364.

Byrne, R. F., Charnes, A., Cooper, W. W., & Kortanek, K. O. (1969). A discrete probability chance-constrained capital budgeting model-I. *Opsearch, 6*(3), 171–198.

Freeland, J. R., & Rosenblatt, M. J. (1978). An analysis of linear programming formulations for the capital rationing problems. *The Engineering Economist, 23*(Fall), 49–61.

Hillier, F. S. (1969). *The evaluation of risky interrelated investments*. Amsterdam: North Holland.

Levary, R. R., & Seitz, N. E. (1990). *Quantitative methods for capital budgeting*. Cincinnati, OH: South-Western Publishing.

Näslund, B. (1966). A model of capital budgeting under risk. *Journal of Business, 39*(2), 257–271.

Richard, S. F. (1979). *A generalized capital asset pricing model* (Studies in the management sciences, Vol. 11, pp. 215–232). Amsterdam: North Holland.

Seitz, N., & Ellison, M. (2005). *Capital budgeting and long-term financing decisions* (4th ed.). Hillsdale, IL: Dryden Press.

Weingartner, H. M. (1967). *Mathematical programming and the analysis of capital budgeting problems*. Chicago: Markham Publishing.

CASE

Computer-aided software-systems engineering.

See

- ▶ Systems Analysis

CDF

Cumulative distribution function.

Center for Naval Analyses

Carl M. Harris
George Mason University, Fairfax, VA, USA

Introduction

In the pre-World War II year of 1940, many scientists believed that organizing the nation's scientific research would strengthen national defense. As a result, the National Defense Research Committee (NDRC) was established by Presidential Executive Order.

The NDRC was placed under the direction of the newly created Office of Scientific Research and Development (OSRD), which reported directly to the president. NDRC's contact with British researchers indicated that studying actual operations was an essential part of any assessment process. Because the need for operations research was particularly pressing

in the area of antisubmarine warfare (ASW), the Navy created the Antisubmarine Warfare Operations Research Group (ASWORG). In 1942, comprising at first fewer than a dozen scientists, it was the first civilian group engaged in military operations research in the United States. The Center for Naval Analyses (CNA) traces its origins to ASWORG.

Today, CNA analysts provide the Navy and Marine Corps with objective studies of a wide variety of operations, systems and programs. Such studies range from the support of training and testing activities to the evaluation of new technologies and alternative force structures for top-level decision-makers. The following short history of CNA recounts the high-lights of its evolution and contribution to national security.

World War II

During the 1940s, the United States was preoccupied first with the war in Europe and then with the war in the Pacific. As soon as the United States entered the war, German submarines began to patrol the U.S. East Coast and Atlantic shipping lanes in earnest. The Navy's immediate focus was on the U-boat threat and the Battle of the Atlantic.

In Britain, Professor P.M.S. Blackett had demonstrated the value of operations research in solving military problems. Captain Wilder Baker, leader of the newly formed U.S. Navy Antisubmarine Warfare Unit in Boston, was inspired by Blackett's paper, "Scientists at the Operation Level" (see *Blackett's later work*, 1962). Baker believed that a cadre of civilian scientists could also help the U.S. Navy. He asked Professor Philip M. Morse of MIT to head such a group. ASWORG was formed in April 1942 with a mission to help defeat the German U-boats. The contract for ASWORG was administered by Columbia University, which already had an existing contract with the NDRC that focused on anti-submarine warfare.

ASWORG set a major precedent when it required its analysts to gather field data firsthand. Sending civilian experts to military commands was a delicate matter. In June 1942, the field program began when an ASWORG analyst assisted the Gulf Sea Frontier Headquarters in Miami. Shortly afterward, several analysts were assigned to the Eastern Sea Frontier in New York. The field analysts quickly became accepted; most of ASWORG's noteworthy work was achieved in the field.

In June 1942, ASWORG was assigned to the Head-quarters of Commander in Chief, U.S. Fleet (CominCh). Admiral Ernest J. King was both CominCh and the Chief of Naval Operations (CNO). The Tenth Fleet was formed in 1943 to consolidate U.S. ASW operations. In July 1943, ASWORG became part of the Tenth Fleet.

In October 1944, because of the decrease in enemy submarine activity and the increase in operations research requirements on subjects other than ASW, ASWORG was transferred from the Tenth Fleet to the Readiness Division of the Headquarters of CominCh. It was also renamed the Operations Research Group (ORG) as its analysis efforts had become more diversified.

By the end of the war, ORG had about 80 scientists whose scope of study was all forms of naval warfare. During most of World War II, about 40% of the group was assigned to various operating commands. These field analysts developed immediate, practical answers to tactical and force allocation questions important to their commands. Concurrently, they fed back practical experiences and understanding to the central Washington group, a practice still continued a half century later.

Among its many World War II contributions, ORG devised more effective escort screening plans; determined the optimum size of convoys; developed ASW tactics, such as optimum patterns and altitudes for flying AWS patrol aircraft; developed counter measures to German acoustic torpedoes and snorkeling U-boats; and contributed to the use of airborne radar.

Post-War Period

In August 1945, Admiral King, in a letter to Secretary of the Navy James V. Forrestal, recommended and requested that ORG be allowed to continue into peacetime at about 25% of its wartime size. Secretary Forrestal gave his approval shortly thereafter.

Both Admiral King and Secretary Forrestal concluded that much of ORG's unique value was due to its ability to provide an independent, scientific viewpoint to a broad range of Navy problems.

Consequently, in extending the service of ORG into peacetime, it was decided that its character could best be preserved by perpetuating the wartime arrangement through a contract with an academic institution. Such a contract was entered into with MIT in November 1945. At that time, ORG was renamed the Operations Evaluation Group (OEG), with Dr. Jacinto Steinhardt as its first director. OEG was to assist the Navy and its research laboratories in analyzing and evaluating new equipment, tactical doctrine and strategic warfare. OEG established a policy that all of its (male) analysts must spend time assigned to fleet operations, a practice that is partially maintained to this day by CNA.

After the war, OEG published several comprehensive reports on important naval operations, which included many new methodologies. Although some were originally classified secret, they later appeared in Morse and Kimball's *Methods of Operations Research*, Bernard Koopman's *Search and Screening*, and Charles Sternhell and Alan Thorndike's *Anti-submarine Warfare in World War II*. Taken together, these reports provided a record of vital lessons learned in World War II, as well as important operations research methods. With the Korean War and the intensification of the Cold War, the role of analysis in defense planning expanded in the 1950s. Once the Soviets had detonated their first thermonuclear device, the United States had to revise its thinking on many critical defense issues. As the consequences of nuclear war loomed and the cost of military preparedness escalated, the government, more than ever, needed reliable scientific information on which to base its strategic decision-making.

Before the Korean War, OEG began a slow but steady buildup. By 1950, the research staff had grown to about 40. As the war began, OEG received requests for analysts from combat commands. These analysts collected data, solved tactical problems and recommended improvements in procedures, improvements that were sometimes used immediately. OEG expended its major efforts on such specific tactical problems as: selection of weapons for naval air attack on tactical targets; scheduling of close air support; analysis of air-to-air combat; naval gunfire in shore bombardment; blockade tactics; and interdiction of land transportation. By the end of the war, OEG had 60 research staff members.

After the war, OEG continued to grow, albeit slowly. Analysts participated with naval forces in all post-Korean crises. The most important changes in the nature of the group's post-Korean activity were the results of major technological advances, particularly in the field of atomic energy and guided missiles. Issues were broadened to include the possible enemy use of nuclear weapons and the effect of U.S. policies and weapon system choices on the nature of wars the United States would have to be prepared to fight. During this period, the Navy also established the Long-Range Studies Project of MIT; it was later renamed the Institute for Naval Studies (INS).

Defense Management

By the 1960s, advances in weapons technology were causing defense costs to rise dramatically, and the increasing tempo of the Vietnam War later in the decade would cause the defense budget to balloon still further. The swearing in of Secretary of Defense Robert S. McNamara in 1961 marked the beginning of a new philosophy of defense management. Emphasis began to be placed on cost as well as effectiveness. McNamara believed that integrated systems analysis throughout the defense establishment was required to achieve a balanced, affordable military structure.

In 1961, MIT established an Economics Division within OEG because the cost of weapon systems was becoming a dominant factor in military decision-making. Until 1961, the Marine Corps had only one OEG analyst. By the early 1960s, however, Marine Corps requirements for operations research had increased substantially. The Marine Corps Section of OEG was established in December 1961.

By 1962, the Secretary of the Navy wanted to consolidate the study efforts of OEG and INS and began to look for a contractor. MIT, which had managed OEG since 1945, declined an invitation to manage this proposed new enterprise. The Navy then selected the Franklin Institute to administer the contract for the new organization. In August 1962, OEG and INS were brought under the common management of a new entity, the Center for Naval Analyses (CNA).

Center for Naval Analyses

Shortly after CNA was formed, OEG (now as a division) again became involved in an actual naval operation. In October 1962, it helped the Office of the Chief of Naval Operations (OPNAV) develop plans for the naval quarantine of Cuba and assessed the effectiveness of surveillance operations.

As combat escalated in Southeast Asia, so did the number of CNA field representatives providing direct support to the naval operating forces. CNA participated in the study of many operations, such as interdiction campaigns in North Vietnam and infiltration rates in South Vietnam. Also, a large data base on war-related activities was being developed and maintained in CNA's Washington office. In August 1967, management of the CNA contract transferred from the Franklin Institute to the University of Rochester.

Because the war in Vietnam was escalating, the Navy needed more combat analysis. As a result, the Southeast Asia Combat Analysis Group (SEACAG) was established within OPNAV. Shortly thereafter, the Southeast Asia Combat Analysis Division (SEA-CAD) was established within OEG. SEACAD's role was to support SEACAG and to increase the amount of war-related analysis that CNA was performing. CNA analyzed various operations of the Southeast Asian conflict, including combat aircraft losses, interdiction, strike warfare and carrier defense, surveillance and naval gunfire support.

In the 1970s, as the war in Vietnam wound down, military budgets, forces and equipment began to deteriorate. To maintain effectiveness in the face of reduced budgets, the Navy increased its emphasis on analysis. As new systems became available, the Navy needed to determine how best to exploit their capabilities. With old systems that were already deployed, the Navy needed to develop tactics that overcame technical shortcoming.

Military Buildup

The 1980s witnessed a major buildup of U.S. forces in response to the growth of Soviet military power during the 1970s. For the Navy, this meant not only more ships and aircraft but also more emphasis on a maritime strategy and on specific concepts of operations for employing the Fleet in a global war. These efforts matured by 1987, just as Gorbachev unleashed the forces that would lead to the razing of the Berlin Wall and, ultimately, the demise of the Soviet Union.

In 1982, CNA began a major study of concepts of operations for employing the Atlantic Fleet in a global war. This work involved issues ranging from Soviet objectives and intentions in a war to actions the Navy could take to counter Soviet strategy, as well as theater-level tactics that would be executable in the face of a concerted Soviet threat. The results of this work were put into practice in 1984 by Commander, Second Fleet, who also added important tactical innovations. The resulting interaction and cooperation of Washington and the Fleet (and of CNA-Washington and the field analysts) set the tone for similar efforts at other fleet commands.

By December 1982, differences concerning the management of CNA had arisen between the Department of the Navy and the University of Rochester. The Secretary of the Navy decided to open the CNA contract to competition, and several universities and nonprofit research organizations responded. In August 1983, the Navy announced that the Hudson Institute had been awarded the contract for the management of CNA, effective October 1983.

New World Order

The 1990s ushered in an entirely new security environment. In light of the collapse of the Soviet Union and the new emphasis on Third World threats, the Navy and Marine Corps are reevaluating their structure. Unlike the threat of the Cold War era, these new threats are smaller and more diffuse. They require smaller units that can operate jointly in distant areas where the United States often has a limited number of forces and restricted access to bases. Developing these types of forces and operations is a continuing theme for defense planning in the 1990s.

During the 1980s, some significant events had solidified CNA's stature in the analytical field. Demands for CNA's analytical assistance had grown, particularly from senior Navy and Marine Corps

leaders. CNA had become more involved in critical issues and issues of concern to top-level decision-makers, and CNA's staff had increased in size and quality to meet those growing demands.

Organizationally, CNA had changed often over the years to meet the demands of a changing world and a changing military environment. In the spring of 1990, CNA's management, the Board of Overseers, the Navy, and the Hudson Institute all agreed that CNA could function as an independent organization. On October 1, 1990, CNA became independent and began operating under a direct contract with the Department of the Navy, ready to help the Navy and Marine Corps cope with the impending changes in national security policy, defense strategy, defense budgets and defense management practices.

After Iraq annexed Kuwait in August 1990, the CNO asked CNA to track and document the events in the Middle East, to analyze activities, and to develop a lessons-learned data base. CNA had up to 20 field representatives providing support to various naval commands in the Middle East, including Commander, U.S. Naval Central Command.

After the Persian Gulf War, CNA was designated the Navy's lead agency for Desert Shield/Storm data collection and analysis. The Navy believed that future force composition, systems design and budget decisions would be shaped by events of the war and the subsequent analysis. CNA led the reconstruction of Desert Shield/Storm and provided the Navy with a 14-volume report. In addition, CNA is continuing its analysis of the war and is archiving all the fleet data for the National Archives.

During Desert Storm, the value of concepts that CNA had analyzed for the Navy and Marine Corps — the Tomahawk cruise missile, the air-cushioned landing craft (LCAC), the maritime prepositioning — became evident. The Tomahawk land-attack missile was one of the high-tech "stars" of the war; the LCAC played an important role in creating fear of an amphibious assault; and maritime prepositioning allowed two brigades of Marines to deploy to the Gulf in record time.

In the 1990s, CNA's most important task was to help the Navy and Marine Corps make the transition to a post-Cold War security environment. To do this, CNA's research program plan emphasized areas of immediate importance to this transition: the new security environment, littoral operations, communications, warfare area adjustments, training and education, investment alternatives, force structure, and economies and efficiencies.

See

- ▶ Field Analysis
- ▶ Military Operations Research
- ▶ Operations Research Office and Research Analysis Corporation
- ▶ RAND Corporation

References

Blackett, P. M. S. (1962). *Studies of war*. London: Oliver and Boyd.

Center for Naval Analyses. (1993). Victory at sea: A brief history of the center for naval analyses. *OR/MS Today, 20*(2), 46–51.

Kreiner, H. W. (1992). *Fields of operations research*. Baltimore: Operations Research Society of America.

Morse, P. M., & Kimball, G. E. (1946). *Methods of operations research, OEG Report 54, Operations Evaluation Group (CNA)*. Washington, DC: U.S. Department of the Navy.

Tidman, K. (1984). *The operations research group*. Annapolis, MD: Naval Institute Press.

Certainty Equivalence

Jeffery L. Guyse
University of California, Irvine, USA

The certainty equivalent of a gamble or lottery is the sum of money for which, in a choice between the money and the gamble, the decision maker is indifferent between the two. Certainty equivalents are used to determine decision makers' attitudes toward risk, which can then be reflected in the shape of their utility functions. Certainty equivalents can also be used to order a set of alternatives. Classic examples of operationalizations of certainty equivalents used in the literature are minimum selling price, maximum buying price, and cash equivalent. Buying and selling prices may be theoretically different though, due to income effects.

By definition, the utility of the certainty equivalent must be equal to the expected utility of the gamble. With this in mind, the relationship between the certainty equivalent (CE) and the expected value (EV) of a gamble can reveal the decision maker's attitude toward risk. If CE < EV, then the individual is said to exhibit a risk-averse attitude. In this case, the difference between the expected value and the certainty equivalent (EV − CE) is known as the "risk premium" that the decision maker is willing to pay in order to avoid the risk associated with the gamble. If CE > EV, a risk-prone (or risk-seeking) attitude is displayed. Finally, if the two values are equal (CE = EV), then the decision maker is risk-neutral. By assessing CEs, decision analysts can calibrate the utility function of the decision maker to reflect risk attitude in the decision process. For a formal discussion, see Keeney and Raiffa (1976).

Certainty equivalents are also used to elicit a preference order on a set of alternatives. It is assumed that the order induced by assigning certainty equivalents reveals the true preference order of the individual. If one alternative has a higher certainty equivalent than a second alternative, one would expect the individual to choose the former over the latter when asked to make a choice between the two. The method by which the certainty equivalents are elicited has been an area of ongoing research. It was once believed that subjects could simply state their certainty equivalent to a gamble, in which case their response is known as a judged certainty equivalent.

Recent empirical studies have provided evidence that the judged certainty equivalent may not necessarily equal the true certainty equivalent elicited through a choice mechanism. Such a violation of procedure invariance is examined in the stream of research on preference reversals. Subjects provide both judged certainty equivalents and then make choices between pairs of gambles. By carefully selecting the gambles, researchers have been able to elicit judged certainty equivalents that produce an ordering on the set, while the same subject's choices results in the reverse ordering (Grether and Plott 1979; Lichtenstein and Slovic 1971, 1973; Lindman 1971; Slovic and Lichtenstein 1983). Such a pair of gambles is:

A:	0.99 Win $4.00 0.01 Win $0	B:	0.25 Win $16.00 0.75 Win $0

The expected values of the two gambles are $3.96 and $4 respectively. A large proportion of people will indicate a preference for gamble A when asked to choose between the two, yet place a higher dollar value on B (Grether and Plott 1979, p. 623).

Work by Tversky, Slovic, and Kahneman (1990) as well as Bostic, Herrnstein, and Luce (1990) has shown that these preference reversals virtually disappear when the certainty equivalents are elicited through a choice mechanism, such as the Parameter Estimation by Sequential Testing (PEST) procedure. For a review of preference reversals, see Tversky and Thaler (1990).

See

- ▶ Decision Analysis
- ▶ Lottery
- ▶ Risk
- ▶ Utility Theory

References

Bostic, R., Herrnstein, R. J., & Luce, R. D. (1990). The effect on the preference-reversal phenomenon of using choice indifference. *Journal of Economic Behavior & Organization, 13*, 193–212.

Grether, D., & Plott, C. (1979). Economic theory of choice and the preference reversal phenomenon. *The American Economic Review, 69*, 623–638.

Keeney, R. L., & Raiffa, H. (Eds.). (1976). *Decisions with multiple objectives: Preference and value trade-offs*. New York: Wiley and Sons.

Lichtenstein, S., & Slovic, P. (1971). Reversals of preference between bids and choices in gambling decisions. *Journal of Experimental Psychology, 89*, 46–55.

Lichtenstein, S., & Slovic, P. (1973). Response-induced reversals of preference in gambling: An extended replication in Las Vegas. *Journal of Experimental Psychology, 101*, 16–20.

Lindman, H. R. (1971). Inconsistent preferences among gambles. *Journal of Experimental Psychology, 89*, 390–397.

Slovic, P., & Lichtenstein, S. (1983). Preference reversals: A broader perspective. *The American Economic Review, 73*, 596–605.

Tversky, A., Slovic, P., & Kahneman, D. (1990). The causes of preference reversals. *The American Economic Review, 80*, 204–217.

Tversky, A., & Thaler, R. (1990). Anomalies: Preference reversals. *The Journal of Economic Perspectives, 4*, 201–211.

Certainty Factor

A numeric measure of the degree of certainty about the goodness, correctness, or likelihood of a variable value, an expression (e.g., premise) value, or conclusion.

See

► Expert Systems

Chain

A chain in a network is a sequence of arcs connecting a designated initial node to a designated terminal node such that the direction (orientation) of flow in the arcs is from the initial node to the terminal node.

See

► Cycle
► Markov Chains
► Path

Chance Constraint

A constraint that restricts the probability of a certain event to a prespecified range of values. Under certain conditions, chance constraints can be incorporated into mathematical-programming problems.

See

► Chance-Constrained Programming
► Linear Programming
► Stochastic Programming

Chance-Constrained Programming

A mathematical-programming problem in which the parameters of the problem are random variables and for which a solution must satisfy the constraints of the problem in a probabilistic sense. Here the usual linear-programming constraints are given as probability statements of the form $\Pr\{\sum_{j=1}^{n} a_{ij} x_j \leq b_i\} \geq \alpha_i$ for $i = 1, \ldots, m$, where the $\{\alpha_i\}$ are given constants between zero and one. Some forms of the chance-constrained programming problem can be transformed to an equivalent linear-programming problem.

See

► Linear Programming
► Stochastic Programming

References

Charnes, A., & Cooper, W. (1959). Chance-constrained programming. *Management Science, 6*, 73–79.
Prékopa, A. (1995). *Stochastic programming*. Dordrecht: Kluwer.

Chaos

A mathematical term describing a situation in which arbitrarily small variations in independent variable values can produce large variations in the dependent variable. The term is most typically used to characterize the behavior of deterministic, nonlinear, differentiable dynamic systems. The term is sometimes used to describe situations in which true mathematical chaos is not present, but where the results are similarly disturbing. The disturbing effect in battle modeling, for example, is the apparent loss of deterministic behavior.

Chapman-Kolmogorov Equations

In a parameter-homogeneous Markov chain $\{X(t)\}$ with state space S, define $p_{ij}(t)$ as the probability that $X(t+s) = j$, given that $X(s) = i$ for $s, t \geq 0$. Then, for all states i, j and index parameters $s, t \geq 0$,

$$p_{ij}(t+s) = \sum_{k \in S} p_{ik}(t) p_{kj}(s)$$

are the Chapman-Kolmogorov equations. There is a comparable definition when the state space is instead continuous.

See

- ▶ Markov Chains
- ▶ Markov Processes

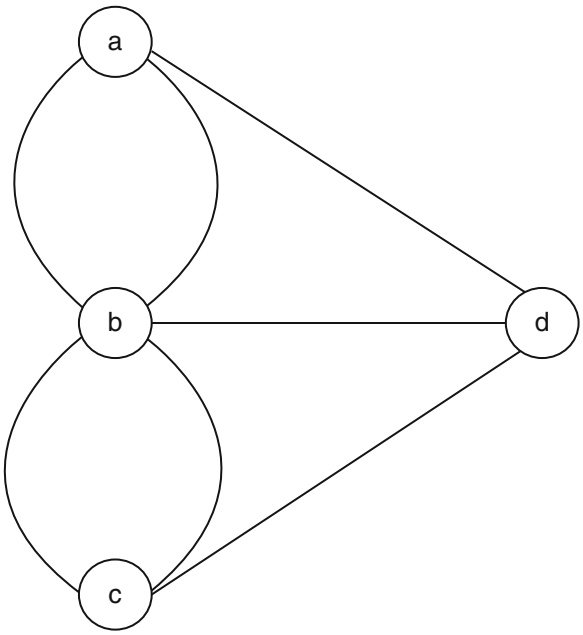
Characteristic Function

For a random variable X, the characteristic function is given by $\phi_X(t) = E[e^{itX}]$, where i denotes the imaginary number $\sqrt{-1}$.

Chinese Postman Problem

William R. Stewart Jr.
College of William and Mary, Williamsburg, VA, USA

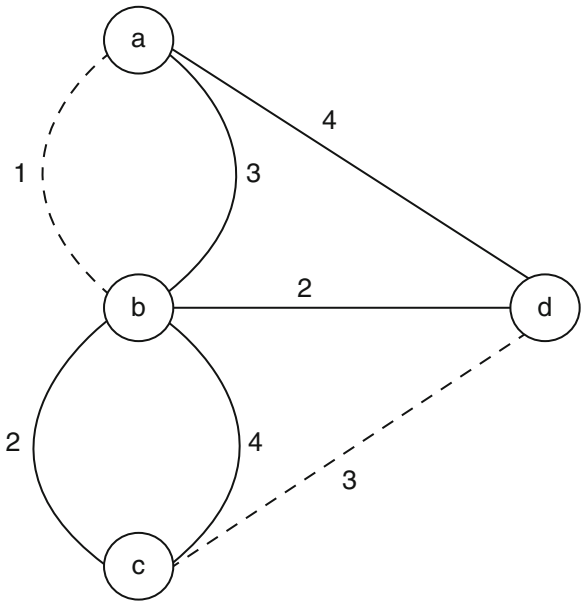
The Chinese Postman Problem acquired its name from the context in which it was first popularly presented. The Chinese mathematician Mei-Ko Kwan (1962) addressed the question of how, given a postal zone with a number of streets that must be served by a postal carrier (postman), does one develop a tour or route that covers every street in the zone and brings the postman back to his point of origin having traveled the minimum possible distance. Researchers who have followed on Kwan's initial work have since referred to this problem as the Chinese Postman Problem or CPP. In general, any problem that requires that all of the edges of a graph (streets, etc.) be traversed (served) at least once while traveling the shortest total distance overall is a CPP. Like its cousin, the traveling salesman problem, that seeks a route of minimum cost that visits every vertex of a graph exactly once before returning to the vertex of origin, the CPP has many real world manifestations, not the least of which is the scheduling of letter carriers. Such problems as street sweeping, snow plowing, garbage collection, meter reading and the inspection of pipes or cables can and have all been treated as CPPs.



Chinese Postman Problem, Fig. 1 A graph of Euler's Konigsberg bridge problem

In the following discussion, the terms tour and cycle will be used interchangeably to refer to a route on a graph that begins and ends at the same vertex and that traverses all of the edges of that graph at least once. Unless otherwise noted, the edges are assumed to be undirected (i.e., they may be traversed in either direction).

The CPP and its many variants have their roots in the origins of mathematical graph theory. The problem of finding a cycle (tour/route) on a graph which traverses all of the edges of that graph and returns to its starting point dates back to the mathematician Leonid Euler and his analysis in 1736 of a popular puzzle of that time, the Königsberg Bridge problem. Euler's problem of traversing all of the bridges of Königsberg and returning to his starting point without retracing his steps is equivalent to asking if there is a tour of the graph shown in Fig. 1 that traverses all of the edges exactly once. Euler showed that such a cycle exists in a graph if and only if each vertex in the graph has an even number of edges connecting to it or, in mathematical terms, each vertex is of even cardinality. This follows logically from the observation that, in a tour that traverses all of the edges exactly once, each vertex must be exited the same number of times it is entered. Tours that traverse each edge of a graph exactly



Chinese Postman Problem, Fig. 2 The Konigsberg bridge problem with edge costs

once are termed Euler cycles or tours, and graphs that contain an Euler cycle are appropriately called Eulerian. When costs are assigned to each of the edges, the problem of finding a minimum cost tour is a CPP.

When a graph is Eulerian, the cost of a tour is just the sum of the costs of all of the edges in the graph, and the solution to the CPP is any Eulerian tour, of which there are usually many. In general, an Eulerian tour can easily be found when one exists. When a graph has more than one odd cardinality vertex (exactly one such vertex is impossible), the CPP is the problem of finding which of the edges must be traversed more than once in order to produce a minimum cost tour. The graph shown in Fig. 1 has four vertices with odd cardinality, and a tour of this graph requires that one or more of the edges be crossed more than once. Figure 2 shows hypothetical costs on each edge, and the dashed lines indicate the edges that must be traversed twice in order to achieve a minimal cost tour. This tour will have a total cost of 23, the cost of crossing each edge once plus the cost of crossing edges (a, b) and (c, d) a second time each.

In mathematical terms, the CPP can be stated as follows: given a graph $G = \{V, E\}$, where V is a set of n vertices, E is a set of edges connecting these vertices, and each edge (i, j) connecting vertices i and j has a nonnegative cost, c_{ij}, find x_{ij}, the number of times that edge (i, j) is to be traversed from i to j so that the total cost of traversing all of the edges in E at least once is a minimum. The sum of x_{ij} and x_{ji} is the total times that the edge between vertices i and j must be traversed in an optimal tour.

$$\text{Minimize} \sum_i \sum_j c_{ij} x_{ij} \tag{1}$$

$$\text{Subject to} \sum_i x_{ik} - \sum_j x_{kj} = 0, \quad \text{for } k = 1, \ldots, n, \tag{2}$$

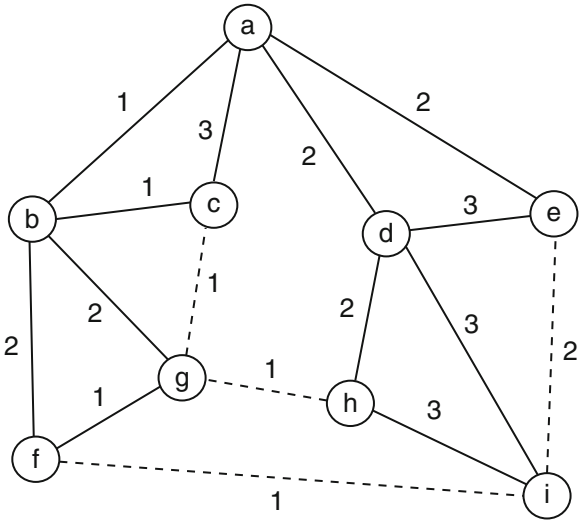
$$x_{ij} + x_{ji} \geq 1, \quad \text{for all } (i,j) \text{ and } (j,i) \in E, \tag{3}$$

$$x_{ij} \geq 0, \quad \text{and integer, for all } (i,j) \in E. \tag{4}$$

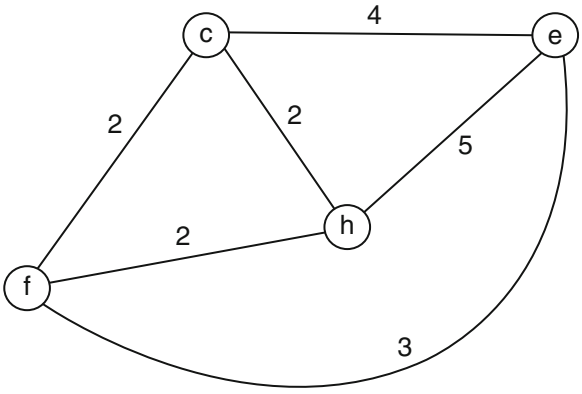
For ease of exposition, this formulation assumes that there is a maximum of one edge between any two vertices. As can be seen in the illustration in Figs. 1 and 2, this may not always be the case. However, cases where there are multiple edges between the same pair of vertices do not complicate the treatment, since those cases can easily be transformed into the form shown in (1)–(4).

As pointed out by Edmonds and Johnson (1973) and Christofides (1973), when there are odd cardinality vertices in the graph, the CPP reduces to the problem of finding a minimum cost matching among the odd cardinality vertices. A minimum cost matching on a graph is a pairing of the vertices on that graph such that each vertex is paired with exactly one other vertex and the total cost of the edges connecting the pairs is a minimum. When no edge exists between a pair of vertices, the cost of pairing them is the cost of the shortest path running between the pair. Replicating the edges that connect each pair of odd cardinality vertices in the minimum matching produces an Eulerian graph (i.e., all vertices now have even cardinality) where the total cost of all the edges, the edges in the original graph plus the edges that have been replicated as a result of the matching, is the cost of the optimal tour of the original graph.

To illustrate the general solution process, Figure 3 presents a graph with four odd cardinality vertices (c, e, f, h). None of the four vertices is directly connected to another of the four. To find the required minimum cost matching requires the construction of the graph G', shown in Fig. 4, which consists of the



Chinese Postman Problem, Fig. 3 The Graph G



Chinese Postman Problem, Fig. 4 The Graph G'

four odd cardinality vertices connected by edges whose costs are the cost of the shortest path between each pair on the original graph. The problem is then to find a minimum cost matching on the graph G'. This matching will determine which edges must be traversed twice to achieve a minimum cost tour on G. A quick inspection of G' shows that the edges (c, h) and (e, f) constitute a minimal matching on G'. The paths $(c\text{-}g\text{-}h)$ and $(e\text{-}i\text{-}f)$ on graph G in Fig. 3 correspond to this matching, and the edges along these paths will be traversed twice each in an optimal tour and are shown as dashed lines in Fig. 3.

Solving the CPP requires two operations, both of which can be performed in polynomial time. A matching of the odd cardinality vertices must be found and the corresponding edges replicated that results in an Eulerian graph. An Eulerian tour of this expanded graph must then be found. The complexity of the CPP is dominated by the complexity of solving the minimum cost matching problem, which can be solved in at most $O(n^3)$ time. Variations of the basic CPP, briefly described below, are generally not as tractable.

Variations of The Chinese Postman Problem

The CPP has many variations that can and do occur on a regular basis. In the CPP, the edges are undirected and they may be traversed in either direction. The most obvious variation of the CPP is the directed postman problem where each of the edges has a direction associated with it. This is often encountered when an edge represents a one way street in a routing problem, or an edge must be traversed twice, once in each direction, as might occur in routing a street sweeper. In this latter case, each street would be represented in the graph by two edges, one in each direction. Like the CPP, the directed postman problem can be solved in polynomial time. In a sense, it is even easier than the CPP since it requires a network flow algorithm rather than a matching algorithm.

When the graph contains a mixture of both directed and undirected edges, the problem of finding a minimum cost tour is called the mixed postman problem. The mixed postman problem has been shown to be NP-hard. The rural postman problem is a variation of the CPP where a subset of the edges in the graph must be traversed. The rural postman problem has been shown to be equivalent to a traveling salesman problem and, as such, it is also an NP-hard problem (see Lawler et al. 1985). Finally, the capacitated Chinese Postman Problem recognizes that each edge may have a nonzero demand for service and that the server (postman) may have a finite capacity for supplying service. In the general case, multiple servers must be assigned to routes such that the demands on all of the edges are met and no server is assigned a route that exceeds his capacity. This then is the problem of partitioning the edges of the graph into subsets and assigning a server (postman) to each subset in such a way that all capacity constraints are met and the total distance covered by all of the servers is a minimum. As with the directed and rural postman problems, the capacitated postman problem has been shown to be NP-hard.

See

- ▶ Combinatorics
- ▶ Computational Complexity
- ▶ Graph Theory
- ▶ Integer and Combinatorial Optimization
- ▶ Matching
- ▶ Network
- ▶ Traveling Salesman Problem
- ▶ Vehicle Routing

References

Christofides, N. (1973). The optimal traversal of a graph. *Omega, 1*, 719–732.

Edmonds, J., & Johnson, E. (1973). Matching, Euler tours, and the Chinese postman problem. *Mathematical Programming, 5*, 88–124.

Kwan, M. K. (1962). Graphic programming using odd or even points. *Chinese Mathematics, 1*, 273–277.

Lawler, E. L., Lenstra, J. K., Rinnooy Kan, A. H. G., & Shmoys, D. B. (Eds.). (1985). *The traveling salesman problem: A guided tour of combinatorial optimization*. Chichester, UK: Wiley.

Choice Strategies

The different approaches people use to combine deterministic information in their mind; sometimes referred to as combination rules.

See

- ▶ Choice Theory
- ▶ Decision Analysis
- ▶ Decision Making and Decision Analysis

Choice Theory

Leonard Adelman
George Mason University, Fairfax, VA, USA

Introduction

There is no one descriptive theory of human choice. Instead, there are different theoretically and empirically-based approaches for describing choice behavior. This article briefly overviews five approaches: bounded rationality, prospect theory, choice strategies, recognition-primed decision making, and image theory. These approaches are descriptive in the sense that they describe certain aspects of how people actually make choices. They contrast with prescriptive approaches, such as decision analysis or other economic-based theories (or models) of choice behavior, which prescribe how one should make decisions, but do not necessarily describe choice behavior.

Bounded Rationality

The concept of bounded rationality is attributed to Nobel laureate Herbert Simon (Simon 1955, 1979; Hogarth 1987), who argued that humans lack both the knowledge and computational skill required to make choices in a manner compatible with economic notions of rational behavior. The rational model's requirements are illustrated by the concept of a payoff matrix, an example of which is presented in Table 1.

The rows of the matrix represent all the different alternatives available to the decision maker for solving a choice problem. The columns represent all of the different states of the world, as defined by future events, that could affect the attractiveness of the alternatives. The $p_1,\ldots, p_k$ values represent the probabilities for each state of the world. The cell entries in the matrix indicate the value or utility of the outcome or payoff for each combination of alternatives and states of the world. Each outcome represents a cumulative payoff comprised of perceived advantages and disadvantages on multiple criteria of varying importance to the decision maker. Finally, the rational decision maker is required to select the alternative that maximizes expected utility, which is calculated for each alternative by multiplying the values for the outcomes by the probabilities for the future states, and then summing the products.

Numerous studies have shown that, unaided, people do not employ the above decision matrix due to the complex, dynamic nature of the environment and to basic human information acquisition and processing limitations. Therefore, how does unaided human choice remain purposeful and reasonable? Simon

Choice Theory, Table 1 The rational economic model's decision making requirements as represented in a payoff matrix

	States of the world			
Alternatives	S_1 (p_1)	S_2 (p_2)	. . .	S_k (p_k)
A	a_1	a_2	. . .	a_k
B	b_1	b_2	. . .	b_k
.	.	.	. . .	.
N	n_1	n_2	. . .	n_k

suggested that people employ three simplification strategies, which result in a bounded rationality. First, people simplify the problem by only considering a small number of alternatives and states of the world at a time. Second, people simplify the problem by setting aspiration (or acceptability) levels on the outcomes. And, third, people choose the first alternative that satisfies the aspiration levels. In other words, people do not optimize (i.e., choose the best of all possible alternatives), but satisfice (i.e., choose the first satisfactory alternative). In this way, people can reduce information acquisition and processing demands and act in a purposeful, reasonable manner.

Prospect Theory

Like Simon's bounded rationality, prospect theory is juxtaposed against expected utility theory. For example, this prospect (or choice) is taken from Kahneman and Tversky (1979):

$$\text{Choice } A: (\$4000 \text{ with } p = .8;\ \$0 \text{ with } p = .2), \text{ or}$$
$$\text{Choice } B: (\$3000 \text{ for sure; that is, } p = 1.0)$$

The majority of participants will select Choice B. Yet, Choice A has the greater expected value; that is, $\$4000 \times .8 = 3200$. Now, consider the following prospect:

$$\text{Choice } C: (-\$4000 \text{ with } p = .8;\ \$0 \text{ with } p = .2), \text{ or}$$
$$\text{Choice } D: (-\$3000 \text{ for sure; that is, } p = 1.0).$$

The only change in the second prospect is that the sign has been reversed so that one is now considering losses, not gains. In this case, however, the majority of the subjects picked Choice C. That is, they would now be willing to take a gamble of losing $4000 with a probability of.8, which has an expected value of losing $3200, instead of taking a sure loss of $3000. Again, they selected the choice with the lower expected value. In addition, they switched from the sure thing to preferring the gamble.

What Kahneman and Tversky (Tversky and Kahneman 1981) have shown is that the way the choice problem is presented (or framed) significantly affects how people evaluate it, such that information that should result in the same choice from the perspective of expected utility theory actually results in different choices. In particular, people perceive outcomes as gains or losses from a reference point rather than from final states (e.g., of wealth), as assumed by economic-based models of choice. The current position is usually considered as the reference point. However, the location of the reference point and, in turn, the coding of outcomes as either gains or losses, can be affected by how the choices are framed.

This framing is particularly important for choice because, as the example presented above indicates, people tend to be risk adverse when considering gains and risk seeking when considering losses, particularly if one of the prospects is certain. Moreover, the value function is steeper for losses than for gains, consistent with the observation that losses loom much larger than gains. For these reasons, many people are willing to gamble to avoid a sure loss, but unwilling to gamble when they have a sure gain, even when both choices have a lower expected value than another choice.

Choice Strategies

Substantial research has focused on describing the different strategies people use to combine information when facing a choice. In contrast to bounded rationality and prospect theory, these strategies are used when people (a) have information on a number of different dimensions (or attributes) describing the alternatives, and (b) do not consider probabilities, either in terms of different states of nature or the reliability (or accuracy) of the information. A representative type of problem is making a purchase decision, such as choosing a car.

The literature (Beach 1990; Hogarth 1987) makes a distinction between two classes of choice strategies: compensatory and noncompensatory. Compensatory

strategies are used when one trades-off (e.g., via relative importance weights) a low value on one attribute for a high value on another. For example, when choosing among cars, one may trade-off gas mileage for comfort. Non-compensatory strategies do not employ trade-offs but, rather, employ thresholds (or cut-offs) that need to be achieved for choice of an alternative. For example, one eliminates all cars that do not get at least 25 miles per gallon, regardless of comfort. Some of the strategies identified in the literature are defined below.

The literature cites three different types of compensatory models:

1. *Linear, additive strategy* — the value of an alternative is equal to the sum of the products, over all the dimensions, of the relative weight times the scale value for the dimension.
2. *Additive difference strategy* — the decision maker evaluates the differences between the alternatives on a dimension by dimension basis, and then sums the weighted differences in order to identify the alternative with the highest value overall.
3. *Ideal point strategy* — is similar to the additive difference model, except the decision maker compares the alternatives against an ideal alternative instead of each other.

The literature cites four different types of noncompensatory strategies:

4. *Dominance strategy* — select the alternative that is at least as attractive as the other alternatives on all the dimensions, but is better than them on at least one dimension. Although the dominance strategy is easier for an unaided decision maker to use, all three compensatory strategies will also identify the dominant alternative. Moreover, the compensatory strategies can be used if there is no dominant alternative; the dominance strategy cannot.
5. *Conjunctive strategy* — select the alternative that best passes some critical threshold on all dimensions. This is the satisficing strategy when one selects the first option that passes a threshold on all dimensions. The conjunctive strategy is often used to reduce the set of alternatives by eliminating all alternatives that fail to pass a threshold on all dimensions.
6. *Lexicographic strategy* — select the alternative that is best on the most important dimension. If two or more alternatives are tied, select among them by choosing the alternative that is best on the second most important dimension, and so on.
7. *Elimination by aspects* —sequentially identify different dimensions, either according to their importance or some more probabilistic scheme. Eliminate all alternatives that fail to pass the threshold or aspect for each dimension until only one alternative is left.

Research (Payne et al. 1993) has shown that people often use multiple strategies when considering choice alternatives. Typically, they use noncompensatory strategies to reduce the number of alternatives and dimensions under consideration. To use a job selection example, a person might first eliminate all alternatives that fail to pass a specific threshold on security, which may no longer be as important when considering the reduced set of alternatives. Then, after the set of alternatives and dimensions have been reduced to a smaller, more manageable set, people often employ a compensatory strategy where they weigh the strengths and weaknesses of the remaining alternatives in order to select the one which best satisfies their values.

Recognition-Primed Decision Making (RPD) and Image Theory

Some descriptive theories have been developed to explain the choice behavior of experts working in naturalistic settings (Zsambok and Klein 1997). These descriptive theories of choice behavior are farther removed from the basic rational economic man model than the three presented thus far. Two are presented here, RPD and image theory, for illustrative purposes.

The RPD model (Klein 1993) emphasizes four critical cognitive processes:

- *Situation recognition* — experienced decision makers know what cues (or indicators) to focus on and, often, simply recognize (or perceive) the situation they are facing through an automatic, feature (or pattern)-matching process, much like perceptual objects in our environment are recognized. People also are quite capable of using explanation-based reasoning to understand a situation when there uncertainties and anomalies in it. In fact, these stories are often, but not always, the causal explanations for the feature-matching process that appears to operate so automatically.

- *Decision option generation* — Once a situation is recognized, decision makers typically generate only one option for consideration, not multiple options.
- *Evaluation through mental simulation* — The initial option tends to be quite good for dealing with the (recognized) situation. Decision makers, however, may evaluate it by mentally simulating the consequences of implementing the option. Although the mental simulation will use intuitive and analytical thought processes, depending on the consequences being evaluated during the simulation, the option will seldom, if ever, be evaluated by a formal analysis on a set of attributes (e.g., by a decision matrix).
- *Use of a decision rule* (emphasizing acceptability, not optimality) — The mental simulation may result in modifications to the proposed option to address problems uncovered during the mental simulation, or even a new option, but the option will be accepted once it is deemed satisfactory; it does not have to be optimal. Thus, RPD explicitly incorporates Simon's (1955) satisficing concept.

Beach (1990, 1993) developed the concept of images to convey the notion that decision makers bring certain knowledge structures to bear on a problem that constrain (or frame) how they evaluate it. In particular, Beach discussed three images: value, trajectory, and strategic, as follows:

- *Value Image* — this is composed of the overriding principles that guide one's behavior or that of one's organization. They "serve as rigid criteria for the rightness or wrongness of any particular decision about a goal or plan" (Beach 1993, p. 151).
- *Trajectory Image* — this consists of previously adopted goals, the timetable for achieving them, and the ideal future once they are achieved.
- *Strategic Image* — this is composed of the plans for achieving the goals in the trajectory image. The plans consist of specific tactics (or actions) for implementing the more abstract plan, and forecasts of what will happen if specific tactics are implemented. These forecasts change in light of new information. "By monitoring these forecasts (or expectations) in relation to the goals on the trajectory image, the decision maker can evaluate his or her progress toward realization of the ideal agendum on the trajectory image" (Beach 1993, p. 152).

In many ways, Beach's image theory is another way for describing the cognitive processes emphasized in Klein's RPD model. Image theory also emphasizes: (a) monitoring behavior; (b) expectations and goals; (c) situation recognition through feature matching and explanation-based reasoning; (d) automatic generation of a decision option to deal with the recognized situation; (e) mental simulation to evaluate it; (f) satisficing; and (g) processes for monitoring and managing the decision process. It is different, however, in its emphasis of three things.

The first difference in emphasis is that the images strongly frame the interpretation of how well the situation is going or even what the problem is, as emphasized in Prospect Theory. (Keeney (1992) also emphasized framing in his approach, called value-focused thinking, but from more of a prescriptive than descriptive perspective.) Second, routine progress decisions are made to compare the current situation and future forecasts with the ideal future. And, third, adoption decisions are routinely made to modify plans, tactics, and expectancies (i.e., the elements of Strategic Image) — and, less frequently, goals, timetables, and ideal future (i.e., the elements of Trajectory Image) — in response to progress decisions. Depending on the situation and person, these adoption decisions are made using one or more of the choice strategies described above. Although frames, progress decisions, and adoption decisions may be concepts that are inherent in Klein's RPD model, they are strongly and explicitly emphasized in Beach's image theory. In addition, Beach's image theory explicitly incorporates the many descriptive choice strategies found in the literature. Thus, image theory integrates many of the concepts in the choice theory literature to describe how people choose to react to changing situations.

Concluding Remarks

In closing, there is a need to emphasize again that there is no one descriptive theory of human choice. Instead, there are different theoretically and empirically-based approaches for describing choice behavior. This article provided brief overviews of five of them: bounded rationality, prospect theory, choice strategies, recognition-primed decision making, and image

theory. These approaches were contrasted with prescriptive approaches, such as decision analysis or other economic-based theories of choice, which prescribe how one should make decisions, but do not necessarily describe choice behavior.

See

- ▶ Decision Analysis
- ▶ Decision Making and Decision Analysis
- ▶ Preference Theory
- ▶ Utility Theory

References

Beach, L. R. (1990). *Image theory: Decision making in personal and organizational contexts*. New York: Wiley.

Beach, L. R. (1993). Image theory: Personal and organizational decisions. In G. Klein, J. Arisen, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods*. Norwood, NJ: Ablex.

Hogarth, R. M. (1987). *Judgment and choice*. New York: Wiley.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision making under risk. *Econometrica, 47*, 263–289.

Keeney, R. L. (1992). *Value-focused thinking*. Cambridge, MA: Harvard University Press.

Klein, G. (1993). A recognition-primed decision (RPD) model of rapid decision making. In G. Klein, J. Arisen, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods*. Norwood, NJ: Ablex.

Payne, J. W., Bettman, J. R., & Johnson, E. J. (1993). *The adaptive decision maker*. New York: Cambridge University Press.

Rubinstein, A. (1998). *Modeling bounded rationality*. Cambridge, MA: MIT Press.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99–118.

Simon, H. A. (1979). Rational decision making in business organizations. *American Economic Review, 69*, 493–513.

Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science, 211*, 453–458.

Wakker, P. P. (2010). *Prospect theory: For risk and ambiguity*. Cambridge, UK: Cambridge University Press.

Zsambok, C. E., & Klein, G. (1997). *Naturalistic decision making*. Hillsdale, NJ: Erlbaum.

Chromatic Number

In a graph, the minimum of colors needed to ensure that adjacent nodes receive different colors.

See

- ▶ Graph Theory

Chromosome

In genetic algorithms, a chromosome represents a potential solution to the problem at hand.

See

- ▶ Evolutionary Algorithms

CIM

Computer integrated manufacturing.

See

- ▶ Automation in Manufacturing and Services

Circling

- ▶ Cycling

Classical Optimization

- ▶ Unconstrained Optimization

Closed Network

A queueing network in which there is neither entrance nor exit but only a fixed number of customers endlessly circulating.

See

- ▶ Networks of Queues

Closed-Loop Supply Chains

Gilvan C. Souza
Indiana University Bloomington, Bloomington, IN, USA

Introduction

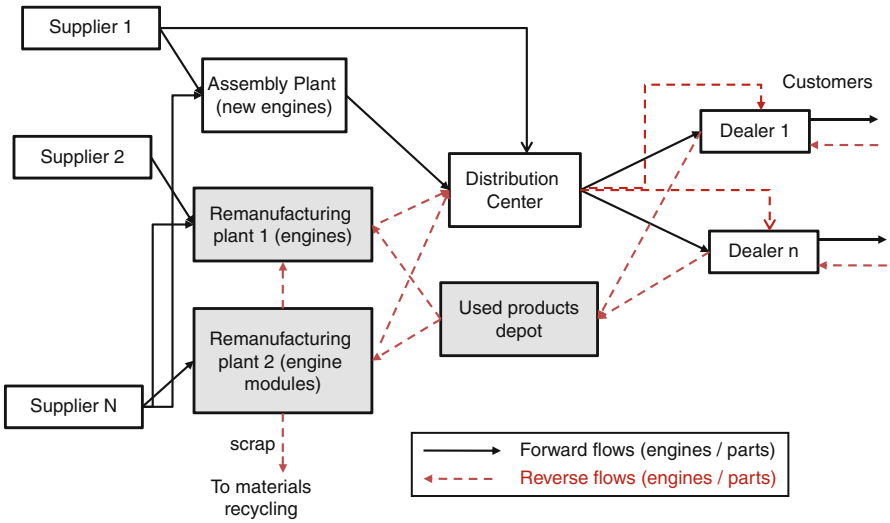
In a (regular) supply chain, there are physical flows of products, components or subassemblies from suppliers to subassembly manufacturers, from subassembly manufacturers to Original Equipment Manufacturers (OEMs), and from OEMs to customers through a distribution system. The distribution system could be comprised of a combination of distribution centers, central, and regional warehouses, and resellers. In the traditional supply chain literature, these physical flows are assumed to be unidirectional, from suppliers to customers. There are, however, bi-directional financial and information flows, for example, orders and payments placed from one tier in the supply chain (e.g., distributors) to its immediate upper tier (OEMs).

Closed-Loop Supply Chains (CLSC)

In a closed-loop supply chain (CLSC), there are, in addition to the forward physical flows described above, reverse physical flows of (used) products or components from customers to manufacturers (and possibly suppliers). As an example of closed-loop supply chain, consider the supply chain for diesel engines and parts for a major North American OEM (Fig. 1). Figure 1 depicts the main physical flows in this supply chain in a simplified manner; the flows are differentiated between forward and reverse flows. Forward flows consist of new parts and/or engines, and reverse flows consist of used parts and/or engines, and remanufactured parts or engines. Remanufacturing is the process of restoring a used product (post consumer use) to a common operating and aesthetic standard, sometimes with upgrades to the original product's functionality. For a diesel engine or module, remanufacturing consists of six different steps: (i) full disassembly to the part level, (ii) thorough cleaning of each part (often through multiple sequential techniques), (iii) making a disposition decision for each part (keep for remanufacturing or dispose the part for materials recycling), (iv) salvaging if necessary (value added work that restores functionality to that of a new part), (v) re-assembly, and (vi) testing. Other terms commonly used for remanufacturing include refurbishing, rebuilding, and overhauling, depending on the industry (no such distinctions are made in this article).

New engines are produced and assembled from new parts, some of which are manufactured and shipped by the OEM's many suppliers. Those suppliers also supply the firm's distribution center with spare parts. New engines are shipped to a (central) distribution center; they are then shipped from this (central) distribution center to several regional distribution centers (not depicted in Fig. 1), and from there to over 3,000 dealers in North America. Customers, say a trucking company, buy new (or remanufactured) diesel engines or engine modules, say a water pump or a turbocharger, from dealers due to replacement needs. They receive a dollar credit from returning the old engine or module upon purchasing a new (or remanufactured) engine or module; the dollar credit can be as high as 30% off the purchase price. Remanufactured engines or modules sell at a 35% discount relative to the corresponding new engine or module. Used modules or engines are shipped from dealers to one of 30 or so different consolidation points in North America (not depicted in Fig. 1), and from there to the OEM's main used products depot. At the depot, shipments are unpacked, customers are given the proper credit for returning the used module or engine; engines and modules are then shipped to one of two plants (or put into inventory for later shipment when needed): engine remanufacturing (plant 1), or part (or module) remanufacturing (plant 2). Remanufactured engines are shipped from plant 1 to the main distribution center, joining new engines or parts for distribution to the dealers. Remanufactured parts or modules are shipped from plant 2 to either the distribution center, or to the engine remanufacturing plant 1, depending on forecasts and current needs. Used parts not suited for remanufacturing are sold to recyclers. The flows depicted in Fig. 1 are simplified, but they convey the major flows in this CLSC. Used products are typically referred to as returns or cores.

The supply chain in Fig. 1 illustrates two major disposition decisions for cores: remanufacturing and recycling. Recycling means materials recovery, i.e., the



Closed-Loop Supply Chains, Fig. 1 CLSC for diesel engines and parts (simplified)

geometry of the used part or product is not preserved. Recycling occurs when remanufacturing is not possible or not economical, for example, the core is highly damaged (e.g., an engine block with a hole on it), there is significant wear and tear of the part (piston rings), or the part is technologically obsolete (this is common in consumer electronics and computers). In addition to these two disposition options, the firm can also disassemble a core, and use some of the resulting parts as spare parts for, say, fulfilling warranty claims or servicing products under service contracts; in these cases there may be no cleaning or salvage needed. Dismantling for spare parts is common in electronic goods industries, such as computers, and IT networking equipment (servers, routers, switches). Dismantling for spare parts can be an attractive alternative when it produces significant savings compared to procuring a new part from a supplier, when demand for remanufactured products is weak, or when the part supplier is no longer active. Other disposition decisions include incineration (which can recover energy, but there can be pollution concerns), and dumping in landfills (which is illegal for some materials known to contaminate water and soil). This article focuses on remanufacturing, as it presents significant operational challenges due to the natural mismatch between supply of cores, which for most firms is not certain, and demand for remanufactured products, which is also uncertain; as a result remanufacturing is a natural candidate for application of OR models.

In addition, Fig. 1 illustrates a CLSC where the main source of cores are end-of-use returns, where the product has undergone a full cycle of use with a customer, but the product still has significant value left for recovery. In addition to end-of-use returns, there are end-of-life returns, which are products that have reached the end of their useful life, mostly due to obsolescence, and whose main disposition decision is recycling; examples include very old computers, monitors, VCRs, and very old cars. Finally, there are consumer returns, which are products that have undergone little or no use by consumers—they are returned by consumers to retailers as a result of liberal returns policies by powerful retailers primarily in North America; most consumer returns are not defective. For example, about 80% of deskjet printers returned to retailers by consumers in the U.S. are not defective; reasons for return include remorse, and lack of product fit with consumer needs (Ferguson et al. 2006).

For the design and operation of a CLSC, the decision making is classified into three buckets, as is the case for regular supply chains: strategic, tactical, and operational. Examples of these three types of decisions are shown in Table 1. As Table 1 suggests, there are many different decisions in CLSC management that are amenable to the use of OR tools and techniques, including mathematical programming, Markov decision processes, and simulation.

Next, a brief description is given on how OR techniques are applied to two decisions: the decision to remanufacture or not by an OEM, and network design. A more complete review of the basic models and extensions for the other decisions is given in Souza (2008) and Ferguson and Souza (2010).

Closed-Loop Supply Chains, Table 1 Strategic, Tactical and Operational Decisions in CLSCs

Decision Type	Examples
Strategic	• **Remanufacturing or not:** Should an OEM remanufacture? • **Network design:** What is the location of remanufacturing plants, recycling plants, collection points, and consolidation points? Should used products be collected through retailers, or directly from consumers? Should forward and reverse flows be combined, or should the forward and reverse supply chain be separate? • **Leasing:** Should the firm lease or sell to customers? • **Strategic alliances:** Should the firm enter into partnerships with third-parties for remanufacturing or collection of its products? • **Design for recovery:** How should a firm design a product if there is remanufacturing at the end of use?
Tactical	• **Product acquisition:** How many used cores should the firm acquire, when, in which quality, and at what price? • **Remanufacturing planning and disposition:** Given a supply of cores, demand forecasts, relevant costs and revenues, what should a firm do with a core (remanufacture, recycle, dismantle for parts), and when?
Operational	• **Disassembly planning:** What is the sequence and depth of disassembly for a core? • **Shop floor scheduling and control**: What is the routing and scheduling priority for remanufacturing orders in the job shop?

Should an OEM Remanufacture?

On the surface, the decision to remanufacture or not by an OEM appears to be simple: if the price the firm can sell a remanufactured product far exceeds the variable cost to remanufacture a core, which includes collection and transportation of the core, disassembly, cleaning, salvaging, re-assembly, testing, and remarketing, then the firm should remanufacture, after properly accounting for any upfront fixed remanufacturing costs (e.g., building a facility and acquiring equipment). This simple revenue vs. cost accounting may not capture all of the facets of the problem, however. There are other factors that favor remanufacturing, such as: (i) extending the OEM's product line and offering a product, priced most likely lower than the corresponding new product, and therefore reaching a customer segment that would not otherwise be reached; (ii) allowing brand protection, given that many third party firms offer remanufactured products—if the OEM offers a certified remanufactured product, then it communicates to consumers that only its version of remanufactured product has the appropriate quality level; (iii) using as a deterrent to market entry of third-party remanufacturers; and (iv) value recovery for used products returned after leases, trade-in programs, or consumer returns. On the other hand, there are factors that do not favor remanufacturing; chiefly among them are: (a) the fear of cannibalization of sales of a (typically more expensive and more profitable) new product by a (typically less expensive and less profitable) remanufactured product; and (b) the ability to reliably collect an appropriate pipeline of cores to sustain a remanufacturing operation. Discussions with manufacturing managers indicate that factor (a) is of significant concern to firms in the IT equipment industry, while factor (b) is of significant concern to firms in automotive parts remanufacturing.

The following simple analytic model provides some insights into the answer to the critical strategic question, "Should an OEM remanufacture?" Consider an OEM selling a new product (say, a diesel engine model X), and considering the decision to offer its remanufactured counterpart (a remanufactured diesel engine model X). The firm has to decide whether to offer the remanufactured product, and if so, how to set the prices of remanufactured and new products, denoted by p_r and p_n respectively. First, consider a monopolist under a single period model (extensions are discussed below.) The model is based on some assumptions about consumer behavior. Specifically, assume that the consumer base is heterogeneous, so that consumers differ in their intrinsic valuation for the new product. A consumer such as a third-party logistics company, for example, with an extensive fleet of large trucks, has a high valuation for the new diesel engine; whereas an operator of small gasoline-powered delivery trucks has a low intrinsic valuation for a new diesel engine. This intrinsic valuation for the new product, which differs across the heterogeneous consumer base, is referred to as willingness-to-pay (the maximum amount a consumer is willing to pay), or w.t.p., for a new

product. Each consumer has an intrinsic w.t.p., a random variable that is denoted by θ. Thus, a consumer has a unique association with its w.t.p. θ and can be referred to simply as consumer θ. Further, it is assumed θ is uniformly distributed between a lower and an upper bound, where the bounds are normalized to be zero dollars and one dollar, respectively. Mathematically, $\theta \sim U[0,1]$. Thus, all consumers (potential customers) are distributed uniformly in the real line between \$0 and \$1. This assumption is common in the marketing and operations literature, because it results in linear demand curves, as shown below (it also allows analytical tractability). A consumer θ's w.t.p. for a remanufactured product is, however, $\delta\theta$, where $0 \leq \delta \leq 1$. If $\delta = 0$, then consumers do not consider the remanufactured product as a potential substitute for the new; this is a limiting case. An example that approaches this limit is retreaded passenger car tires in the U.S., where many consumers perceive them as unsafe, and there are many cheap imports that are priced quite low but are new. If $\delta = 1$, then consumers perceive remanufactured and new products as perfect substitutes. One example of this is retreaded truck tires used in commercial fleets in the U.S., where fleet owners have service contracts with certain dealers and pay them by each mile of service a tire provides to the fleet owner. These firms are insensitive as to whether the dealers use retreaded or new tires to keep the truck running. Most products fall in between, i.e., $0 < \delta < 1$; [see Hauser and Lund (2003), and Souza (2008) for a complete discussion]. For diesel engines, for example, $\delta \cong 0.65$; for power tools $\delta \cong 0.85$.

Suppose the firm only offers the new product at a price $p_n \leq 1$ (the firm would never offer a new product priced higher than \$1 because the maximum w.t.p. in the consumer base is \$1). Then, only those consumers with w.t.p θ higher than p_n buy the product, because they are the only ones with a non-negative net utility $(\theta - p_n)$ for the product. Because consumers' w.t.p. are distributed uniformly between 0 and 1, then the number of consumers that buy the new product (q_n) is $q_n = M \cdot \Pr\{\theta - p_n \geq 0\} = M \cdot \Pr\{\theta \geq p_n\}$, where M is the overall size of the consumer base (number of potential customers). Normalizing $M = 1$, and because $\theta \sim U[0,1]$, then $\Pr\{\theta \geq p_n\} = (1 - p_n)/1 = 1 - p_n$; as a result the firm sells $q_n = 1 - p_n$ new products. Now, suppose the firm offers both new and remanufactured products at prices p_n and p_r, respectively. A consumer θ's net utility for a new product is $\theta - p_n$, and for a remanufactured product is $\delta\theta - p_r$. Consumers whose net utilities are higher for a new than for a remanufactured product, i.e., $\theta - p_n > \delta\theta - p_r$, buy a new product; solving for θ yields $\theta > (p_n - p_r)/(1 - \delta)$. If $\theta < (p_n - p_r)/(1 - \delta)$, then consumers have a higher net utility for a remanufactured than a new product; they will buy remanufactured if their net utility is positive, that is, $\delta\theta - p_r > 0$, or $\theta > p_r/\delta$. Consumers with w.t.p. θ lower than p_r/δ will not buy anything. Given the uniform distribution for θ, the quantities of new and remanufactured products sold, given their prices, are $q_n = 1 - (p_n - p_r)/(1 - \delta)$, and $q_r = (p_n - p_r)/(1 - \delta) - p_r/\delta$, respectively. These two expressions constitute the demand curves for new and remanufactured products given respective prices; the demand curves are linear, assuming a uniform w.t.p. distribution (a different distribution results in a different demand curve shape). For a period with R cores available for remanufacturing, denote the remanufacturing yield—the percentage of cores that are found fit for remanufacturing—by μ. Further, assume that the remanufacturing cost per unit is constant at c_r, and the manufacturing cost per unit (new) is c_n. Then, the OEM's decision problem can be formulated as:

$$\max_{p_n, p_r} \Pi = q_n(p_n - c_n) + q_r(p_r - c_r), \tag{1}$$

$$\text{s.t. } q_r \leq \mu R, \tag{2}$$

$$q_n, q_r \geq 0, \tag{3}$$

Equation (1) is the OEM's per period profit; equation (2) is a constraint that limits the availability of cores for remanufacturing, and equation (3) is a logical constraint. Note that the decision variables are the prices p_n and p_r, thus, one needs to substitute the corresponding expressions for the quantities $q_n = 1 - (p_n - p_r)/(1 - \delta)$, and $q_r = (p_n - p_r)/(1 - \delta) - p_r/\delta$ in (1–3). This is a non-linear optimization problem, which can be solved analytically. The solution comprises several regions, depending on which constraints are binding or not. It can be shown that if $c_r < c\delta$ (and $R > 0$), then

the firm remanufactures, i.e., $q_r > 0$. Thus, the decision to remanufacture in this simple model is dependent upon the unit remanufacturing cost relative to new, the consumer's perception of remanufactured products relative to new (δ), and the availability of cores.

The model described above is very stylized, and does not include the following factors that are (typically) present in real life:

1. Competition: This transforms the decision problem into a game. There is a second decision maker with an objective function similar to the second term in (1). The demand functions now become significantly more complicated. See Atasu et al. (2008) for a way to incorporate competition.
2. Non-linear recovery costs: the model above assumes a constant marginal remanufacturing cost c_r. In practice, there is a cost of collection that is convex increasing in the quantity of cores collected, because it is increasingly more difficult to improve collection rates. Remanufacturing cost per se (i.e., disassembly, cleaning, salvage, testing) may also be convex increasing in the quantity remanufactured, because as remanufacturing quantity increases, the firm needs to dig deeper into the pile of cores, and remanufacture cores in worse quality condition, which demands more labor and materials. The combination of convex collection and remanufacturing costs implies that c_r in (1) is substituted with αq_r^2, where α is a positive constant. See Ferguson and Toktay (2006) for an analysis of this problem.
3. Availability of cores is dependent on sales in previous periods: simply put, the number of cores available for recovery are a function of sales in previous periods. To capture this dimension, one needs a multi-period model, so that all decision variables are defined for each period t: $p_{n,t}$, $p_{r,t}$, $q_{n,t}$, and $q_{r,t}$. If a product can only be remanufactured once (for example, if the product becomes technologically obsolete after the third generation is introduced), and L is the lag, in periods, between the sale of a new product, and its collection as a core post-consumer use, then (2) should be rewritten as $q_{r,t} \leq \mu q_{n,t-L}$, for each $t > L$. (See Ferrer and Swaminathan (2006), and Debo et al. (2005) for examples of models incorporating this dynamic aspect).

CLSC Network Design

As Table 1 indicates, designing a CLSC network requires deciding upon the locations of manufacturing and remanufacturing plants, warehouses (or distribution centers), points of sale, and consolidation centers for shipping cores from points of sale to remanufacturing plants, among other facilities. To help design such a network, a mixed-integer linear program (MILP) is described next based on the modeling framework by Fleischmann et al. (2001). For a review of CLSC network design, see Ammons et al. (2001) and Pochampally et al. (2008).

Assume an OEM that manufactures and remanufactures products, similar to the diesel engine CLSC shown in Fig. 1. The supply chain comprises four levels: (i) manufacturing and remanufacturing plants (a facility can do one or both), (ii) warehouses for distribution of manufactured and remanufactured products, (iii) consolidation centers for consolidating shipments of used products originating from resellers for shipment to plants, and (iv) resellers, who are independent entities that sell manufactured and/or remanufactured products to customers, in addition to collecting used products from customers.

Indexes

i Potential plant locations, $i \in I, I_0 = I \cup \{0\}$, where $i = 0$ is the disposal option.
j Potential warehouse locations, $j \in J$
k Fixed reseller locations, $k \in K$
l Potential consolidation center locations, $l \in L$

Variables

X_{ijk}^f Fraction of reseller k's demand served from plant i through warehouse j
X_{kli}^r Fraction of reseller k's returns returned to plant i through consolidation center l
U_k Unsatisfied fraction of reseller k's demand
W_k Uncollected fraction of reseller k's returns
Y_i^p Indicator variable for opening plant i ($= 1$ if plant is open; 0 otherwise); Y_j^w and Y_l^r are similarly defined

Costs

c_{ijk}^f Unit cost (transportation, production, handling) of serving k from i via j

c^r_{kli} Unit cost of returns (transportation, handling) from k to i via l

c^r_{kl0} Unit disposal cost (including collection, transportation, handling) for k via l

c^u_k Unit penalty cost for not serving reseller k's demand

c^w_k Unit penalty cost for not collecting reseller k's returns

f^p_i Fixed cost for opening plant i (f^w_j and f^r_l similarly defined)

Parameters

D_k Demand for reseller k

R_k Returns from reseller k

γ Minimal disposal fraction

Note that in this formulation, the continuous decision variables (e.g., X^f_{ijk} and X^r_{kli}) are defined in terms of *fractions* of demand and returns at each reseller, and as a result they are all bounded below by zero and above by one. An alternative formulation would have the continuous decision variables defined simply as quantities shipped. The firm's network design problem can be formulated mathematically as a MILP as follows:

$$TC = \min \sum_{i\in I} f^p_i Y^p_i + \sum_{j\in J} f^w_j Y^w_j + \sum_{l\in L} f^r_l Y^r_l + \sum_{i\in I}\sum_{j\in J}\sum_{k\in K} c^f_{ijk} D_k X^f_{ijk} + \sum_{k\in K}\sum_{l\in L}\sum_{i\in I_0} c^r_{kli} R_k X^r_{kli} + \sum_{k\in K}\left(c^u_k D_k U_k + c^w_k R_k W_k\right) \tag{4}$$

$$\text{s.t.} \sum_{i\in I}\sum_{j\in J} X^f_{ijk} + U_k = 1, \forall k \tag{5}$$

$$\sum_{l\in L}\sum_{i\in I_0} X^r_{kli} + W_k = 1, \forall k \tag{6}$$

$$\sum_{k\in K}\sum_{l\in L} R_k X^r_{kli} \le \sum_{j\in J}\sum_{k\in K} D_k X^f_{ijk}, \forall i \tag{7}$$

$$\gamma \sum_{i\in I_0} X^r_{kli} \le X^r_{kl0}, \forall k, \forall l \tag{8}$$

$$\sum_{j\in J} X^f_{ijk} \le Y^p_i, \forall i, \forall k \tag{9}$$

$$\sum_{i\in I} X^f_{ijk} \le Y^w_j, \forall j, \forall k \tag{10}$$

$$\sum_{i\in I_0} X^r_{kli} \le Y^r_l, \forall k, \forall l \tag{11}$$

$$Y^p_i, Y^w_j, Y^r_l \in \{0,1\}\ \forall i, \forall j, \forall l \tag{12}$$

$$0 \le X^f_{ijk}, X^r_{kli}, U_k, W_k \le 1, \forall i, \forall j, \forall k. \tag{13}$$

The objective function (4) minimizes total cost, comprised of fixed costs of opening and operating the facilities, and variable distribution costs. Constraints (5–6) represent basic flow constraints, which indicate that, for each reseller k, shipments plus unsatisfied demand are equal to total demand; similarly for reseller k's returns. Constraint (7) represents flow balancing at each plant, where the difference between incoming returns and outgoing shipments represent manufacturing of new products. Constraint (8) indicates that the number of disposed products should be larger than a given fraction of all returned products; in this model disposal is meant to represent material recycling or dismantling for spare parts. Thus, constraint (8) indicates the extent remanufacturing should take place (for example if $\gamma = 1$, then there is no remanufacturing, and all cores are recycled, or dismantled for spare parts). Constraints (9–11) are logical constraints—there is no shipment to/from a facility if that facility is not open. Constraint (12) states the binary decision variables for the problem, and constraint (13) represents the non-negativity constraints, and the fact that the continuous variables in this problem are defined as fractions of total demand or total returns at each reseller.

As described in Fleischmann et al. (2001), this formulation is very general and can accommodate many different scenarios. For example, if the firm has two separate networks for forward and reverse flows, then it can set R_k and D_k equal to zero, respectively. The values of c^f_{ijk} relative to c^r_{kli} and c^u_k can model different production scenarios for each plant, such as whether a plant only produces new products or it only produces remanufactured products, or both. In this model, demand can be met through remanufactured or new products. If there are separate demand streams for these products, then one can add another index, say t, to the decision variables and parameters of the problem to indicate the product type. For example, D_{kt} would be demand at customer k for product type t, where $t \in$ {remanufactured, new}. Finally, this

problem has been aggregated in that there is only one "aggregate" product being sold at the resellers. In the case of the CLSC for diesel engines, modules and parts remanufacturing are treated differently than engine remanufacturing; to accommodate this scenario one can again simply add another index to indicate product type (say, entire engines, or modules).

The discussion above was centered on remanufacturing as the key recovery activity taking place in the network. The same formulation can be used to design a network for recycling; an example is paper recycling, studied by Bloemhof-Ruwaard et al. (1996).

CLSC OR Applications

Closed-loop supply chain management provides numerous opportunities for application of OR methodology. The two applications discussed above are at the strategic level: an OEM's decision to engage in remanufacturing, and the design of a network of remanufacturing and manufacturing plants, distribution centers, and consolidation centers. The first model analyzes an OEM's decision to remanufacture—it is based on a model of consumer behavior that, in essence, implies linear demand curves for remanufactured and new products, and where they are partial substitutes for each other, so that each customer values a remanufactured product less than a corresponding new product. Based on these demand curves, and relevant costs, the firm chooses prices for remanufactured and new products that maximize its profit by solving a non-linear program. The second model is more of a decision support-type model—given relevant fixed costs of opening new facilities, relevant distribution costs, and demand and return points, the firm designs its CLSC network to minimize its fulfillment costs.

One significant area of application of OR models not covered in this article concerns the match between supply of cores and demand for remanufactured products and parts. The problem of product acquisition—acquiring the right amount of cores at the right price at the right quality at the right time, and its corresponding disposition decision—deciding what to do with a core: disassemble, remanufacture, recycle, or put it in inventory for future use, given relevant demand forecasts, and underlying costs is of significant importance to firms, at the tactical level. See Souza (2008) and Ferguson and Souza (2010) for more information on these models.

For the range of CLSC applications, especially from a policy maker's perspective, the design of environmental legislation is of significant concern. Specifically, the decision maker (say, the government) is interested in designing environmental legislation that sets appropriate collection and recycling levels to maximize society's welfare. This is comprised of total profits across all manufacturers impacted by legislation, consumer surplus, and environmental benefits of the legislation (e.g., lower pollution levels); notice that environmental benefits must be measured in dollar terms. Again, OR models can be used to help design such legislation (Ferguson and Souza 2010).

See

- ▶ Industrial Applications
- ▶ Integer and Combinatorial Optimization
- ▶ Supply Chain Management

References

Ammons, J. C., Realff, M. J., & Newton, D. J. (2001). Decision models for reverse production system design. In C. N. Madu (Ed.), *Handbook of environmentally conscious manufacturing*. Boston: Kluwer Academic Publishers.

Atasu, A., Sarvary, M., & Van Wassenhove, L. N. (2008). Remanufacturing as a marketing strategy. *Management Science, 54*, 1731–1746.

Bloemhof-Ruwaard, J., Van Wassenhove, L. N., Gabel, H., & Weaver, P. (1996). An environmental life cycle optimization model for the European pulp and paper industry. *Omega, 24*, 615–629.

Debo, L. G., Toktay, L. B., & Van Wassenhove, L. N. (2005). Market segmentation and production technology selection for remanufactured products. *Management Science, 51*, 1193–1205.

Ferguson, M. E., Guide, V. D. R., Jr., & Souza, G. C. (2006). Supply chain coordination for false failure returns. *Manufacturing & Service Operations Management, 8*, 376–393.

Ferguson, M. E., & Souza, G. C. (Eds.). (2010). *Closed-loop supply chains: New developments to improve the sustainability of business practices*. Boca Raton: CRC Press.

Ferguson, M. E., & Toktay, B. (2006). The effect of competition on recovery strategies. *Production and Operations Management, 15*, 351–368.

Ferrer, G., & Swaminathan, J. (2006). Managing new and remanufactured products. *Management Science, 52*, 15–26.

Fleischmann, M., Beullens, P., Bloemhof-Ruwaard, J., & Van Wassenhove, L. N. (2001). The impact of product recovery on logistics network design. *Production and Operations Management, 10*, 156–173.
Hauser, W., & Lund, R. (2003). *The remanufacturing industry: Anatomy of a giant*. Boston: Boston University, Department of Manufacturing Engineering Report.
Pochampally, K. K., Nukala, S., & Gupta, S. (2008). *Strategic planning models for reverse and closed-loop supply chains*. Boca Raton, FL: CRC Press.
Souza, G. (2008). Closed-loop supply chains with remanufacturing. In Z. L. Chen & R. Raghavan (Eds.), *Tutorials in operations research*. Hanover, MD: INFORMS.

Cluster Analysis

Jay E. Aronson[1] and Lakshmi S. Iyer[2]
[1]The University of Georgia, Athens, USA
[2]The University of North Carolina at Greensboro, Greensboro, NC, USA

Introduction

Cluster analysis is a generic term for various procedures that are used objectively to group entities based on their similarities and differences. In applying these procedures, the objective is to group the entities (elements, items, objects, etc.) into mutually exclusive clusters so that elements within each cluster are relatively homogeneous in nature while the clusters themselves are distinct. The key purposes of cluster analysis are reduction of data, data exploration, determination of natural groups, prediction based on groups, classification, model fitting, generation and testing of hypotheses (Everitt 1993; Aldenderfer and Blashfield 1984; Lorr 1983).

Due to the importance of clustering in different disciplines such as psychology, zoology, botany, sociology, artificial intelligence and information retrieval, a variety of other names have been used to refer to such techniques: Q-analysis, typology, grouping, clumping, classification, numerical taxonomy, and unsupervised pattern recognition (Everitt 1993). In fact, as Jain and Dubes (1988) noted: "I.J. Good (1977) has suggested the new name botryology for the discipline of cluster analysis, from the Greek word for a cluster of grapes."

Though clustering techniques have existed for many years, profuse work in this area has been accomplished only in the past two decades. The primary stimuli for this were the founding of the Classification Society in 1970 and the publication of the *Principles of Numerical Taxonomy* by Sneath and Sokal (1973; also see, Lorr 1983). Other reasons for the rapid growth in cluster analysis literature are the basic importance of classification as a scientific procedure, prolific developments in high-speed computers, and the need to solve large, real-world problems efficiently. The complexity of clustering methods are known to increase tremendously with increase in problem sizes. With the availability of sophisticated computing power, the handling of large practical problems is of less concern now.

Applications of Cluster Analysis

Clustering methods are applied in a variety of fields including psychology, biology, medicine, economics, marketing research, pattern recognition, weather prediction, environmental science, linguistics, information systems design, electronic brainstorming and flexible manufacturing systems. Some interesting cluster analyses include analyzing large engineering records collections (Homayoun 1984), measuring welfare and quality of life across countries (Hirschberg et al. 1991), management of cutting tools in flexible manufacturing systems (DeSouza and Bell 1991), clustering as a quality management tool (Spisak 1992), identifying the structure and content of human decision making (Allison et al. 1992), mapping consumers' cognitive structures (Hodgkinson et al. 1991), information systems design (Aronson and Klein 1989; Karimi 1986; Klein and Aronson 1991), vehicle routing, production scheduling and sampling (Romesburg 1984), income tax bracket determination (Mulvey and Crowder 1979), management team construction, and idea grouping to handle information overload in electronic brainstorming. The maximum diversity problem forms clusters based on maximizing the differences (distances) among the items rather than the similarities. Applications include forming a single, diverse group from a larger set in which the objective function is imposed only on those items in the group (Kuo et al. 1993), and multiple diverse groups consisting of all items (Weitz and Lakshminarayanan 1997, 1998). Punj and Stewart (1983) provide a good description of the applications of cluster analysis,

including some details on the various clustering packages and programs that are available.

Clustering Techniques

Authors such as Everitt (1993), Cormack (1971), Aldenderfer and Blashfield (1984), Hartigan (1975), and Anderberg (1973) have provided good reviews on existing clustering methods. Nevertheless, there has been no unique classification of the various clustering methods. In fact, this is one of the pitfalls of cluster analysis. Due to work by Cormack (1971), Punj and Stewart (1983), and Everitt (1993), the following five categories have been accepted as a basis:

1. Hierarchical methods,
2. Optimization techniques,
3. Density search techniques,
4. Clumping methods, and
5. Other techniques.

Hierarchical methods: Hierarchical procedures are tree-like structures in which elements are first separated into broad classes. These classes are further subdivided into smaller classes and so on until the terminal classes are not further subdivisible. These methods are most frequently used in the biological sciences. The hierarchical methods are basically of two types — agglomerative and divisive.

The agglomerative methods begin by making each item its own cluster. In subsequent iterations two or more closest clusters are combined to form a new, aggregate cluster. Eventually, all items are grouped into one large cluster. Hence, these methods are some times referred to as build-up methods (Hair et al. 1987).

In contrast to agglomerative methods are the divisive methods that begin with one large cluster. Groups of items that are most dissimilar are removed and placed into smaller clusters. The process continues until each item becomes a one-element cluster. Cormack (1971), Everitt (1993), Aldenderfer and Blashfield (1984) and Hair et al. (1987) have provided comprehensive descriptions of the various agglomerative and divisive procedures.

Optimization techniques: These methods allow relocation of items during the clustering process, improving from an initial solution to optimality. The number of clusters must be decided a priori, although some methods allow for changes (manually or automatically) while solving. There are differences in optimization techniques due to the different methods used for obtaining an initial solution and different objective criteria (Everitt 1993).

Since most of the objective criteria of the optimization techniques are based on those of the well-established statistical concepts, very few mathematical programming approaches have been developed to solve these problems. The statistical methods have proved adequate for many situations, because (1) the solutions found are believed to be reasonably close to the optimum; (2) the solutions typically involve human analyst intervention to determine when an appropriate number of clusters have been identified; and (3) the combinatorial nature of clustering makes it difficult to solve a large problem to a guaranteed optimum. Mulvey and Crowder (1979) developed a subgradient method coupled with a simple search procedure for solving the clustering problem. However, their method did not yield an exact optimum. Though heuristics generally seem efficient, the need to obtain optima to problems such as effective information systems design (Klein et al. 1988) make heuristics less attractive. Klein and Aronson (1991) developed a mixed-integer programming model and method to obtain an optimal solution to clustering problems, where the objective function is the sum of pairwise interactions among all items in each cluster. No metric space nor median are used. Their method is based on the implicit enumeration method of Balas (1965). Extensions including precedence and group size limits are discussed by Aronson and Klein (1989). Earlier, Gower and Ross (1969) and Rohlf (1974) showed that there is a direct relationship between some common cluster formulations and certain types of well-known graph theoretic problems, primarily that of the minimum spanning tree. A further expansion on the use of graph theoretic techniques in cluster analysis may be found in Matula (1977).

Density search techniques: This concept, proposed by Gengerelli (1963), depicts the items as points in a metric space. Parts of the space where the distribution of points is very dense but separated by parts of low density suggest natural clusters. Everitt (1993) describes the different types of density search techniques.

Clumping techniques: These techniques are most popular in language studies where words that tend to have several meanings, when classified based on their

meaning, belong to several groups. Thus, in general, clumping techniques allows for overlapping clusters. This terminology was introduced by Jones, Needham, and co-workers at the Cambridge Language Research Unit (Everitt 1993). This method attempts to partition entities into two groups based on the similarity matrix from the original data. The Needham (1967) criterion is to minimize the cohesion function between the two groups. Other clumping procedures are also discussed in Rohlf (1974) and Everitt (1993).

Other techniques: This comprises all clustering techniques that do not fall in to the above four categories. For example, there is inverse "Q" factor analysis that is commonly used in behavioral sciences (Cattell 1952). The "R" factor analysis is a type of "Q" factor analysis that utilizes the correlations between variables. Gower (1966) provided a good review of the properties of various "Q" and "R" factor analysis techniques. Everitt (1993) and Aldenderfer and Blashfield (1984) included a good summary of various other clustering methods.

Issues of Concern

Though initially the concepts of cluster analysis seem to be intuitive, one can encounter a host of problems while performing an actual analysis. Some of the problems include selection of data units and variables, knowing exactly what to cluster, distance or similarity measures, transformation of measures, clustering criterion, the clustering method to use, the number of clusters and interpretation of the results (Anderberg 1973). Authors such as Aldenderfer and Blashfield (1984), Everitt (1993), Hair et al. (1987) and Anderberg (1973) have addressed some of the issues in great detail. A few of the more critical issues are discused next.

Measurement of distance or similarity matrix: The relationship between elements are represented by using either a similarity or distance measure. While similarity measures (indicating cohesion) take values between 0 and 1, distance measures can be any positive value. The output of any clustering method depends on the type of input measure used. One of the most commonly used measures is the Euclidean distance. This concept can be easily generalized for additional variables (Hair et al. 1987).

Another measure which allows for correlations between variables was originally proposed by Mahalanobis in 1936 (Everitt 1993). This is similar to Euclidean distance measure using standardized variables when the correlations are zero. The Mahalanobis distance measure has been used by McRae (1971). Everitt (1993), Hair Jr et al. (1987) and Hartigan (1975) have provided some discussions on other types of distance measurements. The clustering model for computer-assisted organization presented by Klein and Aronson (1991) accounts for total pairwise interactions independent of a metric. The need to consider all interactions among items in each cluster led to the formulation of a mixed-integer model for optimal clustering based on scaled, pairwise distance (Klein and Aronson 1991).

Which clustering method to use: The problem of choosing an appropriate clustering method generally arises after one has determined the variables, distance measure and criterion for clustering. A number of software packages and programs are available for clustering. Punj and Stewart (1983) and Anderberg (1973) identified some early programs for clustering; now, all major statistical packages contain one or more routines to do such analyses effectively. For selecting the best clustering method one should be aware of the performance characteristics of the various methods (Hair et al. 1987).

Appropriate number of clusters: One of the practical issues of concern in clustering is choosing the number of clusters. Some algorithms find the best fitting structure for a given number of clusters while others, like the hierarchical methods, provide configurations from the number of entities to one large cluster, that is, the entire data set as one cluster. However, if the number of clusters cannot be predetermined, a range of clusters can be selected, solving the problem for each of those cluster sizes, and then selecting the best alternative (Hair et al. 1987).

See

- ▶ Data Mining
- ▶ Decision Making and Decision Analysis
- ▶ Graph Theory
- ▶ Information Systems and Database Design in OR/MS
- ▶ Integer and Combinatorial Optimization
- ▶ Minimum Spanning Tree Problem
- ▶ Vehicle Routing

References

Abonyi, J., & Feil, B. (2007). *Cluster analysis for data mining and system identification*. Basel: Birkhäuser.

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. California: Sage Publications.

Allison, S. T., Jordan, A. M. R., & Yeatts, C. E. (1992). A cluster-analytic approach toward identifying the structure and content of human decision making. *Human Relations, 45*, 49–73.

Anderberg, M. R. (1973). *Cluster analysis for applications*. New York: Academic.

Aronson, J. E., & Klein, G. (1989). A clustering algorithm for computer-assisted process organization. *Decision Sciences, 20*, 730–745.

Balas, E. (1965). An additive algorithm for solving linear programs with zero–one variables. *Operations Research, 13*, 517–546.

Cattell, R. B. (1952). *Factor analysis: An introduction and manual for the psychologist and social scientist*. New York: Harper.

Cormack, R. M. (1971). A review of classification. *Journal of the Royal Statistical Society (Series A), 134*, 321–367.

DeSouza, R. B. R., & Bell, R. (1991). A tool cluster based strategy for the management of cutting tools in flexible manufacturing systems. *Journal of Operations Management, 10*, 73–91.

Everitt, B. (1993). *Cluster analysis* (3rd ed.). New York: Halsted Press.

Gengerelli, J. A. (1963). A method for detecting subgroups in a population and specifying their membership. *Journal of Psychology, 5*, 456–468.

Good, I. J. (1977). The botryology of botryology. In J. Van Ryzin (Ed.), *Classification and clustering*. New York: Academic.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika, 53*, 325–338.

Gower, J. C., & Ross, G. J. S. (1969). Minimum spanning trees and single linkage cluster analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 18*, 54–64.

Hair, J. F., Jr., Anderson, R. E., & Tatham, R. L. (1987). *Multivariate data analysis* (2nd ed.). New York: Macmillan.

Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley.

Hirschberg, J. G., Maasoumi, E., & Slottje, D. J. (1991). Cluster analysis for measuring welfare and quality of life across countries. *Journal of Econometrics, 50*, 131–150.

Hodgkinson, G. P., Padmore, J., & Tomes, A. E. (1991). Mapping consumer's cognitive structures: A comparison of similarity trees with multidimensional scaling and cluster analysis. *European Journal of Marketing, 25*, 41–60.

Homayoun, A. S. (1984). The use of cluster analysis in analyzing large engineering records collection. *Records Management Quarterly, October*, 22–25.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliffs: Prentice Hall.

Karimi, J. (1986). An automated software design methodology using CAPO. *Journal of Management Information Systems, 3*, 71–100.

Klein, G., & Aronson, J. E. (1991). Optimal clustering: A model and method. *Naval Research Logistics, 38*, 447–461.

Klein, G., Beck, P. O., & Konsynski, B. R. (1988). Computer aided process structuring via mixed integer programming. *Decision Sciences, 19*, 750–761.

Kuo, C.-C., Glover, F., & Dhir, K. S. (1993). Analyzing and modeling the maximum diversity problem by zero–one programming. *Decision Sciences, 24*, 1171–1185.

Lorr, M. (1983). *Cluster analysis for social scientists*. California: Jossey-Bass Publishers.

Matula, D. W. (1977). Graph theoretic techniques for cluster analysis algorithms. In J. Van Ryzin (Ed.), *Classification and clustering*. New York: Academic Press.

McRae, D. J. (1971). MICKA, A FORTRAN IV iterative K-means cluster analysis program. *Behavioural Science, 16*, 423–424.

Mulvey, J., & Crowder, H. (1979). Cluster analysis: An application of lagrangian relaxation. *Management Science, 25*, 329–340.

Needham, R. M. (1967). Automatic classification in linguistics. *The Statistician, 17*, 45–54.

Punj, G., & Stewart, D. W. (1983). Cluster analysis in marketing research: Review and suggestions for application. *Journal of Marketing Research, 20*, 134–148.

Rohlf, F. J. (1974). Graphs implied by the Jardine-Sibson overlapping clustering methods. *Journal of the American Statistical Association, 69*, 705–710.

Romesburg, H. (1984). *Cluster analysis for researchers*. Belmont, CA: Lifetime Learning Publications.

Sneath, P. H. A., & Sokal, R. R. (1973). *Principles of numerical taxonomy*. San Francisco: W.H. Freeman.

Spisak, A. W. (1992). Cluster analysis as a quality management tool. *Quality Progress, 25*, 33–38.

Weitz, R. R., & Lakshminarayanan, S. (1997). An empirical comparison of heuristic and graph theoretic methods for creating maximally diverse groups, VLSI design, and exam scheduling. *Omega, 25*, 473–482.

Weitz, R. R., & Lakshminarayanan, S. (1998). An empirical comparison of heuristic methods for creating maximally diverse groups. *Journal of the Operational Research Society, 49*, 635–646.

Cobb-Douglas Production Function

► Economics and Operations Research

COEA

Cost and operational effectiveness analysis.

See

► Cost Analysis

Coefficient of Variation

The ratio of the standard deviation to the mean of a random variable.

Cognitive Mapping

A graphical notation for capturing concepts in use by decision makers for understanding a problematic situation. Concepts are fixed by reference to polar opposites, and directed arcs indicate perceived causal relationships.

See

► Problem Structuring Methods

Coherent System

► System Reliability

COIN-OR Computational Infrastructure for Operations Research

► Open-Source Software and the Computational Infrastructure for Operations Research (COIN-OR)

Column Generation

A technique that permits solution of very large linear-programming problems by generating the columns of the constraint matrix only when they are needed. It is typically employed when the constraint matrix is too large to be stored, or when it is only known implicitly. Column generation, as imbedded in the revised simplex method, has been used to solve the trim problem and other such problems in which the columns are formed from combinatorial considerations.

See

► Integer and Combinatorial Optimization
► Lagrangian Relaxation
► Trim Problem

Column Vector

One column of a matrix or a matrix consisting of a single column.

See

► Matrices and Matrix Algebra

Combat Model

A model whose object is military combat or some aspect of some combat. Three associated terms are often used as synonyms, but are frequently used to differentiate three common aspects of combat modeling: combat model, combat simulation, and war game. When combat model is used as a discriminator it often is used to mean that the model in question is an analytic combat model.

See

► Analytic Combat Model
► Battle Modeling

Combat Simulation

A type of model whose object is military combat or some aspect of combat. Combat simulation is used as a discriminator to emphasize the time or process aspect of the model in question.

See

► Battle Modeling

Combinatorial Auctions

Karla L. Hoffman
George Mason University, Fairfax, VA, USA

Introduction

The advent of the Internet has led to the creation of global marketplaces in which sales of everything from low-cost used merchandise to billion dollar government procurements are conducted through auctions. This article concentrates on designs where many items are auctioned simultaneously and where bidders have the flexibility to combine the goods into packages. The discussion (1) highlights alternative combinatorial auction designs and provides the reader with multiple references to resources that describe more fully the underlying theory of these designs., and (2) describes the mechanisms used to evaluate the efficacy of such approaches in terms of their efficiency, equity, and cognitive complexity, and presents some examples of the use of combinatorial auctions for high-value government lease rights, as well as the use of such auctions for supply-chain procurement. These auctions require knowledge of both game theory and combinatorial optimization.

General Concepts

Governments throughout the world use auctions to lease the right to explore and extract minerals, fuel, and lumber on government properties, to use the airwaves for mobile or broadcast communications, or to control emissions through cap and trade regulations. In addition, the use of business-to-business auctions (often called supply chain auctions) has become a billion-dollar industry. In each of these cases, the need to be able to bundle buys and sells has resulted in new auction theory and designs that enable the simultaneous selling or buying of items using mechanisms that allow participants to indicate their value for the entire package which may have a greater value than the sum of the items within that package. In addition, such auction designs allow users to specify quantity discounts, to indicate budget constraints on the total procurement, and to define other goals of the auction, e.g. social welfare goals in a government auction. These auction designs are computationally more complex for all participants and require languages that allow bidders to express their willingness to participate at a given price for a collection of objects. Such auctions have been termed combinatorial auctions. There are many books that describe the history of auctions, auction theory and its relationship to game theory, and others that are focused exclusively on combinatorial auction designs. For further reading on the subject, see: McMillan (2002) on the history of markets, Krishna (2002) on auction theory, Steiglitz (2007) on the success and pitfalls of EBAY auctions, Klemperer (2004) on auction theory and practice, and Milgrom (2004) and Cramton et al. (2005) on combinatorial auctions. In this review, only the major topics of the field are described, but multiple references are provided for further reading.

In what follows, one-sided auctions are considered and are restricted to the case where there is a single seller and multiple buyers (two-sided auctions are often referred to as exchanges, see Milgrom (2007), Parkes et al. (2001), and Hoffman and Menon (2010) on exchange designs). Since the multiple-sellers/single-buyer case and the multiple-buyers/single-seller case are symmetric, the discussion emphasizes the latter, but all results follow for either case. The concentration is on auction designs where that there are multiple items being sold. For at least some of the buyers, a collection of items must be procured to have a viable business plan; consideration is given only to auction designs that allow the packaging of collections of items. Such designs can provide greater efficiency, as well as greater revenue to the seller than the sequential selling of items individually. These designs are sufficiently general to allow bidders to express a value on a package where the collection of items may have a value greater than the individual items (i.e. the goods are complements), as well as on a package where a buyer can express a quantity discount for buying more of the good (i.e. the goods are substitutes).

Why are auctions such a popular mechanism for buying and selling valuable objects? With the advent of the Internet, auctions are capable of reaching many more possible participants. Here, the potential buyers wish to determine the minimum price that they must pay given that they must compete with others for the ownership of a good or collection of goods. From the

seller's perspective, submitting goods to an auction may increase the number of buyers, thereby increasing the potential for competitive bidding and higher selling prices. Thus, an auction is a mechanism to determine the market-based price, since the bidders set the price through the competition among the bids. This mechanism is dynamic and reacts to changes in market conditions. The determination of selling prices by an auction is perceived as fairer than if the price were set by bilateral negotiations because all buyers must adhere to the same set of rules. Most importantly, if the rules are well designed, the result will have the goods allocated to the entity that values them the most.

The two basic classes of auctions are described next: (1) sealed bid auctions whereby there is only a single opportunity to provide bids to the auction, and (2) multi-round auctions where bids are taken over a period of time and any high bid can be overtaken whenever a new bid is received that increases the overall revenue to the seller.

Sealed Bid Auctions

One common auction mechanism is the first-price (sealed bid) auction**.** In this design, all bidders submit their bids by a specified date. The bids are examined simultaneously and the auctioneer determines the set of bidders that maximizes the revenue to the seller. The optimization problem that determines a collection of package bids that do not often overlap and produce the maximum revenue is known as the Winner Determination Problem (WDP). Mathematically, the problem can be stated as follows:

$$WDP_{OR} : Max \sum_{b=1}^{\#Bids} BidAmount_b x_b \quad (1)$$

$$subject\ to:$$

$$Ax \leq 1$$

$$x \in \{0, 1\} \quad (2)$$

where x_b is a zero–one variable which indicates whether bid b loses or wins, respectively. A is an n x m matrix with m rows, one for each item being auctioned. Each of the n columns represents a bid where there is a one in a given row if the item is included in the bid and zero otherwise. Constraint set (1) specifies that each item can be assigned at most once. Set (1) constraints are equations when the seller chooses to put a minimum price on each item and is unwilling to sell any item below that price. In this case, there is a set of m bids each with only a single item in the package and a bid price at a price slightly below the minimum opening bid price. In this way, the seller will keep the item rather than allow it to be won by a bidder at less than the opening bid price.

In this formulation of the WDP, the bidder can win any combination of bids, as long as each item is awarded only once; this is referred to as the "OR" language. The problem with this language is that it creates a type of exposure problem, that of winning more than the bidder can afford. When multiple bids of a single bidder can be winning, it is incumbent on the software to highlight the maximum exposure to the bidder. This calculation requires that a combinatorial optimization problem be solved for each bidder that calculates the dollar exposure, creating new computational issues for the auctioneer and may result in packages that are not best for the bidder.

The most natural alternative to this "OR" language is the "XOR" language. In this case, the user supplies every possible combination of bids of interest along with a maximum bid price that she is willing to pay for that package. This language removes the dollar exposure problem, since the maximum number of bids that a bidder can possibly pay is the highest bid amount of any of its bids. The problem with the XOR language is that it places a new burden on the bidder: the bidder is forced to enumerate all possible combinations of packages of interest and their associated values. Clearly, as the number of items in an auction increase, the number of possible bids goes up exponentially. When the XOR bidding language is used the Winner Determination Problem (WDP_{XOR}) becomes:

$$WDP_{xor} : Max \sum_{b=1}^{\#Bids} BidAmount_b x_b \quad (3)$$

$$subject\ to:$$

$$x = 1$$

$$\sum_{b \in S_B} x_b \leq 1 \text{ for each bidder B} \quad (4)$$

$$x_b \in \{0, 1\} \quad (5)$$

Where S_B is the set of bids of bidder B, and constraint set (4) specifies that at most one of these bids can be in the winning set.

Fujishima et al. (1999) proposed a generalization of the OR language that does not require the enumeration of all possible combinations. They label this language OR*. Here, each bidder is supplied dummy items (these items have no intrinsic value to any of the participants). When a bidder places the same dummy item into multiple packages, it tells the auctioneer that the bidder wishes to win at most one of these collections of packages. This language is fully expressive, as long as bidders are supplied sufficient dummy items. This language is also relatively simple for bidders to understand and use, as was shown in a Sears Corporation supply-chain transportation auction. In that auction, all bids were treated as "OR" bids by the system. Some bidders cleverly chose a relatively cheap item to place in multiple bids thereby making these bids mutually exclusive, Ledyard et al. (2002). There have been a number of alternative bidding languages that have been proposed; see Fujishima et al. (1999), Nisan (2000), Boutilier and Hoos (2001), and Boutilier et al. (2001) for descriptions of alternative languages.

One serious flaw in a first-price sealed-bid design is that the bidder can experience what is referred to as the winner's curse, i.e., the winning bidder may pay more than was necessary to win since the second highest bid price was far less than the winning bid amount. For this reason, sealed-bid first price auctions encourage bidders to shave some amount off of the bid price. From a game-theoretic perspective, one wants an auction design that encourages straight-forward honest bidding.

An alternative that overcomes this problem is the second price (sealed bid) auction whereby the bidder that has submitted the highest bid is awarded the object (package), but the bidder pays only slightly more (or the same amount) as that bid by the second-highest bidder. In second price auctions with statistically independent private valuations, each bidder has a dominant strategy to bid exactly his valuation. The second price auction also is often called a Vickrey auction (1961).

In a second-price auction, one solves the same winner determination problem as one does for the first-price sealed-bid case, but the winners do not necessarily pay what they bid. Instead, one determines the marginal value to the seller of having this bidder participate in the auction. To do this, for each winning bidder, one calculates the revenue that the seller would receive when that bidder participates in the auction and when that bidder does not, i.e. when none of the bids of this bidder are in the winner determination problem. The difference in the two objective function values is known as the Vickrey-Clarke-Groves discount, named after the three authors, Vickrey (1961), Clarke (1971), and Groves (1973). Each of these authors wrote separate papers producing certain attributes that this auction design has as it relates to incentivizing bidders to reveal their truth value of the goods demanded, and the bidder pays the bid price minus the discount. When winners pay this amount, the auction is known as the Vickrey-Clarke-Groves (VCG) Mechanism.

Although it can be shown that the VCG mechanism encourages truthful bidding, it is almost never used in practice. For a complete list of reasons for it being impractical, see Ausubel and Milgrom (2006) and Rothkopf (2007). In essence, the prices provided by this mechanism may be very low. Worse yet, when items have complementary values, i.e. the package is worth more to the bidder than the sum of the values of the individual items, the outcome may price the items so low that there is a coalition of bidders that would prefer to renege on the auction and negotiate privately with the seller, and the seller may respond by reneging on the sale since both the seller and the coalition of buyers will be better off. Ausubel and Milgrom (2002) argue that prices should be set high enough so that no such coalitions exist. In game theoretic terms, the prices are set such that the outcome is in the core of a coalitional game. These authors introduced an auction design known as the ascending proxy auction in which the bidders provide all bids as if in a sealed-bid auction. Each bidder is provided with a proxy that bids for the bidder in a straightforward manner during an ascending auction. The proxy only announces bids to the auctioneer that maximize the bidder's profit, (i.e. bid price minus announced price) in any given round. The auction continues as an ascending package-bidding auction until, in some round, there are no new bids. Thus, the auction simulates, through proxy bidders, an ascending auction where the increment in each round is infinitesimally small and each bidder, through the use of its proxy, bids in a straight-forward manner.

This auction design is very similar to the iBundle design of Parkes and Ungar (2000).

Hoffman et al. (2005) provide a computational approach toward speeding up the calculations associated with this proxy auction design, and Day and Raghavan (2007) provide an elegant mechanism to obtain minimal core prices directly. The direct mechanism of Day and Raghavan sequentially solves winner determination problems to determine losing coalitions that could supply more revenue to the seller at the current prices. When the solution to this optimization problem yields revenue greater than what the VCG mechanism would provide, the prices of the winning bid set are raised so that the total price paid by winning bidders is equal to this new revenue. To determine these new prices, one must be sure that any winning bidder that forms part of this blocking coalition does not have its price raised from its prior price since it would not be willing to join a coalition if it were to lose revenue relative to its prior offer by the seller. The algorithm is an iterative cutting plane algorithm that forces the prices higher at each iteration until one can find no coalition that can increase revenue to the seller. Therefore, the algorithm finds prices for each winning bidder that are in the core. Since there may be many such minimum core prices, Day and Milgrom (2008) suggest that, in order to encourage sincere bidding, one choose the minimum core prices that are closest in Euclidean distance from the VCG prices. Alternatively, Erdil et al. (2009) argue for a different set of minimum core prices that are based "on a class of 'reference rules' in which bidders' payments are, roughly speaking, determined independently of their own bids as far as possible."

These core-selecting second-price sealed-bid mechanisms have the following properties: They are in the core, they eliminate the exposure problem, and they encourage bidders to bid sincerely. As with all sealed-bid auctions, they make collusion and punishment for not adhering to tacit agreements extremely difficult.

There are, however, negatives associated with this auction, as well as for all sealed-bid auction designs, in that it puts a significant burden on the bidders. Each bidder needs to assess, for every possible combination of items, whether it is a package of interest and then, for all such packages, determine the maximum it is willing to pay. In addition, such mechanisms do not provide any information about how the packages submitted might fit with packages submitted by other bidders. To overcome these problems, a number of authors have suggested simultaneous ascending combinatorial auction designs that allow users price information during the auction.

Multi-round Auctions

Often the value of the good or package of goods being auctioned is not completely known and/or private. Instead, there is a common component to the bid value, that is, the value of the item is not independent of the other bidders, but rather there is a common underlying value as well. In such situations, each agent has partial information about the value. Many high-stakes auctions, such as government auctions for spectrum, oil exploration, and land use, fall into this class. In the case of package-bidding auctions, when there is a common component and bidders want to assess how much others are willing to pay for that item or package of items, the auction is usually an ascending auction with multiple rounds. A round consists of a given time period where bidders have the opportunity to submit new bids. When the round ends, all bids are collected and the winner determination problem is solved. This optimization problem determines the packages that provide the seller with the maximum revenue. The bids that are in the winning set are labeled "provisionally winning," i.e. they would be winning if the auction ended in this round. Thus, in an ascending combinatorial auction, all items are sold simultaneously and a bidder can bid on any collection of items in a given round. To overcome the current set of provisionally winning bids, a bidder must submit a bid that increases the total revenue to the seller.

There are a number of design question that must be answered to have a complete combinatorial auction design:

1. How does the auction end?
2. Must bidders participate in every round?
3. Are bids from previous rounds part of the bids considered by the winner determination problem?
4. How are the prices set in each round?
5. What do bidders know about the bids of other bidders?
6. What other rules might be necessary to ensure that collusion is avoided, to make reneging costly, and to encourage bidders to act truthfully?

Of importance is how to assure that the auction ends in a reasonable period of time and that price discovery (the main reason for a multi-round auction) is accomplished. Most package-bidding auctions have discrete time periods, called rounds, and in each round, the auctioneer provides a price to the user that is the minimum price that the bidder must supply in order to place a new bid. One can choose either a fixed stopping rule or a stopping rule that is determined dynamically. A fixed time stopping rule specifies that the auction will end at a given time. With a fixed stopping time, bidders are encouraged to not provide any bids until the very last seconds of the auction, called sniping. The purpose of sniping is to give other bidders no chance of responding to an offer. In this way, a bidder can acquire price information from other bidders but does not reciprocate, since throughout most of the auction, the bidder is silent. If all bidders chose to snipe and provide no bids until the end of the auction, the auction essentially becomes a first-price sealed-bid auction. To overcome the problem of sniping and to encourage price discovery, most package bidding auctions use an alternative stopping criteria whereby the auction ends when no new bids are presented within a round.

Often, for high-stakes multi-round auctions, there are also activity rules that require a bidder to bid in a consistent way throughout the auction. Activity rules force bidders to maintain a minimum level of bidding activity to preserve their eligibility to bid in the future. Thus, a bidder desiring a large quantity at the end of the auction (when prices are high) must bid for a large quantity early in the auction (when prices are low). If the bidder cannot afford to bid on a sufficient number of items to maintain current eligibility, then eligibility will be reduced so that it is consistent with current bidding. Once eligibility is decreased, it can never be increased. As the auction progresses, the activity requirement increases, reducing a bidder's flexibility. The lower activity requirement early in the auction gives the bidder greater flexibility in shifting among packages early on when there is the most uncertainty about what will be obtainable. Precisely how the activity and eligibility rules are set matters and must be depend upon the type of auction – the value of the items being auctioned, the projected length of the auction, the number of participants, etc. In many high-stakes auctions, such as spectrum or electricity, these activity rules have proven highly successful, Klemperer (2002), McMillan (2002), and Milgrom (2004).

In an ascending multi-round auction design, the auctioneer must provide information about the current value of each package. This information is used for two related purposes: (1) to specify the minimum bid for each item or package in the next round and (2) to provide valuation information to bidders so that they can determine what might be required for a bid to be winning in a subsequent round. While pricing information is easy to ascertain in single item auctions or in simultaneous multi-round auctions without package bidding, (i.e. where bids can be placed on only single items), pricing information for combinatorial auctions is not well defined. Bidders provide only aggregate package prices without providing the information about how each of the individual components that made up the bundle contributes to the overall price. Attempting to disaggregate these bundles into single item prices unambiguously is not possible. Also, since there are many ways that some bundle might partner with other packages to create a winning set, determining the minimal cost partnering for a given package by a given bidder is a complex problem.

To further complicate the pricing issue, bidders may view certain items as substitutes and other items as complements. In the case where items are substitutes, bidders are likely to express sub-additive values for their packages. That is, the value of a package of items is less than or equal to the sum of the values of the items that make up the package. In the complementary case, bidders are likely to express super-additive values for packages. In this case, the value of a package of items is greater than or equal to the sum of the values of the items that make up the package. When items can be both substitutes and complements for bidders, providing unambiguous, complete and accurate price information is an unsolved problem. The non-convex nature of the problem means that the linear prices (i.e. the sum of a package is equal to the sum of the individual items that make up the package) that can be obtained from dual prices from the linear relaxation of the WDP problem will overestimate the true values of the items. In most auctions, one adjusts the dual prices so that the prices are modified so that when one sums the items in each of the winning packages, the prices on those packages exactly equal the prices bid by the provisionally winning bidders (i.e. the winners at the end of the current round). Rassenti et al. (1982) terms these prices pseudo-dual prices.

(For theoretical issues with duals associated with non-convex problems see Wolsey(1981), and for non-anonymous non-linear prices see deVries and Vohra (2003) and Bikhchandani and Ostroy (2002).

Although linear pricing cannot accommodate all aspects of the pricing associated with the non-linear, non-convex, winner determination problem, there are still good reasons for considering its use for determining future bid requirements. First, even perfect pricing is only correct when all other aspects of the problem remain fixed, i.e. when bid amounts remain the same on all other bids and when no new bids are submitted. Second, a dual price associated with a given constraint is only correct when one changes this single restriction (the right-hand-side of the associated constraint) by a very small amount. In the case of combinatorial auctions, the item is either won or it is not. Changes to a constraint would either remove the item entirely from consideration or create a second identical item. Thus, even non-linear, non-anonymous pricing has serious limitations in the context of the winner determination problem since the removal of a single item from the auction (e.g. the removal of the New York City market from consideration in a nationwide spectrum auction) may change the willingness of bidders to participate.

Finally, in an ascending bid auction, bidders need pricing information that is easy to use and understand, and is perceived to be fair. In this situation, easy to use means that bidders can quickly compute the price of any package, whether or not it had been previously bid. Often, bidders want to know what it would take for such a bid to be competitive, i.e. have some possibility of winning in the next round. Bidders may also perceive such prices to be fair since all bidders must act on the same information. Linear prices are likely to move the auction along and deter such gaming strategies as parking (parking is an approach whereby the bidder bids on packages that currently have very low prices knowing that these packages have a very low probability of winning). Bidding on such low-priced packages allows a bidder to maintain eligibility (by maintaining activity), while hiding interest in the packages that are really desired until later in the auction). Thus, virtually all ascending combinatorial auctions use pseudo-dual pricing. For more on alternative pricing within this general framework and the testing thereof, see (Dunford et al. (2003), Bichler et al. (2009) and Brunner et al. (2011).

In 1999, DeMartini et al. proposed an auction design labeled The Resource Allocation Design or RAD where the WDP is solved each round and all losing bidders can only bid on packages where the package price is the sum of the pseudo-dual prices plus some increment (as announced by the auctioneer). There is no activity rule for this auction design. In 2002, the Federal Communications Commission (FCC) announced a similar package bidding design but proposed refinements to the pseudo-price calculations that attempts to limit fluctuations (both positive and negative) in prices. A related design was proposed by Bichler et al. (2009) and is called the Approximate Linear Pricing Scheme (ALPS). It also uses similar rules but chooses the ask price to better balance prices across items. Note that all of these pricing procedures allow prices to both increase and decrease depending upon the packages that are in the winning set. In virtually all of these designs, any bid submitted in any round is considered active throughout the auction. This rule works well with the XOR language since only one bid of a bidder can be in an optimal set and bidders should be willing to win bids placed in early rounds of the auction, when prices were low. This rule forces bidders to provide sincere bids throughout the auction.

A very different ascending package bidding design was proposed by Porter et al. (2003). It is called the combinatorial clock auction. In this design, the auctioneer provides prices for each unique good (if there are multiple identical items, then the bidder indicates that number of units of that item they desire) based solely on whether there is more demand for the item than for supply; no WDP problem is solved. There is no concept of a provisionally winning bidder. Instead, prices increase whenever demand for a given item is greater than supply. Bidders indicate the single package bid that is best given the per-unit prices announced by the auctioneer. All bidders must rebid on any item that they wish to procure in each round. The only information provided to bidders at the end of each round is the quantity demanded for each item and the price for the next round. As long as demand exceeds supply for at least one item, the price is increased for those items with excess demand. If there are no new bids in a round and supply equals demand, then the auction ends. However, it may happen that when there are no new bids, demand has been reduced to below supply. If this occurs, a WDP is solved using all bids from all rounds. If the computed

prices do not displace any bids from the last round, then the auction ends. Otherwise, the auction resumes with the prices determined by using the pseudo-prices calculated from the WDP. Thus, for most rounds, the computation has been drastically reduced to merely increasing prices by a given increment. Only, when demand has dropped below supply is the WDP solved.

Other approaches are the auction designs that simplify the problem by only allowing a few pre-defined packages (Harstad et al. 1998) for which the WDP is polynomially solvable. This idea of only allowing a certain pre-determined set of packages (called hierarchical packages, Goeree and Holt 2010) was used in the 2009 FCC auction for broadband spectrum that brought over $19B into the U.S. Treasury. In that design, all bids were additive (the OR language applied) and the WDP was solved in linear time. When it is possible, in advance, to understand the needs of the bidders and when the packages most desired can be represented in a hierarchical fashion, then one obtains an auction design that is both simpler and quite efficient. However, if the demand for packages does not take on this hierarchical structure, then imposing such structure on the problem for the sake of computability will likely lead to less efficient outcomes.

Hybrid Designs

Ausubel et al. (2005) have argued for a hybrid design that reduces the computational burden on both the bidder and the auctioneer. Here, one first uses a combinatorial clock design followed by a last round second-price sealed-bid approach. The combinatorial clock is similar to that proposed by Porter et al. (2003) with the further enhancement that bidders who find the increment too high are able to place a bid at a price between the old price and the new price that indicates the maximum amount the bidder is willing to pay for that combination of items. In this way, the efficiency loss due to increment size is lessened. This phase of the auction ends when demand is less than or equal to supply or when demand on most items has trailed off. When demand does not exactly equal supply on all items, a sealed-bid phase is initiated. Here, the ascending proxy auction of Ausubel and Milgrom (2002) is imposed. When these two auction designs are merged, one must be careful that the activity rules work well for both phases of the auction. One wants tight activity rules in the ascending phase of the auction to ensure that the bidders are forced to bid sincerely. However, these rules may need to be relaxed or altered during the final sealed-bid phase or a straightforward bidder may be precluded from providing all of the packages that bidder values during the sealed-bid round. Also, theory dictates that in order to guaranteed an efficient outcome, losing bidders (i.e. bidders who dropped out prior to the final phase) must also provide all of the bids that they value in the final phase. Thus, although this hybrid auction is promising in that it is likely to speed up combinatorial auctions, research is still necessary to better understand how the rules of these two disparate auctions should be set so that they mesh well. For more on testing of this design, see Bichler et al. (2011).

Complexity of Combinatorial Auctions

As the previous discussion illustrates, most combinatorial auction designs requires considerable computation and most of the computational burden falls to the auctioneer. This seems appropriate since the auctioneer wants an auction that allows much participation; bidders should not be required to understand combinatorial optimization in order to participate. In terms of these computations, commercial software, such as CPLEX, GUROBI, or XPRESS have shown their ability to solve such problems in reasonable times (less than 30 minutes). Thus, although there is much in the literature that argues against combinatorial auctions because of the computational burden, the optimization software has proven up to be capable of handling the problems that are currently being considered applicable for this type of auction. For more on the computational issues in computing winner determination problems, see Leyton-Brown et al. (2005) and Bichler et al. (2009).

Since multi-item auctions are complex and require bidders to consider multiple alternative bid options, it is important that the computer software used for communication between the bidder and the auctioneer be easy to use and understand. Good graphical user interfaces help bidders to feel comfortable that they understand the current state of the auction (they have been able to find the current

price information, the items they are winning, the amount of bidding necessary to remain eligible, their dollar exposure based on what they have bid, etc.). The system must also provide easy ways for bidders to input their next moves and confirm that they have provided the system with the correct information. As the use of auctions is spreading, computer interfaces for such processes continue to improve and to provide better ways of displaying information to the users through charts, graphs and pictures. There is likely to be continued improvement in this area.

These tools do not, however, help the bidder determine the optimal combination of items to bundle as a package and the optimal number of packages to supply to the system. Since bidders face the serious problem of determining which bids are most likely to win at prices that are within the their budgets, tools that assist bidders in understanding the state of the auction is important. In both supply-chain auctions and in high-stakes government auctions (such as spectrum auctions), bidder-aided tools are often developed to assist the bidder in determining the package or packages to submit in any given round. In the case of supply-chain auctions, the auctioneer often suggests packages to the suppliers that will fit well with other bidder's bids (e.g. by either adding or removing a single item from the package, or by considering a quantity discount for supplying more of an item). Such tools have been found to be very useful and also computationally tractable; see An et al. (2005), Dunford et al. (2003), and Boutilier et al. (2004). Day and Raghavan (2005) and Parkes (2005) provide alternative ways for bidders to express preferences that do not require that the bidder specify particular packages to the auctioneer.

Applications of Combinatorial Auctions

There are many examples of governments' using auctions for the allocation of valuable assets. In most of these auctions, the government is allocating a good and uses auctions to determine both the price and the allocation. Since 1994, governments throughout the world have been using simultaneous multi-round auctions for the allocation of spectrum. For spectrum, a government has the goal of allocating the good to the entities that value it the most with the hope that the bid cost will encourage the build-out of the services. To assure that there is sufficient competition in the telecommunications industry, the U.S. government has, in the past, set spectrum caps for each region. These auctions have been copied globally and are now the standard way that spectrum is allocated. Recently, a number of different package-bidding designs are being tried including the hierarchical ascending auction, the combinatorial clock auction, or the clock-proxy design. As of 2005, these auctions have resulted in revenues in excess of $200 billion dollars worldwide (Cramton 2005).

Within the power industry, there has also been an evolutionary movement toward auctions for the determination of who can supply power to the electricity grid and at what price. Most of the allocation is determined one day ahead of the demand. The auction reflects the unique characteristics (both physical and structural) of the industry. The allocation is determined by a complicated optimization that evaluates the demands at various nodes of the networks and prices power generation at each such node. The spot market corrects this allocation for any last minute changes due to weather, plant outages, etc. Long term contracts make this process work.

Similarly, auctions have been used to bring market-based forces to control air pollution. Here, a government entity (either nationally run or regionally administered) establishes a fixed number of tradable allowances each of which represents the legal right for its owners to emit a fixed quantity of pollution. A firm holding an allowance can emit the fixed quantity and surrender the allowance to the government, or if the firm can abate its emissions, it can profit by selling the allowance to another polluter than cannot so inexpensively abate emissions. The establishment of the fixed quantity is the cap. The exchange of allowances (credits) between polluters is the trade. See Ellerman et al. (2003) and Tietenberg (2006) for a general overview of cap and trade ideas.

The use of combinatorial auctions for the procurement of goods in services has also been growing. Some of these auctions are sealed-bid auctions, while most are moving toward multi-round auction designs. In such auctions, the providers of the goods and services are pre-screened and are then allowed to provide bids for collections of good and/or services as all or nothing packages. For a general

survey of supply-chain auctions, see Bichler et al. (2006). The three applications described next highlight a few examples to show how such auctions differ from government auctions.

1. The first use of a combinatorial auction within the transportation industry was an auction conducted by Sears. Here, suppliers of freight delivery were allowed to bundle multiple lanes together into a single bid thereby allowing carriers to coordinate multiple businesses and reduce empty or low value backhaul movements. It also provided a means to incorporate surge demand contingencies into the longer (3-year) contracts, thereby lessening the need to renegotiate contracts whenever demands changed; Ledyard et al. (2002).
2. Mars Incorporated used a combinatorial auction mechanism to procure the necessary goods from multiple suppliers allowing bidders to specify complex bid structures that indicated quantity discounts, minimum supply, and multiple goods collected within a single bid. No bidder was allowed to supply more than a certain percentage of the overall quantity needed and newer suppliers were limited more severely than their suppliers they had used over a number of years. The algorithm also assured that there were multiple suppliers in the solution for each critical entity. These auctions are not simple, but work to match the needs of the procurer, Mars, with the capabilities of the suppliers (often farmers). The allocation considers geographic, volume and quality factors. The suppliers liked the auction mechanism because of its transparency, shorter negotiation time and fairness; Honer et al. (2003).
3. Motorola Corporation used auctions for the procurement of the multitude of parts needed for cellular devices. Motorola needed to reduce both the time and the effort required to prepare for and conduct negotiations with its suppliers, simplify their coordination, and optimize contract awards across sectors, in order to save costs; Metty et al. (2005).

Governments are moving toward procuring their goods and services in a similar fashion. One such example is the use of auctions to determine the suppliers of lunches in a large school system. Chile spends around US$180 million a year to feed 1,300,000 students from low income families. To improve the quality of the goods and services being provided to the school system and to save money, the government chose to assign catering contracts in a single-round sealed-bid combinational auction. This auction resulted in a transparent and objective allocation approach, thereby generating competition among firms. It also allowed the companies to build flexible territorial bids to include their scale of economies, leading to more efficient resource allocation. This new methodology improved the price-quality ratio of the meals with yearly savings of around US$40 million, equivalent to the cost of feeding 300,000 children during one year; Epstein et al. (2002).

In supply-chain auctions, rules are designed to assure a certain diversification in suppliers and to assure the reliability of the supply chain. In each case, are goals other than revenue maximization or efficiency that drove the auction design. In addition, the auction design must consider the nature of the investment. For spectrum, where there was both uncertainty in the long-term use of the technologies and where the cost of build-out are high, long-term leases were chosen. For energy, auctions are used for a much shorter decision problem. The U. S. Treasury uses multiple auctions for short, medium and long-term debt allocation. Oil and gas exploration must have a relatively long-term horizon where payments for wildcatting are based on the bid price and a yearly rent, whereas payments for extraction are based on bid price and royalties.

Thus, one must consider carefully the application when designing the allocation mechanism and the payment scheme. Auction theory and its use is growing because of its proven value. It provides price discovery and signals where more capacity is needed. It is often a fairer and more transparent process for the allocation of goods and services.

Concluding Remarks

Combinatorial auctions are appropriate for problems where the bidders need to procure a collection of items that contribute to their having a viable business plan. When evaluating alternative designs, one is likely to want to satisfy the following goals:

1. The property rights are well-defined.
2. Bidders are able to, through their bids, announce the entire collection of objects that they need for a given business plan.

3. The auction results in maximum revenue to the seller.
4. The auction results in an efficient outcome i.e. all items are collectively allocated to the bidders that value these items the most.
5. The auction is perceived as fair to all bidders.
6. The auction ends in a reasonable amount of time.
7. The auction has limited transaction costs, i.e. the rules are not so difficult or the bidding so complicated that a straightforward bidder finds it difficult to participate.
8. The auction cannot be gamed, i.e. truthful bidding is an optimal strategy for all bidders.
9. The auction allows price discovery.
10. The auction is computationally feasible and scalable.

It is not possible to have all such attributes obtain simultaneously. For each applications, some of these goals will be more important than others. One should, however, keep all of these goals in mind when evaluating a mechanism.

In addition, the auction mechanism should consider any application-specific issues that might arise. For example, in government auctions one might want to consider how market power impacts the outcome, whether there will be sufficient participation, and whether the outcome will limit future competition in the industry. In certain situations, there may need to be a transition period that allows the market to adjust to a change in the way rights are allocated; One may have to consider the associated rights that a bidder would need to be able to use the right being sold or leased in the auction; The seller needs to determine if the rights are paid for over time or at the end of the auction; The money obtained may need to be designated for a specific use in order for the government to obtain the approval of all constituents. The auction design may also need to satisfy other social goals specific to the application (e.g. reducing emissions, increasing competition, incentivizing innovation, improving multi-modal transportation). Similarly, in supply chain auctions, a variety of goals need to be considered– quality of the goods, price, historical dependability of the supplier, among others.

See

► Auction and Bidding Models
► Integer and Combinatorial Optimization

References

An, N., Elmaghraby, W. J., & Keskinocak, P. (2005). Bidding strategies and their impact on auctioneer's revenue in combinatorial auctions. *Journal of Revenue and Pricing Management, 3*(4), 337–357.

Ausubel, L. M., Cramton, P., & Milgrom, P. (2005). The clock proxy auction. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions* (pp. 113–136). MIT Press.

Ausubel, L. M., & Milgrom, P. (2002). Ascending auctions with package bidding. *Frontiers of Theoretical Economics, 1*, 1–42.

Ausubel, L. M., & Milgrom, P. (2006). The lovely but lonely Vickrey auction. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions*. Cambridge, MA: MIT Press.

Bichler, M. (2011). Auctions: Complexity and algorithms. In *Wiley encyclopedia of operations research and management science*. John Wiley and Sons.

Bichler, M., Shabalin, P., & Wolf, J. (2011). *Efficiency, auction revenue, and bidding behavior in the combinatorial clock auction*. Technical Report available from M. Bichler.

Bichler, M., Davenport, A., Hohner, G., & Kalagnanam, J. (2006). Industrial procurement auctions. In P. Crampton, Y. Shoam, & R. Steinberg (Eds.), *Combinatorial auctions* (pp. 593–612). MIT Press.

Bichler, M., Shabalin, S., & Pikovsky, A. (2009). A computational analysis of linear price iterative combinatorial auction formats. *Information Systems Research, 20*(1), 33–59.

Bikhchandani, S., DeVries, S., Schummer, J., & Vohra, R. (2002). Linear programming and Vickrey auctions. In B. Dietrich & R. Vohra (Eds.), *Mathematics of the internet: E-auctions and markets* (pp. 75–115).

Bikhchandani, S., & Ostroy, J. M. (2002). The package assignment model. *Journal of Economic Theory, 107*, 337–406.

Boutilier, C., & Hoos, H. H. (2001). Bidding languages for combinatorial auctions. *Seventh International Joint Conference on Artificial Intelligence (IICAI-01)*, 1211–1217.

Boutilier, C., Sandholm, T., & Shields, R. (2004). Eliciting bid taker non-price preferences in "Combinatorial Auctions". In V. Khu-Smith & C. J. Mitchell (Eds.), *Proceedings of the national conference on artificial intelligence* (pp. 204–211). San Jose, CA.

Brunner, C., Goeree, J. K., Holt, C. H., & Ledyard, J. O. (2011). An experimental test of flexible combinatorial spectrum auction formats. *American Economic Journal: Microeconomics, 2*, 39–57.

Cason, T. N. (1993). Seller incentive properties of EPA's emission trading auction. *Journal of Environmental Economics and Management, 25*, 177–195.

Clarke, E. (1971). Multipart pricing of public goods. *Public Choice, 8*, 19–33.

Cramton, P. (2005). Simultaneous ascending auctions. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions* (pp. 99–114). MIT Press.

Cramton, P., Shoham, Y., & Steinberg, R. (Eds.). (2005). *Combinatorial auctions* (pp. 99–114). MIT Press.

Day, R., & Milgrom, P. (2008). Core-selecting package auctions. *International Journal of Game Theory, 36*(3), 393–407. Springer.

Day, R. W., & Raghavan, S. (2007). Fair payments for efficient allocations in public sector combinatorial auctions. *Management Science, 53*(9), 1389–1406.

Day, R., & Raghavan, S. (2005). *Assignment preferences and combinatorial auctions*. Working paper, Operations and information management school of business, University of Connecticut. http://users.business.uconn.edu/bday/index.htm

DeMartini, C., Kwasnica, A. M., Ledyard, J. O., & Porter, D. (1999). *A new and improved design for multi-object iterative auctions*, Social Working Paper. Pasadena, CA: Division of the Humanities and Social Sciences, California Institute of Technology.

DeVries, S., & Vohra, R. (2003). Combinatorial auctions: A survey. *INFORMS Journal on Computing, 15*(3), 284–309.

Dunford, M., Hoffman, K., Menon, D., Sultana, R., & Wilson, T. (2003). *Price estimates in ascending combinatorial auctions*, Technical Report. Fairfax, VA: George Mason University, Systems Engineering and Operations Research Department.

Ellerman, A. D., Joskow, P. L., Montero, J., Schmalensee, R., & Bailey, E. M. (2000). *Markets for clean air: The U.S. acid rain program*. Cambridge University Press.

Ellerman, A. D., David, H., & Paul L. J. (2003). *Emissions Trading: Experience, Lessons, and Considerations for Greenhouse Gases*. Washington, D.C.: Pew Center for Global Climate Change.

Epstein, R., Henriquez, L., Catalan, J., Weintraub, G., & Martinez, C. (2002). A combinatorial auction improves school meals in Chile. *Interfaces, 32*(6), 1–14.

Erdil, A., Klemperer, P., Cramton, P., Dijkstra, G., Goeree, J., Marszalec, D., Meyer, M., Milgrom, P., Pagnozzi, M., & Parkes, D. C. (2009). *A new payment rule for core-selecting package auctions*. Technical report available on Paul Klemperer's website.

Friedman, D., & Rust, J. (Eds.). (1993). *The double auction market: Institutions, theories and evidence* (Santa Fe Institute studies in the sciences of complexity, Vol. XIV). Addison Wesley.

Fujishima, Y., Leyton-Brown, K., & Shoham, Y. (1999). Taming the computational complexity of combinatorial auctions: Optimal and approximate approaches. *Proceedings of IJCAI 1999*, 548–553.

Goeree, J. K., & Holt, C. A. (2010). Hierarchical package bidding: A paper & pencil combinatorial auction. *Games and Economic Behavior, 70*(1), 146–169.

Groves, T. (1973). Incentives in teams. *Econometrica, 41*, 617–631.

Harstad, R., Pekec, A., & Rothkopf, M. H. (1998). Computationally manageable combinatorial auctions. *Management Science, 44*, 1131–1147.

Hoffman, K., Menon, D., van den Heever, S. A., & Wilson, T. (2005). Observations and near-direct implementations of the ascending proxy auction. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions* (pp. 415–450). MIT Press.

Hoffman, K., & Menon, D. (2010). A practical combinatorial clock exchange for spectrum licenses. *Decision Analysis, 7*(1), 58–77.

Hoffman, K., Menon, D., & van Den Heever, S. A. (2008). A package bidding tool for the FCC's spectrum auctions and its effect on auction outcomes. *Telecommunications Modeling Policy and Technology: Operations Research/ Computer Sciences Interfaces Series, 44*, 153–189.

Holt, C. A., Shobe, W., Burtraw, D., Palmer, K., & Goeree, J. (2007). *Auction design for selling CO_2 emission allowances under the regional greenhouse gas initiative. Regional greenhouse gas initiative*. Technical Report to RGGI.

Honer, G., Rich, J., Ng, E., Reid, G., Davenport, A., Kalagnanam, J., Lee, H. S., & An, C. (2003). Combinatorial and quantity discount procurement auctions benefit mars, incorporated and its suppliers. *Interfaces, 33*(1), 23–35.

Klemperer, P. (1999). Auction theory: A guide to the literature. *Journal of Economic Surveys, 13*(3), 227–286.

Klemperer, P. (2002). What really matters in auction design. *Journal of Economic Perspectives, 16*, 169–189.

Klemperer, P. (2004). *Auctions: Theory and practice* (The toulouse lectures in economics). Princeton, NJ: Princeton University Press.

Koboldt, C., Maldoom, D., & Marsden, R. (2003). *The first combinatorial spectrum auction*. Ofcom Technical Report describing the results of the 2003 Nigerian Spectrum Auction, available on of com website.

Krishna, V. J. (2002). *Auction theory*. Academic Press, 200pp.

Kwasnica, A. M., Ledyard, J. O., Porter, D., & DeMartini, C. (2005). A new and improved design for multi-object iterative auctions. *Management Science, 51*, 419–4234.

Ledyard, J. O., Olson, M., Porter, D., Swanson, J. A., & Torma, D. P. (2002) The first use of a combined value auction for transportation services. *Interfaces, 32*, 4–12.

Lehmann, D., Mueller, R., & Sandholm, T. (2005). The winner determination problem. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions*. Cambridge, MA: MIT Press.

Leyton-Brown, K., Nudelman, E., & Shoham, Y. (2005). Empirical hardness models. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions* (pp. 479–503). MIT Press.

McMillan, J. (2002). *Reinventing the bazaar: A natural history of markets*. Norton Press, 278pp.

Metty, T., Harlan, R., Samelson, Q., Moore, T., Morris, T., & Sorenson, R. (2005). Reinventing the supplier negotiation process at motorola. *Interfaces, 35*(1), 7–23.

Milgrom, P. (2004). *Putting auction theory to work*. Cambridge Press, 368pp.

Milgrom, P. (2007). Package auctions and exchanges. *Econometrica, 75*(4), 935–965.

Nisan, N. (2000). Bidding and allocation in combinatorial auctions. *Proceedings of the 2nd ACM Conference on Electronic Commerce*, 1–12.

O'Neill, R. P., Helman, U., Hobbs, B., Stewart, W. R., & Rothkopf, M. (2007). The joint energy and transmission rights auction: A general framework for RTO market designs. *Power Engineering Review, IEEE, 22*(10), 59–68.

Parkes, D. C. (2005). Auction design with costly preference elicitation. *Annals of Mathematics and AI, 44*, 269–302.

Parkes, D. C., Kalagnanam, J., & Eso, M., (2001). Achieving budget-balance with Vickrey-based payment schemes in combinatorial exchanges. *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI-01)*, 1161–1168.

Parkes, D. C., & Ungar, L. H. (2000). Iterative combinatorial auctions: Theory and practice. *Proceedings of the 17th National Conference on Artificial Intelligence* (AAAI-00), 74–81.
Pekec, A., & Rothkopf, M. H. (2006). Non-computational approaches to mitigating computational problems in combinatorial auctions. In P. Cramton, Y. Shoham, & R. Steinberg (Eds.), *Combinatorial auctions*. M.I.T. Press.
Porter, D., Rassenti, S., Roopnarine, A., & Smith, V. (2003). Combinatorial auction design. *Proceedings of the National Academy of Sciences, 100*(19), 11153–11157.
Porter, D., & Smith, V. (2006). FCC license experiment design: A 12-year experiment. *Journal of Law Economics and Policy, 3*, 63–80.
Rassenti, S., Smith, V., & Bulfin, R. I. (1982). A combinatorial mechanism for airport time slot allocation. *Bell Journal of Economics, 13*, 402–417.
Rothkopf, M. H. (2007). Thirteen reasons why the Vickrey-Clarke-Groves process is not practical. *Operations Research, 55*(2), 191–197.
Steiglitz, K. (2007). *Snipers, shills and sharks: eBay and human behavior*. Princeton University Press, 298pp.
Tietenberg, T. H. (2006). *Emissions trading: Principles and practice* (2nd ed.). Washington: RFF Press.
Vickrey, W. (1961). Counter-speculation, auctions and competitive sealed tenders. *Journal of Finance, 16*, 8–37.
Wolsey, L. A. (1981). Integer programming duality: Price functions and sensitivity analysis. *Mathematical Programming, 20*(1), 173–195.
Wurman, P. R., & Wellman, M. P. (1999). *Equilibrium prices in bundle auctions*, Sante Fe Institute Working Papers (Paper: 99-09-064).

Combinatorial Explosion

The phenomenon associated with optimization problems whose computational difficulty increases exponentially with the size of the problem. One common paradigm is the traveling salesman problem.

See

- ▶ Combinatorics
- ▶ Curse of Dimensionality
- ▶ Integer and Combinatorial Optimization
- ▶ Traveling Salesman Problem

Combinatorial Optimization

- ▶ Integer and Combinatorial Optimization

Combinatorics

Eugene L. Lawler

Combinatorics is the branch of mathematics that deals with arrangements of objects, usually finite in number. The term arrangement encompasses, among other possibilities, selection, grouping, combination, ordering or placement, subject to various constraints.

Elementary combinatorial theory concerns permutations and combinations. For example, the number of permutations or orderings of n objects is $n! = n(n-1) \ldots (2)(1)$, and the number of combinations of n objects taken k at a time is given by the binomial coefficient $\binom{n}{k} = n!/[k!(n-k)!]$. In order to compute the probability of throwing a seven with two dice, or of drawing an inside straight at poker, one must be able to count permutations and combinations, as well as other types of arrangements. Indeed, combinatorics is said to have originated with investigations of games of chance. Combinatorial counting theory is the foundation of discrete probability theory as is exists today.

Experimental design provides the motivation for another classic area of combinatorial theory. Suppose five products are to be tested by five experimental subjects over a period of 5 days, with each subject testing one product per day. Labeling the subjects A, B, C, D, E, the products 1,2,3,4,5, and the days M, Tu, W, Th, F, one way to schedule the tests is as follows:

	M	*Tu*	*W*	*Th*	*F*
1	A	B	C	D	E
2	B	C	D	E	A
3	C	D	E	A	B
4	D	E	A	B	C
5	E	A	B	C	D

A square array of symbols, with each symbol occurring in each row exactly once and in each column exactly once, is called a Latin square.

Now suppose each of the tests is to be performed by a subject in the presence of an observer. In order to reduce the effects of bias due to subject-observer interactions, the Latin square should represent the

schedule for the subjects to be combinatorially orthogonal to the Latin square for the observers. This means that when the two Latin squares are superimposed, each of the 25 possible subject-observer pairs appears exactly once in the resulting array, called a Graeco-Latin square. Labeling the observers a, b, c, d, e, a 5×5 Graeco-Latin for our experiment is as follows:

Aa	Bb	Cc	Dd	Ee
Bc	Cd	De	Ea	Ab
Ce	Da	Eb	Ac	Bd
Db	Ec	Ad	Be	Ca
Ed	Ae	Ba	Cb	Dc

Leonhard Euler observed that no 2×2 Graeco-Latin square exists and found he was able to construct examples of $n \times n$ Graeco-Latin squares for n up to five, but had trouble with six. In 1782 Euler conjectured the nonexistence of such an arrangement for any $n = 4k + 2$, where k is an integer. About 1900, Euler's conjecture was confirmed, by systematic examination of cases, for $n = 6$. However, his more general conjecture remained unsettled until 1959 when Bose, Shirkhande and Parker exhibited a 22×22 Graeco-Latin square. Shortly after, these same investigators (Euler's Spoilers) demolished what remained of Euler's conjecture by establishing that Graeco-Latin squares do exist for all n other than two and six. Their work made use of results of number theory, a branch of mathematics with which combinatorics exists in happy symbiosis.

Another investigation of Euler turned out to have considerable importance for combinatorial mathematics. In the old city of Königsberg in Eastern Prussia the River Pregel divided into two branches surrounding an island. The river was spanned by seven bridges. It is said that the people of Königsberg entertained themselves by trying to find a route around the city that would cross each of the bridges exactly once. In 1736, Euler provided a definitive answer to the Königsberg bridge problem, and any related instances: "If there are no more than two areas to which an odd number of bridges lead, then such a journey is not possible. If, however, the number of bridges is odd for exactly two areas, then the journey is possible if it starts in either of these areas. If, finally, there are no areas to which an odd number of bridges leads, then the required journey can be accomplished from any area." This result has been viewed as the oldest theorem of what is now known as graph theory.

With the advent of digital computers and operations research, the emphasis of combinatorics shifted from problems of counting and existence of arrangements to problems of optimization. Modern combinatorics may be said to have come of age with the development of network flow theory by Lester Ford and Ray Fulkerson in the 1950s. This remarkable theory enables a great variety of practical optimization problems to be solved by efficient algorithms. A number of elegant duality results follow directly from Ford and Fulkerson's Max-Flow Min-Cut Theorem. For example, consider the König-Egervary Theorem, which can be stated as follows: Let us call a subset of elements of a matrix independent if no two of the elements lie in the same row or the same column. Let all elements be 0 or 1. Then the maximum size of an independent set of 1 s is equal to the minimum number or rows and columns containing all the 1 s in the matrix.

In the 1960s Jack Edmonds generalized many of the results of Ford and Fulkerson by exploiting the concept of a matroid, a combinatorial structure abstracting the notion of linear independence. Edmonds also developed a general theory of matching in graphs, where a matching is a subset of edges, no two of which are incident to the same vertex. He also proved a generalization of the König-Egervary Theorem, which may be viewed as a duality theorem for matchings in the special case of bipartite graphs.

Edmonds (1965) further observed that the running time of his general matching algorithm was bounded by a polynomial in the size of the graph it is applied to, and made an eloquent argument for the goodness of polynomial-time bounded algorithms. The significance of polynomial time bounds came to be more fully appreciated with the development of NP-completeness theory by Stephen Cook, Richard Karp and Leonid Levin in 1973. The theory of NP-completeness has been an essential tool for researchers in combinatorial optimization ever since.

Algorithms arising from network flow theory, matroid optimization theory, matching theory, or similar theories, may all be viewed as special-purpose linear programming algorithms. Combinatorial duality results, including the Max-Flow Min-Cut Theorem and the König-Egervary Theorem, are most often special cases of linear programming duality. The term applied

to the general paradigm of formulating and solving combinatorial problems by linear programming techniques is polyhedral combinatorics.

More often than not, combinatorial optimization problems that arise in the real world are too idiosyncratic and complicated to be fully tamed by polyhedral techniques alone. For these problems, it is usually necessary to engage in some form of enumeration of cases if one seeks to find a provably optimal solution. The Traveling Salesman Problem (TSP) is prototypical of a difficult (NP-complete) problem with a real-world flavor. In this problem, one is asked to find a shortest closed tour of n cities (visiting each city exactly once, and ending at the starting point), given an $n \times n$ matrix of intercity distances. The number of possible tours is, of course, finite: $(n-1)!$. But for any interesting value of n, say 100 or 1,000, the number of tours is so astronomically large as to be effectively infinite. An exhaustive enumeration of even a tiny fraction of the tours is out of the question. Hence if the TSP is to be solved by enumeration, the enumeration must be very artfully limited.

The TSP has served as a testbed for algorithmic research. Indeed, the approaches that have been applied to the TSP are representative of the full range of techniques of combinatorial optimization. These include polyhedral and integer linear programming, Lagrangian relaxation, nondifferentiable optimization, heuristic and approximation algorithms, branch-and-bound, dynamic programming, neighborhood search, and simulated annealing. With much effort by many investigators, it is today possible to find optimal, or provably near-optimal, solutions to instances of the TSP with hundreds, even thousands of cities.

Combinatorial optimization has assumed great practical importance, in such diverse problem areas as machine scheduling and production planning, vehicle routing, plant location, network design, VLSI design, among many others. The practical and theoretical importance of this field can only be expected to grow in the future.

See

- ▶ Chinese Postman Problem
- ▶ Computational Complexity
- ▶ Graph Theory
- ▶ Integer and Combinatorial Optimization
- ▶ Traveling Salesman Problem

References

Biggs, N. L., Lloyd, E. K., & Wilson, R. J. (1976). *Graph theory* (pp. 1736–1936). London, UK: Oxford Univ Press.

Edmonds, J. (1965). Paths, trees, and flowers. *Canadian Journal of Mathematics, 17*, 449–467.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to NP-completeness*. San Francisco: W.H. Freeman.

Graham, R. L., Rothschild, B. L., & Spencer, J. H. (1980). *Ramsey theory*. New York: John Wiley.

Lawler, E. L. (1976). *Combinatorial optimization: Networks and matroids*. New York: Holt, Rinehart and Winston.

Lawler, E. L., Lenstra, J. K., Rinnooy Kan, A. H. G., & Shmoys, D. B. (Eds.). (1985). *The traveling salesman problem: A guided tour of combinatorial optimization*. New York: John Wiley.

Lovasz, L. (1979). *Combinatorial problems and exercises*. Amsterdam: North Holland.

Lovasz, L., & Plummer, M. D. (1986). *Matching theory*. Amsterdam: North Holland.

Nemhauser, G. L., & Wolsey, L. A. (1988). *Integer programming and combinatorial optimization*. New York: John Wiley.

Roberts, F., & Tesman, B. (2009). *Applied combinatorics*. New York: CRC Press.

Schrijver, A. (1986). *Theory of linear and integer programming*. New York: John Wiley.

Tucker, A. (2006). *Applied combinatorics* (5th ed.). New York: John Wiley.

Wilson, R. J., & Watkins, J. J. (1990). *Graphs: An introductory approach*. New York: John Wiley.

Common Random Numbers

- ▶ Simulation of Stochastic Discrete-Event Systems
- ▶ Variance Reduction Techniques in Monte Carlo Methods

Common Value Bidding Model

A bidding model in which the value of what is being auctioned, while unknown at the time of the auction, is known to be the same for all bidders. In such a model, bidders must correct for the selection bias, often called

the winner's curse, caused by the fact that winning bidder is likely to have been the one who most overestimated the value.

See

► Bidding Models

Communications Networks

Edward A. Sykes
Make Systems, Inc., Carey, NC, USA

Introduction

Communications networks are systems of electronic and optical devices that support information exchange among their subscribers. Examples of communications networks are abundant in everyday life: telephone networks, broadcast and cable television networks, and computer communications networks such as the Internet. The impacts of communications networking on the individual, society and the planet are staggering, rivaling that of the tall ship and the automobile. In just under two centuries, humanity has been transformed from myriad villages and towns isolated in obscure corners of the continents to one global information village. This transformation is no more evident than in the fact that the very boundaries between information transfer and information processing are increasingly hard to define. The integration of communications networks, computing technology, and end-user devices (e.g., the telephone, television, personal computer) is increasingly being referred to simply as the information infrastructure. This global transformation of the world community is no more evident than in the Internet, which some analysts predict will be the predominate mechanism for conducting business (both consumer and business-to-business) within 5 years.

OR and MS have been major players in the development, deployment and management of information technologies and infrastructure. Applications of OR/MS in modeling, analysis and design of communications networks are among the oldest of the fields, dating from the late nineteenth and early twentieth century. Among the most notable of all work in OR/MS history is queueing modeling of telephony by A.K. Erlang. Modeling, analysis and design of communications networks, moreover, is an area rich in applications of more generic OR/MS work. Communications networks are, fundamentally, networks and thus, almost all generic discussion of networks applies. Analogous remarks are appropriate: in communications network modeling and analysis for topics such as queueing and queueing networks, simulation, and network reliability; and in communications network design for topics such as facility location, topological design and optimization, capacity optimization and allocation. Finally, communications networking problems have a great deal of commonality with problems arising in other domains, for example, modeling, analysis and design of transportation systems, water resource distribution systems, etc.

A discussion of the wealth of communications networking issues arising in the application of OR/MS techniques would be quite extensive. Here the focus is on several classes of modeling, analysis, and design problems arising in a variety of modern communications technologies.

Basic Structure and Concepts

A typical communications network comprises a set of subscribers that offer subscriber-to-subscriber traffic requirements to be supported on the given network architecture. For example, a typical household (subscriber) makes telephone calls (traffic requirements) to be supported on a voice network switching fabric (architecture). In most communications architectures, a hierarchy of communications devices exist to support traffic, but the most basic of these are customer premises equipment, local access equipment, and switching equipment. Customer premises equipment is associated directly or indirectly with the generation of traffic requirements. Local access equipment provides a means of connecting the subscriber to the network, that is, the interface between the subscriber and the network necessary for traffic to enter the network and be routed over it. Switching equipment routes the traffic from its source subscriber to its destination subscriber.

All three types of equipment are determined by the nature of the traffic requirements and their associated technology and architecture. In a voice (i.e., telephone) network, the customer premises equipment is generally just a telephone – in this case, the subscriber is the household whose aggregate traffic (telephone calls) enters and leaves the network at the telephone. The local access equipment in this case is owned and provided by the local telephone company. Although there typically is switching in the local access in this case (for local calls), for purposes of the discussion here, the long haul switching equipment is owned and provided by a common carrier such as AT&T. Analogous examples can be provided for data communications networks, video teleconferencing networks, etc.

Communications networks differ on the manner in which they carry traffic requirements. Considerable attention has been paid to quality-of-service (QoS) issues in communications networks, with major trends in standards and implementations increasingly focused on assuring that different applications receive the QoS they require across the technologies they traverse. Most voice networks set-up calls from source to destination in a circuit switched manner, that is, dedicating capacity along the entire path of the call. Most data networks segment information into streams of packets or cells which are routed independent from one another from source to destination and reassembled into the original information at the destination. Data networks can operate with or without functions that route traffic according to QoS needs and with or without reservation or dedication of capacity along the path. Many variations and hybrids of these basic approaches exist and the evolution of technology is becoming increasingly toward supporting traffic sources with differing traffic characteristics and differing service requirements differently. For example, voice traffic is error tolerant (one can tolerate a little static on the line) but delay sensitive (one cannot tolerate long delays between the time a word is spoken and the time it is received at the destination). Some data traffic (e.g, file transfer) is typically error intolerant but delay insensitive. Consideration of these kinds of issues is addressed in network modeling and simulation.

A common thread among most network modeling, analysis and design conceptualizations is the view of a network as a graph comprising nodes and links. A node is used to abstractly represent a device location (e.g., a subscriber location or a switching location). A link is used to represent connections between subscribers and switches and between switches. A link typically has a capacity for supporting traffic. One can view a link as analogous to a pipe and the capacity of the link as analogous to the diameter of the pipe, but with one caveat. A communications link of a given capacity typically supports traffic at that capacity in both directions, that is, it is more properly viewed as two pipes of equal capacity in parallel, each flowing in a direction opposite the other. In addition, a single link can support many "logical" entities as well – for example, it can have its bandwidth dedicated in some proportions to support different service classes or priorities for purposes of providing differential QoS to traffic applications. Design of communications networks typically addresses selecting the number of and the placement of backbone (central) nodes, selecting and sizing the links between subscribers and backbone nodes, selecting and sizing the links between pairs of backbone nodes, and configuring logical constructs (virtual links sharing physical links, bandwidth allocations among service classes, etc.).

Modeling

Communications networks are large scale systems with enormous complexity. As with most such systems, modeling relies heavily on computer-based techniques and the nature of the models developed depends strongly on the questions the model is intended to answer. For example, a simulation may be used to answer detailed questions regarding the interaction of communications devices or protocols. Often these studies address questions as to the feasibility of a given device or protocol to support certain types of traffic requirements with acceptable performance. Such models can be used to design the devices or protocols as well. Simulation of communications systems typically models the generation, transfer, and disposition of each unit of information (e.g., call, packet, cell), the protocol decisions as the system operates and the physical behavior of the devices that make up the network. As with any simulation, various

aspects of the system may be ignored or aggregated to improve the computational speed of the simulation.

An alternative to simulation approaches is analytical modeling (a classic reference is Kleinrock 1976), which typically implicitly aggregates traffic units into flows whose characteristics are captured using statistical or probabilistic models. The advantage of analytical modeling is that the behavior of a network can be predicted by a system of equations more quickly computed than the operation of the network can be simulated. The disadvantage is in the aggregation and averaging of detail, effectively capturing the behavior of the network on average rather than accurately depicting a realization of performance over time. Most analytical models of communications systems employ individual and network queueing models. Information units (calls, packets, cells) are the customers in these queueing systems and communications devices (switches, links, etc.) are the servers.

Hybrid simulation/analytical modeling is a third and increasingly popular approach to communications network modeling (Sage and Sykes 1994). The tenets of this approach are to use simulation techniques in capturing key protocol decisions in traffic admission, routing, congestion control, and resource allocation, but to use analytical techniques for modeling the behavior of the traffic itself, thus avoiding the computational complexity incurred if each packet or cell were to be simulated individually. Hybrid simulation/analytical models of communications networks also have been described as "flow-based simulations," in which the paths that traffic flows take are simulated while the flows themselves are modeled analytically.

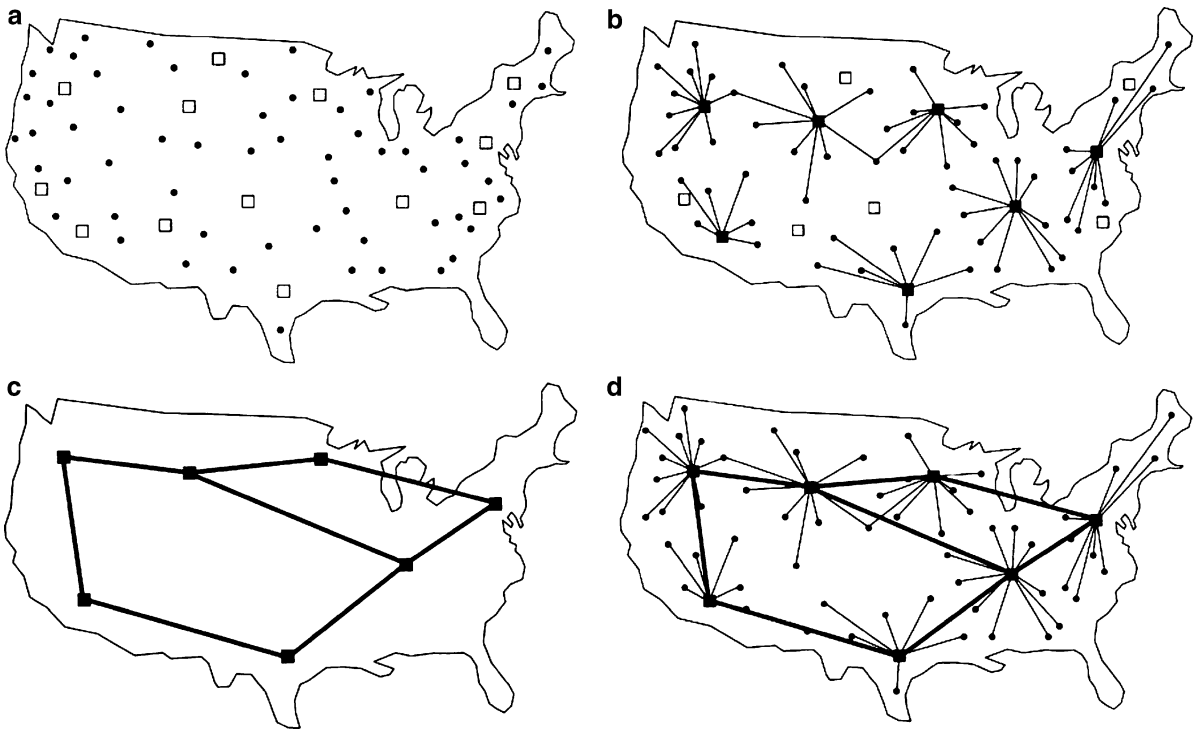
Selection of modeling approach depends strongly on the purpose to which the model is applied. For purposes of protocol or device design, where many replications of realizations of performance are required to observe the entity under a wide variety of operational conditions and circumstances, simulation approaches dominate. For analysis and design purposes, where often the intent is to assess the quality of the design or to compare alternative designs, models which provide average behavior over many potential realizations of performance are useful. Performance can be computed over multiple simulation replications, but analytical tools or simulation analytical hybrids which compute those aver-ages directly and more efficiently are dominant.

Analysis

Network analysis is the application of one or more network models to characterize a communications network. In many communications network design contexts, the central step of the design process is to characterize a design on a number of categories of measures: cost, topological properties, performance, behavior under failures (survivability) being the major ones. For each of these categories of measures, models which compute specific measures of interest can be applied, with the aggregate network analysis being produced in summary from the results of the individual models. Cost measures can include one-time (e.g., device purchase) and recurring costs (e.g., link leasing), often commensurate to the same units. Topological measures are generally technology independent characterizations of the network structure along gross lines (e.g., measures summarizing path availability and diversity, path lengths in number of links or hops from source to destination, etc.). Performance measures are generally technology dependent characterizations of the ability of the network to support the offered traffic and the quality of that support. Survivability measures are indications of what traffic can be supported under various failure scenarios and what the performance of the network will be in those scenarios.

Design

A common paradigm for design of communications networks is one in which the design process is broken down into two phases: *access area design* and *backbone design* (Boorstyn and Frank 1977). Access area design determines the number and location of backbone nodes and homes (i.e., provides a link from) each subscriber to a backbone node. Backbone design determines the interconnections among (links between) backbone nodes. The process is depicted in Fig. 1. Figure 1(a) represents the starting point, where the subscriber locations (black circles) and candidate backbone node locations (squares) are given.



Communications Networks, Fig. 1 (**a**) Network design – starting point. (**b**) Network design – access area design. (**c**) Network design – backbone design. (**d**) Network design – integrated solution

Figure 1(b) represents the completion of the access area design phase, where the black squares are the selected backbone nodes and the lines from the subscribers to the backbone nodes are the homings (implicit in the homings is the assumption that the communications links from each subscriber to its switch is of type and capacity to support the subscriber's offered traffic requirements). The output of the access area design phase is the input to the backbone design phase: the number and location of backbone nodes and the aggregate traffic requirements among the backbone nodes. The aggregate traffic is computed based on the homings. In the backbone phase, the interconnections among backbone nodes are designed to support the backbone traffic with adequate performance, to meet other constraints and typically to minimize cost. Figure 1(c) depicts a backbone design and Fig. 1(d) depicts the final overall solution.

It is notable that solution of the global design problem (including all access and backbone components) is precluded by the computational complexity of the design problem for all but a few special cases which will be ignored here. It also is notable that the structure of the decomposition of the global problem into access area and backbone design phases can lead to gross suboptimalities in the overall solution. To illustrate this assertion, consider a global design problem in which the total cost of the network includes three components:

- homing link costs, the sum of the costs of links homing subscribers to nodes, which can vary for each subscriber-node pair;
- backbone node costs, typically the cost associated with purchasing each node selected as part of the backbone, which typically is uniform over all candidate nodes; and
- backbone link costs, the sum of the costs of links between backbone nodes, which can vary for each subscriber-node pair.

Under fairly general assumptions, the following relationships hold as the number of nodes selected for the backbone increases:

- the access area homing costs tend to decrease (because the access links tend to decrease in length and hence cost);

- the node activation costs increase linearly (directly with the number of nodes selected); and
- the backbone link costs tend to increase (as the number of backbone nodes increases, more backbone links are required).

Thus, if the access area design phase optimizes solely on the basis of homing and node activation costs, it tends to select too many nodes. Two remedies to this pathology are commonly employed: (i) using some estimate of the backbone cost in the access area design problem; or (ii) iterating on the number of backbone nodes selected (i.e., fixing the number of activated nodes to given number, solving the access area and backbone problems in sequence for that number, and computing the total solution cost, but doing so over a wide range of numbers of nodes and selecting the best total cost solution obtained). Neither remedy guarantees global optimality, but both approaches can improve solution costs substantially.

Access area design problems often are formulated as 0-1 integer-programming problems (Fischer et al. 1993) that are strongly related to discrete location problems and/or facility location problems generally (Mirchandani and Francis 1990). In many cases, these integer programs are too large to be solved directly, so a variety of solution approaches are used; for example, linear-programming relaxation methods, Lagrangian relaxation methods, cutting plane and column generation methods, etc. (Ahuja et al. 1993). Alternatively, heuristic algorithms can be used to solve the access problem, and perhaps, more generally, clustering techniques can be used as a solution approach. The basic access area design problem can be stated as follows:

Given:	Subscriber-to-subscriber traffic requirements; Candidate node locations.
Minimize:	Sum of costs of homing each subscriber to a backbone node + Sum of node activation costs.
Over:	Node activations; Subscriber homings.
Subject to:	Node port constraints (limit on the number of subscribers than can be homed to a node); Node traffic constraints (limit on the total amount of subscriber traffic that can be homed to a node); Each subscriber must be homed to a node (occasionally subscribers must be homed to more than one node); (Optionally, a constraint fixing the number of node activations).

Backbone design problems can be formulated as 0-1 or general integer-programming problems (Gavish 1986), however, it is difficult if not impossible to accurately capture or predict network performance in that context. Moreover, many of the critical aspects of the backbone problem that can be captured in the integer-programming formulation (e.g., topological constraints) can also cause a combinatorial explosion in its solution time. Nonetheless, OR/ MS literature is replete with many IP backbone design formulations. In these cases, the solution techniques again typically rely on LP relaxation or Lagrangian relaxation approaches.

An alternative to the mathematical-programming approach to backbone design is commonly employed in interactive software based tools for solving design problems (Stiffler and Sykes 1990; Monma and Shallcross 1989). This iterative approach:

- starts the design process with an initial design;
- analyzes the design using a series of models assessing measures of various aspects of cost, topological properties, performance, and physical constraints on feasibility;
- makes an assessment as to whether the design is satisfactory, stopping if so; and if not
- improves one or more design deficiencies and returns to the analysis step.

This iterative paradigm for backbone design has been used extensively and successfully for design of communications networks with a wide range of architectures (e.g., voice, packet data, multiplexer, asynchronous transfer mode). It also can capture directly a broader set of design objectives and constraints than mathematical-programming methods, as well as be implemented in ways which more accurately predict network performance. All of this is possible through the embedding of the comprehensive network analysis at the core of the process, along with the decomposition of the optimization process into smaller steps aimed at initial design generation and design improvement. Unlike mathematical-programming approaches, which often can be solved to optimality or at least provide bounds from optimality for the solutions they produce, iterative approaches typically cannot guarantee nor bound optimality.

A typical backbone design problem can be stated as follows:

Given:	Backbone node-to-node traffic requirements; Node locations; Link availability and costing.
Minimize:	Sum of link costs.
Over:	Link Placement.

(*continued*)

Subject to:	Topological Constraints, such as – Node connectivity (lower bound on the number of node-disjoint paths available between each node pair); Diameter (upper bound on the minimum number of links a node pair must traverse in order to communicate); and Node port degree (upper bound/physical limit on the number of links that can be incident to each node). Performance Constraints, such as – Constraints on throughput, utilization, delay, blocking, etc., as appropriate for a given network architecture; and Constraints on achieving the QoS requirements of individual traffic demands or, on the achievement of service level agreements computed as a function of QoS of individual traffic demands.

Concluding Remarks

For an overview of telecommunications systems and their operations, see Bertsekas and Gallager (1987); Schwartz (1987); Tanenbaum (2010). For an introduction to network design problems and optimization approaches, see Cahn (1998) or Kershenbaum (1993). For a classical introduction to data communications networking, see Kleinrock (1976), that contains extensive basic modeling and optimization discussions. Also see Pattavina (1998); Ross (1995) or Woodward (1994) for modeling work, and Schmidt and Minoli (1998) or Partridge (1994) for technologies of communications networking. Of a general interest are the books by Ball et al. (1995); León-Garcia and Widjaja (2004), Koster and Muñoz (2010).

See

- ▶ Integer and Combinatorial Optimization
- ▶ Network Optimization
- ▶ Networks of Queues
- ▶ Queueing Theory

References

Ahuja, R. K., Magnanti, T. L., & Orlin, J. B. (1993). *Network flows*. Englewood Cliffs, NJ: Prentice Hall.

Ball, M., Magnanti, T., Monma, C., & Nemhauser, G. (Eds.). (1995). *Network routing*. New York: Elsevier Science.

Bertsekas, D., & Gallager, R. (1987). *Data networks*. Englewood Cliffs, NJ: Prentice Hall.

Boorstyn, R. R., & Frank, H. (1977). Large scale network topological optimization. *IEEE Transactions on Communications, 25*, 29–47. COM.

Cahn, R. S. (1998). *Wide area network design: Concepts and tools for optimization*. San Francisco: Morgan Kaufmann.

Fischer, M. J., Swinsky, G. W., Garland, D. P., & Stanfel, L. E. (1993). A methodology for designing large private line transmission networks with multiple facilities. *Telecommunication Systems, 1*, 243–261.

Gavish, B. (1986). A general model for the topological design of communications networks. *Proceedings GLOBCOM '86*, 1584–1588.

Kershenbaum, A. (1993). *Telecommunications network design algorithms*. New York: McGraw-Hill.

Kleinrock, L. (1975). *Queueing systems, volume I: Theory*. New York: John Wiley.

Kleinrock, L. (1976). *Queueing systems, volume II: Computer applications*. New York: John Wiley.

Koster, A., & Muñoz, X. (Eds.). (2010). *Graphs and algorithms in communication networks*. New York: Springer.

León-Garcia, A., & Widjaja, I. (2004). *Communication networks* (2nd ed.). New York: McGraw-Hill.

Mirchandani, P. B., & Francis, R. L. (Eds.). (1990). *Discrete location theory*. New York: John Wiley.

Monma, C. L., & Shallcross, D. F. (1989). Methods for designing communications networks with certain two-connected survivability constraints. *Operations Research, 37*, 531–541.

Partridge, C. (1994). *Gigabit networking*. Reading, MA: Addison-Wesley.

Pattavina, A. (1998). *Switching theory: Architecture and performance in broadband ATM networks*. Chichester, UK: John Wiley.

Ross, K. W. (1995). *Multiservice loss models for broadband telecommunication networks*. New York: Springer.

Sage, K. M., & Sykes, E. A. (1994). Evaluation of routing-related performance for large scale packet-switched networks with distributed, adaptive routing policies. *Information and Decision Technologies, 19*, 545–562.

Schmidt, A. G., & Minoli, D. (1998). *Multiprotocol over ATM: Building state of the Art ATM intranets*. Greenwich, CT: Manning.

Schwartz, M. (1987). *Telecommunication networks, protocols, modeling and analysis*. Reading, MA: Addison-Wesley.

Stiffler, J. A. & Sykes, E. A. (1990). An AI/OR hybrid expert system for data network design. *Proceedings of the 1990 IEEE International Conference on Systems, Man and Cybernetics*, 307–313.

Tanenbaum, A. S. (2010). *Computer networks* (5th ed.). Englewood Cliffs, NJ: Prentice Hall.

Woodward, M. E. (1994). *Communication and computer networks: Modeling with discrete-time queues*. Los Alamitos, CA: IEEE Computer Society Press.

Community OR

Rebecca Herron
University of Lincoln, Lincoln, UK

Introduction

Community OR is perhaps best understood as a subdiscipline of OR that focuses on communities

as alternative clients for (and users of) OR and related activity.

One defining characteristic of this has been the explicit consideration of two key questions:

1. How can communities benefit from OR approaches, i.e., what can OR offer?
2. Who are the clients/beneficiaries?

These questions have also prompted the subsidiary methodological question: "How should OR practice and theory change in light of this?" These are non-trivial questions which can radically change the view that OR (or management science) as a discipline, only serves traditional managerial interests.

Community OR analysts, or facilitators, as they more frequently refer to themselves, aim to develop engagements within community groups, not-for-profit organizations and/or local multi-agency partnerships so as to provide them analytical support that build capacity and resilience within these communities. Such engagements are frequently related to improving awareness of options and choices available to individuals and groups, and the likely outcomes of various courses of action, e.g. planning and decision-making. They also often involve exploring issues that directly affect people's lives and the strengthening of discourse, dialogue, and local/global agency in relation to these (dialogue, critical awareness, and encouraging self-organization).

The idea of a wider client group for OR and an interdisciplinary approach is not a new idea or pursuit – indeed it has its origins deep in OR's early history – nor is it a concept that has developed identically with different Community OR contributors since. Despite these important differences, common themes have emerged between researchers and practitioners alike that unite the endeavor under the term Community OR.

The practice of Community OR has been organized in many different ways. Much work has been carried out by individuals acting in a voluntary capacity within their own communities, often as a natural extension/application of their personal OR background. Other work has been generated through formalized units and research centers. Both provide contrasting experiences and insights, with different roles and practical and ethical considerations for the Operational Researcher concerned.

Much debate has taken place within the Community OR community about what approaches and methods are appropriate. Given the broad nature of community issues, problems, and choices to be considered, it should not be surprising to find a wide variety of OR/Management Science tools being put to use. While some authors have reported the importance of developing the capacity for quantitative analysis within community groups, a large proportion of the Community OR work has involved softer OR methods that help structure issues: Problem Structuring Methods, understanding inter-related issues and exploring choices; Strategic Choice Approach and decision making/planning, and helping communities reflect on current and future situations and to organize themselves in light of these reflections; and workshops, organizational learning, or creative methods.

Links to systems research are a particularly strong feature of many Community OR studies, but this is not a universal precondition for Community OR. In particular, ideas of socio-technical systems, viable systems modeling (cybernetics), soft systems methodology, boundary critique, and critical systems have played key roles in the development of Community OR.

There is a strong U.K. tradition of Community OR, encouraged by the U.K. OR Society, but global perspectives have been just as important in shaping the development of the subject, including work in Venezuela, Columbia, Ghana, Kenya, New Zealand and Mexico. Many links can also be found in the writings of Community OR researchers and practitioners with Social Justice, Community Development, and Participatory Methods in research and evaluation.

The emphasis of Community OR is often on trying to build meaningful engagements with groups of people in the community and creating increased capacity within the groups. Community OR, therefore, pays great attention to issues of engagement, facilitation, group work, ethical interactions, and sustainability. These are not issues unique to Community OR, but they are in sharp focus in this field.

Another consideration is the question of motivating factors for involvement in Community OR. One factor has already been introduced – a volunteer applying OR skills within his/her own personal communities. Others, however, are motivated to work within new communities, usually ones where there is felt to be some kind of social

inequality or collective issue to address. Community OR researchers tend not to see themselves as the expert in these situations; rather, authors write about bringing another viewpoint/set of skills to a community who are already expert in their own situations. As such, there are some very interesting issues about knowledge and power to unpick in Community OR, and political, religious or ethical beliefs to reflect upon. These have often shaped the way Community OR has been developed by individuals or groups of researchers; seen particularly in the theoretical foundations of the approaches developed and, indeed, in the choices of community issues that are addressed in practice.

History

The desire that OR should be put to use for societal aims is not a new one. As early as the 1930s, Patrick Blackett (often referred to as "the father of OR"), John Bernal and others were calling for science to be put to use for wider social benefit (Blackett 1935; Bernal 1939). In the post-war years in the U.K., OR departments were set up across governmental and nationalized industries.

By the 1970s, some of the initial social objectives of OR seemed to have been somewhat forgotten. There were, however, several key figures who campaigned for redressing this balance and restoring the interdisciplinary, action-focused aspects of the OR enterprise.

In the U.S., Russell Ackoff and colleagues at the University of Philadelphia pioneered work with local residents. In the title of his 1970 paper, "*A black ghetto's research on a university,*" Ackoff made an important distinction that was picked up later by Community OR researchers (Ackoff 1970). It was not the University's research on the ghetto, but the research participants' research on a University. Ackoff continued developing his program and calling for a systems approach to societal problems (Ackoff 1974). Other researchers of the same era famously also called for new perspectives (Churchman 1979) and the recognition and appropriate handling of wicked, i.e. messy & complex, social problems (Rittel and Webber 1973).

In the U.K. there were also calls within the OR community to turn attention again to the social application of OR. The Institute of OR had been established in 1963, later to become attached to the Tavistock Institute of Human Relations, both with an interest in understanding social and public affairs and to work to improve planning processes; Friend and Hickling (2005) discuss some of the long term legacies from this work. Other researchers continued to argue strongly that OR should be used to benefit society and lead to improved well-being (Cook 1973; Thunhurst 1973).

By the 1980s, the movement pressing for OR to move beyond its now well-established scope and client-base to encompass community beneficiaries had gathered considerable momentum. In 1985, under the presidency of Jonathon Rosenhead, the U.K. OR Society wanted to challenge perceived views about who the clients of OR were and to find a significant social role for OR:

> The idea behind the Unit is that is should give extensive experience of how formal problem-structuring approaches can assist non-hierarchical organizations, disposing of few resources, but attempting to represent the interests of their members … we see the unit as extending the range of OR's potential clients … we shall be expanding the domain of rational argument, tackling a new and exciting range of unstructured problems, and contributing to making our society a better one to live in. (Rosenhead 1987, quoted in Parry and Mingers 1991).

The OR Society set up the original Community Operational Research Unit at Barnsley College, Yorkshire (Ritchie et al. 1994, discuss case studies from this period). A new Community OR Network was created the following year by the Society, and a Centre for Community OR was also set up at Hull University, building on the work of Jackson and Keys (Jackson 1987, gives an overview of these 3 initiatives).

Other work in this era also fed into the discussion of the newly emerging concept of Community OR. This included the work of Jones and Eden (1981) and parallel, but related systemic researches such as Ulrich's Critical Heuristics (Ulrich 1983).

By the 1990s, sufficient examples of Community OR were appearing in the literature (Thunhurst 1987; Mar Molinero 1993; Taket and White 1994; Ritchie et al. 1994, and others) to enable the community to reflect on Community OR as an entity and approach in its own right (Jackson 1987;

Mar Molinero 1992; Midgley and Ochoa-Arias 1999; Ochoa-Arias 1994; Parry and Mingers 1991; White and Taket 1993; Wong and Mingers 1994). These papers demonstrate a breadth of OR activity drawn from a range of OR techniques (hard and soft OR methods), drawing on different intellectual traditions (e.g. different forms of systems modeling or political theory), but sharing common aims and interests, such as participation, social justice, and community empowerment.

The following decade saw a period of consolidation for Community OR with the publication of three core OR texts that in different ways document and shape the record of Community OR theory and practice. *Planning under Pressure* (Friend and Hickling 2005) shares the authors' and 21 other contributors' experiences of using the Strategic Choice Approach. *Rational Analysis for a Problematic World Revisited* (Rosenhead and Mingers 2001) brings together the work of several authors associated with the development of Community OR and illustrates their approaches to Problem Structuring in a range of decision contexts. *Community Operational Research, OR and Systems Thinking for Community Development* (Midgley and Ochoa-Arias 2004) focuses specifically on Community OR and brings together several key papers as well as setting the scene for some future development, e.g. in environmental OR (Midgley and Reynolds 2004).

At the end of the 1990's the Community Operational Research Unit (CORU) moved to the University of Lincolnshire and Humberside (soon to become the University of Lincoln). During the subsequent decade researchers from CORU developed citizen learning networks and participatory evaluations - exploring issues for social justice, well-being, community self-organization and social action research (Herron (2006) and Herron & Mendiwelso Bendek (2007) introduce examples in two Special Editions of *OR Insight*, vol. 19 issue 2 and vol. 20 issue 2).

The expansion possibilities of the scope of Community OR, in terms of geographical spread of activity and the issues under consideration, are extensive. There are important global issues that call for the continued and renewed attention: poverty alleviation, social justice, community well-being, environmental responsibility, fair trade, community organization and resilience, as well as local, national, and international calls for greater community participation in local decision making and planning.

Emerging Themes for Community OR

There has been much interesting debate about the variety within Community OR that reflects the range of different professional contexts of researchers and practitioners and different contributing areas of expertise within OR/Systems/Action Research. Accounts of different engagements provide a rich source of case studies of using different OR methods and approaches with different communities (Ritchie et al. 1994; Midgley and Ochoa-Arias 2004).

The nature of the community situations encountered has also clearly had a large impact in what has been done under the banner of Community OR, and creating engagements that are meaningful and have value for those taking part may mean the selection, and modification, customization of different OR methods. It does not seem very practicable to attempt to define Community OR by the choice of methods or tools used, although a familiarity with Soft OR/Problem Structuring Methods is useful for understanding much of the existing literature or to develop skills likely to be of use within a number of community contexts (Rosenhead and Mingers 2001; Friend and Hickling 2005).

Community OR could also be defined as the resulting body of work of a community of practice (the socially constructed definition). But, to define the subject this way provides little insight for the initial enquirer, as it requires a familiarity and further knowledge of the outputs of the community of practice concerned.

Importantly, all this does not mean that it is not possible to identify Community OR themes that have emerged over the past decades. Rather than discussing a single unified methodology for Community OR, it seems productive to describe similarities of approach that transcend discussion of which methods have been developed or applied, and focuses instead on some of the general characteristics and values emergent in Community OR. An introduction would not be complete without at least starting to draw out a few of these themes that recur in many, if not necessarily all, Community OR activities. Alongside the other elements presented above, these will help to provide a fuller introduction to the subject.

Engaging with Communities

Interventions and Interactions with Local Groups

Community OR is generally understood as a form of action research. In this sense, it returns to some of the original intentions and conceptions of OR where interdisciplinary teams work together to support solutions to problematic situations and/or work with problem-owners to explore improvements in how they operate (Jackson 1987). In Community OR this usually always involves some form of engagement with community groups, not-for-profit organizations, or multi-agency partnerships working toward a social/community aim.

To work in meaningful ways with community groups, regardless of the type of method used, requires the establishment of good working relations with the relevant parties and the identification of key stakeholders, particularly those made vulnerable or marginalized by the current situation. Many Community OR approaches start with some form of stakeholder analysis and the scoping of different perspectives and points of view, along with other forms of collective sense-making or mess-structuring activity (Rittel and Webber 1973; Rosenhead and Mingers 2001).

Encouraging the full participation of all the stakeholders identified can be very challenging and requires the building of trust and the careful consideration of issues of access and appropriate delivery. Flexibility and creativity of approach is likely to be valuable – and many Community OR engagements include the need to adapt methods and delivery styles for the particular group concerned. Thus, in addition to content discussions, much Community OR literature also explores these softer issues of group facilitation, interacting with communities and sense-making activities that handle multiple perspectives and conflict of opinion and objective.

Analytical Process Support

A common motivation for undertaking Community OR work is the desire to support a community or communities. Often this support is in terms of providing some structured intervention such as workshop, training, or participant research that helps strengthen the groups' ability to think through and analyze a situation, identify or create new resources, build robust arguments and narratives (sometimes even models), or create improved dialogue and awareness of a situation amongst stakeholders.

Much of the work done in this respect has links to, and has implications for, planning and decision making, awareness of options, and choice of action. Community OR interventions are very often workshop-based or use other forms of community-based learning. Community OR facilitators usually provide process-facilitation and spaces for reflection and dialogue: the aim being to support increased capacity within communities to understand and be more resilient to changes in their external environment.

Emphasizing once again that Community OR is not defined by the application of any particular method, approaches applied in community contexts have included:

- Cognitive Mapping
- Community Visioning
- Critical Systems Heuristics (CSH)
- Drama Theory
- Strategic Choice Approach (SCA)
- Strategic Options Development and Analysis (SODA)
- Soft Systems Methodology (SSM)
- Viable Systems Modeling (VSM)

The exact form of analytical support provided by each Community OR researcher/practitioner will, of course, be context dependent—specific to the exact needs of a particular group and the issue concerned at a certain time—and it will also be bounded by the choice of method or approach chosen. Certain themes, however, emerge in terms of what Community OR may be said to support, including:

- Processes and structures (exploring, building, and rethinking)
- Dialogue and supporting groups to build stronger arguments (logical argumentation)
- Information and enabling groups to make more informed choices (handling information)
- Reasoning about local issues, including exploring links to global issues
- Negotiation and creating rules of engagement
- Engaging relevant others; exploring and extending stakeholders
- Sweeping-in new elements to the model such as individuals, issues, values, ethics

- Critical awareness of learning processes, political impact on decision making, changing environments, and the possible impacts and side effects of choices and actions

Reasons for doing this include:

- Organizational learning – facilitating community-learning processes
- Knowledge transfer – co-creating knowledge and analysis
- Addressing inequalities – providing access to analytical resources
- Increasing choice and resilience for communities - managing uncertainty
- Increasing community knowledge and confidence to act in changing environments

Emancipation and Social Justice

The U.K. OR Society has described OR as "The Science of Better." For Community OR, this highlights the key issues of improvement, i.e., improvement for whom, and how?

Community OR, as much as any subdiscipline of OR, has highlighted the complexity of working for improvement in contexts where there are multiple goals and perspectives. It has provided many examples and ways of proceeding in situations where there are multiple world-views, value systems, and objectives interacting to build up a complex collective logic. In this, Community OR has often been seen as working in the way suggested by Rittel and Webber (1973): continually solving and resolving tricky, messy, and complex situations and providing ongoing support for situations that continue to evolve.

Community OR practice seems to generate several key themes related to making the notion of a Science of Better meaningful for work with communities:

- Ethical dimensions: considering the likely impact on multiple stakeholders
- Locus of control: who has ownership of the goals? processes? outcomes?
- Surfacing issues: creating models participants find authentic and provide insight
- Increasing understanding and fairer dialogue
- Collective improvement: critical reflection from different perspectives
- Consideration of likely side-effects and issues of robustness and sustainability
- Increasing individual and collective control and agency
- Supporting vulnerable people, readdressing inequalities, and rethinking the client
- Exit strategies: building community capacity for learning, analysis, and reflection

Different Community OR facilitators have approached many of these issues in different styles and by using different methodologies, especially SSM, VSM, CSH, SCA, the choice of which depends very much on their experiences, cultural and intellectual traditions, and personal beliefs. However some common emergent themes are worth noting:

- Inclusion of vulnerable groups in decision making and new forms of participation
- Empowerment of communities, emancipation, and addressing social inequality
- Democratic decision making: dialogue, interaction, and community learning
- Handling plurality: multiple realities and objectives
- Self-organization and local control: strengthening civil society
- Feedback, communication, and the interlinking of issues
- Linking local issues to global concerns

Community OR also continues a long tradition in OR of valuing the co-creation of knowledge. The Community OR facilitator will usually have certain process knowledge to contribute (e.g., problem structuring, mess-handling, restructuring, drama theory, game theory), but other participants in the group will have been involved because of their local context knowledge and experience, or other specific knowledge bases. Community OR practice usually explicitly values these other knowledge forms and encourages groups to take ownership of the process of exploration, and idea and solution generation. This is also consistent with the emancipatory interests of many researchers and practitioners, often underpinned by distinct philosophical positions including those shaped by the works of Habermas and Foucault.

Core to any discussion about any Science of Better must also be a critical reflection of who has the power to determine decisions – and the directions chosen for improvement. More critically still, those who do not have any say in these decisions are excluded from the dialogue for any number of reasons, or only able to have a very small voice in the decision-making

processes that affect their lives? Thus, while much of OR strives to remain firmly outside political discourse, Community OR practitioners work in situations where strengthening collective critical awareness of the rules of engagement, and the fairness of these, is seen as an important aspect of the work of the analyst.

See

- ▶ Cognitive Mapping
- ▶ Critical Systems Thinking
- ▶ Deep Uncertainty
- ▶ Delphi Method
- ▶ Developing Countries
- ▶ Group Decision Making
- ▶ Soft Systems Methodology
- ▶ Strategic Assumption Surfacing and Testing (SAST)
- ▶ Strategic Choice Approach (SCA)
- ▶ Strategic Options Development and Analysis (SODA)
- ▶ System Dynamics
- ▶ Systems Analysis

References

Ackoff, R. L. (1970). A black ghetto's research on a university. *Operations Research, 18*, 761–771.

Ackoff, R. L. (1974). *Redesigning the future: A systems approach to societal problems*. New York: Wiley.

Bernal, J. D. (1939). *The social function of science*. London: Routledge.

Blackett, P. M. S. (1935). The frustration of science. In D. Hall et al. (Eds.), *The frustration of science* (pp. 129–144). London: Allen and Unwin.

Churchman, C. W. (1979). *The systems approach and its enemies*. New York: Basic Books.

Cook, S. L. (1973). Operational research, social well-being and the zero-growth concept. *Omega, 1*(6).

Friend, J., & Hickling, A. (2005). *Planning under pressure* (3rd ed.). Oxford: Elsevier/Butterworth-Heinemann.

Herron, R. (2006). Editorial – Special issue for community operational research. *OR Insight, 19*(2), 2–3.

Herron, R., & Bendek, M. (2007). Take part: Active learning for active citizenship contributing to community OR reflections and practices. *OR Insight, 20*(2), 3–7.

Jackson, M. C. (1987). Community operational research: Purposes, theory and practice. *Dragon, 2*(2), 47–73.

Johnson, M. (Ed.). (2011). *Community-based operations research: Decision modeling for local impact and diverse populations*. New York: Springer.

Jones, S., & Eden, C. (1981). OR in the community. *Journal of the Operational Research Society, 32*, 335–345.

Mar Molinero, C. (1992). Operational research: From war to community. *Socio-Economic Planning Sciences, 26*, 203–212.

Mar Molinero, C. (1993). Aldermoor School: The operational researcher on the side of the community. *Journal of the Operational Research Society, 44*, 237–245.

Midgley, G., & Ochoa-Arias, A. E. (1999). Visions of *community OR*. *Omega, 27*, 259–274.

Midgley, G., & Ochoa-Arias, A. E., (Eds.) (2004). *Community operational research, OR and systems thinking for community development*. Kluwer Academic/Plenum.

Midgley, G., & Reynolds, M. (2004). Community and environmental OR: Towards a New Agenda. In G. Midgley & A. E. Ochoa-Arias (Eds.), *Community operational research, OR and systems thinking for community development*. Kluwer Academic/Plenum.

Ochoa-Arias, A. E. (1994). The possibilities for community OR in a third world country. *International Transactions in Operational Research, 1*, 345–352.

Parry, R., & Mingers, J. (1991). Community operational research: Its context and its future. *Omega, 19*, 577–586.

Ritchie, C., Taket, A., & Bryant, J. (Eds.). (1994). *Community works: 26 case studies showing community operational research in action*. Sheffield: Pavic Press.

Rittel, H. J. W., & Webber, M. M. (1973). Dilemmas in a general theory of planning. *Policy Science, 4*, 155–169.

Rosenhead, J. (1987). From management science to workers' science. In M. C. Jackson & P. Keys (Eds.), *New directions in management science*. Aldershot: Gower.

Rosenhead, J. V., & Mingers, J. (Eds.). (2001). *Rational analysis for a problematic world revisited*. Chichester: Wiley.

Taket, A. R., & White, L. A. (1994). Doing community operational research with multicultural groups. *Omega: International Journal of Management Science, 22*(6), 579–588.

Thunhurst, C. (1973). Who does OR operate for?. *Presented at OR Society Conference*, Torquay.

Thunhurst, C. (1987). Doing OR with the community. *Dragon, 2*, 143–153.

Ulrich, W. (1983). *Critical heuristics of social planning*. Berne: Paul Haupt.

White, L., & Taket, A. (1993). The death of the expert. *Journal of the Operational Research Society, 45*, 733–748.

Wong, N., & Mingers, J. (1994). The nature of community OR. *Journal of the Operational Research Society, 45*, 245–254.

Complementarity Applications

Steven A. Gabriel
University of Maryland, College Park, MD, USA

Introduction

This article emphasizes complementarity applications found in the infrastructure industries. A number of

such industries in the U.S. and overseas have been restructured with the goals of making them more competitive and transparent. Examples of these industries in the U.S. include: energy (production, transmission, and distribution), air transportation, and telecommunications. The business model for these industries dramatically changed with a greater emphasis on the microeconomic, game theory effects of individual market players seeking to maximize their own profits or utilities rather than just having a central facility optimizing a stream of regulated profits. Along with these changes came the rise of data-gathering culminating in the present-day Internet which in some cases is nearly real-time. These two sources of change (no doubt with other contributing factors) have led to the rise of complementarity models in the operations research community.

This broad class of mathematical programs includes many optimization problems (via their Karush-Kuhn-Tucker conditions), n-person Nash games, solving nonlinear equations, and many other interesting problems in a variety of engineering and economic settings (Cottle et al. 1992; Facchinei and Pang 2003). A related class of problems, variational inequalities, also benefitted from these contributing factors (see Facchinei and Pang (2003) for a discussion of the relationship between these two problem classes). Collectively, complementarity or variational inequality problems are sometimes called equilibrium problems in that they both seek to arrive at a solution so that the system under study is balanced or has no incentive to change.

One advantage of complementarity problems over traditional optimization is the ability to simultaneously manipulate both primal variables as well as shadow prices for resources, usually expressed as Lagrange multipliers for constraints. This ability, coupled with the complicated picture of restructured industries which often are composed of both regulated and deregulated portions, can usually be approached from the complementarity perspective resulting in richer, more realistic models.

As complementarity models have become more mainstream like linear programs that preceded them, other more complicated and potentially more realistic problem classes have been studied. These extensions of complementarity problems include: mathematical programs with equilibrium constraints (MPECs) which are optimization problems having two or more levels with the bottom ones potentially complementarity problems (Luo et al. 1996), quasi-variational inqualities corresponding to Generalized Nash games (Harker 1991), (Facchinei and Pang 2003), and equilibrium problems with equilibrium constraints (EPECs), problems with two or more levels with an equilibrium at multiple levels, which are some of the hardest problems to solve (Facchinei and Pang 2003).

There are several forms for the complementarity problem, the most common of which is the mixed one abbreviated as MCP (mixed complementarity problem). Note that the term mixed refers to the presence of both equations and inequalities. Having a function $F : R^n \to R^n$, MCP(F) is to find vectors $x \in R^{n_1}, y \in R^{n_2}$ such that the following conditions hold:

$$F_x(x, y) \geq 0, x \geq 0, x^T F_x(x, y) = 0 \tag{1a}$$

$$F_y(x, y) = 0, y \quad \text{free} \tag{1b}$$

where the notation $F_x(x, y), F_y(x, y)$ refers, respectively, to those components of F that match up with the vectors x and y. Equivalently, the first set of conditions (1a) can be stated as $F_i(x, y) \geq 0, x_i \geq 0, x_i \cdot F_i(x, y) = 0$, for $i = 1, \ldots, n$ with the last set referred to as "complementary conditions" (either x_i or $F_i(x, y)$ or both must equal zero). A more compact representation of this first set of conditions is often stated as $0 \leq F_x(x, y) \perp x \geq 0$ with the perpendicular operator $\perp$ denoting the inner product of two vectors equal to zero. The statement of MCP(F) is deceptively simple– just a set of inequalities and complementarity conditions, as well as equations that must simultaneously be satisfied. This very general form, however, includes many problems in optimization, game theory, as well as a host of other areas some of which are described below; additional examples and/or related theory can be found in Cottle et al. (1992), Harker and Pang (1990), Harker (1993), Ferris and Pang (1997), Ferris et al. (2001), Facchinei and Pang (2003), Gabriel et al. (2013).

Discussion

To demonstrate the generality and flexibility of (1), a few representative examples are next shown.

A Simple Complementarity Example

Let $n_1 = 2$, $n_2 = 1$ and $F : R^3 \to R^3$ be defined as:

$$F(x_1, x_2, y_1) = \begin{pmatrix} F_1(x_1, x_2, y_1) \\ F_2(x_1, x_2, y_1) \\ F_3(x_1, x_2, y_1) \end{pmatrix} = \begin{pmatrix} x_1 - x_2 \\ x_1 + y_1 \\ x_1 - y_1 + 1 \end{pmatrix}. \quad (2)$$

The corresponding MCP is to find x_1, x_2, y_1 that simultaneously solve the following conditions:

$$\begin{aligned} F_1(x_1, x_2, y_1) = x_1 - x_2 \geq 0, \\ x_1 \geq 0, (x_1 - x_2) \cdot x_1 = 0; \\ F_2(x_1, x_2, y_1) = x_1 + y_1 \geq 0, \\ x_2 \geq 0, (x_1 + y_1) \cdot x_2 = 0; \\ F_3(x_1, x_2, y_1) = x_1 - y_1 + 1 = 0, \\ y_1 \text{ free}. \end{aligned} \quad (3)$$

The first question with any mathematical programming problem is to try to find the solution set. For small problems like (3), the set of solutions can often be determined by enumeration of several cases and then by some algebra. Doing so additionally provides some insight into the structure of complementarity problems.

The first step is to eliminate the free variable y_1 by using the equation $x_1 - y_1 + 1 = 0 \Leftrightarrow y_1 = x_1 + 1$ and then making the substitution in the remaining two sets of conditions to obtain an equivalent set of conditions:

$$\begin{array}{lll} x_1 - x_2 \geq 0, & x_1 \geq 0, & (x_1 - x_2) \cdot x_1 = 0, \\ 2x_1 + 1 \geq 0, & x_2 \geq 0, & (2x_1 + 1) \cdot x_2 = 0. \end{array} \quad (4)$$

Next, the following four cases can be analyzed to determine the solution set:

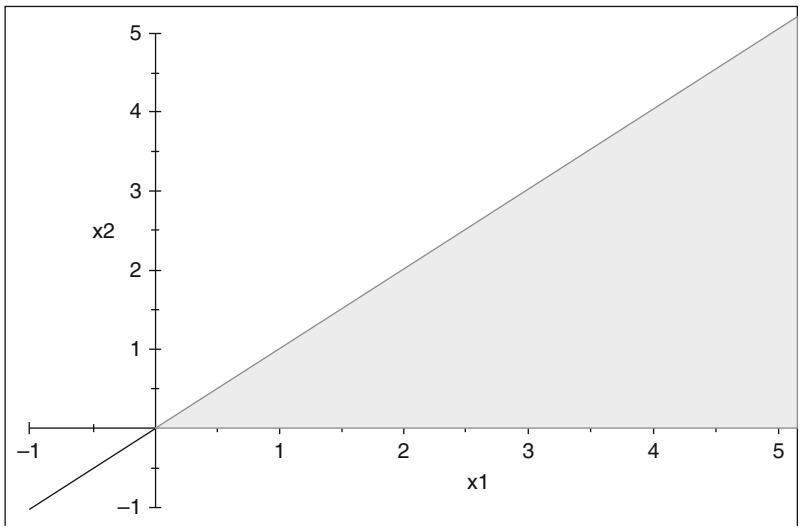
1. $x_1 > 0, x_2 > 0$
2. $x_1 = 0, x_2 > 0$
3. $x_1 > 0, x_2 = 0$
4. $x_1 = 0, x_2 = 0$

Using the complementary conditions, the first case implies that

$$(x_1 - x_2) = 0, (2x_1 + 1) = 0,$$

or that $x_1 = x_2 = -\frac{1}{2}$, which is not possible since both these variables must be positive. Case 2 also is not possible since by complementarity it would imply that $x_1 = -\frac{1}{2} \not\geq 0$. Analyzing Case 3 shows that by complementarity, $x_1 - x_2 = 0$ or that both values must be the same. This is not possible under this case as $x_1 > 0, x_2 = 0$. Lastly, if both values are equal to zero, then all the inequalities as well as complementarity conditions hold. Thus, $(x_1, x_2, y_1) = (0, 0, 1)$ is the unique solution to this linear MCP. This simple, three-variable linear MCP can also be viewed from a geometric point of view in $x_1 - x_2$ space as shown in Fig. 1. The conditions: $x_1 - x_2 \geq 0, x_1 \geq 0, x_2 \geq 0$ represent a polyhedron a sample of which is shown by the shaded region. The condition $2x_1 + 1 \geq 0 \Leftrightarrow x_1 \geq -\frac{1}{2}$ is superfluous given that x_1 must be nonnegative. The first complementarity condition $(x_1 - x_2) \cdot x_1 = 0$ can be interpreted as: if $x_1 > 0$ then the potential solution must be on the line $x_1 = x_2$ which in turn would make $x_2 > 0$. The second complementarity condition would then force $x_1 = -\frac{1}{2}$ which is not in the shaded region. Thus, the only other choice it to have $x_1 = 0$ (satisfying the first complementarity condition) but forcing $x_2 = 0$ by the second one in (4). Hence, the only point in the shaded region that satisfies all the six conditions of (4) is the origin.

While this simple example had a unique solution, in general this will not be the case for MCPs. Indeed, there can be no solutions, one solution, any finite number of solutions or an infinite number since MCPs generalize solving nonlinear (or linear) equations, optimization problems (one or more), or combinations thereof. To see that (1) includes solving equations, consider the case when there are no inequalities, i.e., just $F_y(x, y) = 0, y$ free. In the next sections, the connection with optimization problems and extensions is explored.



Complementarity Applications, Fig. 1 Geometric Depiction of Simple Example

Connection between Optimization and Complementarity Problems

Consider the following standard (primal) linear program and its corresponding dual problem:

$$\min_{x} \boldsymbol{c}^{\boldsymbol{T}}\boldsymbol{x} \tag{5a}$$

$$s.t.\ \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b} \tag{5b}$$

$$\boldsymbol{x} \geq 0 \tag{5c}$$

$$\max_{y} \boldsymbol{b}^{\boldsymbol{T}}\boldsymbol{y} \tag{6a}$$

$$s.t.\ \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{y} \leq \boldsymbol{c} \tag{6b}$$

$$\boldsymbol{y} \geq 0 \tag{6c}$$

where $\boldsymbol{A}$ is a real-valued $m \times n$ matrix, $\boldsymbol{c} \in R^n, \boldsymbol{b} \in R^m$. The m primal constraints $\boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{b}$ are associated with the dual vector $\boldsymbol{y} \in R^m$ and likewise the n dual constraints match up with the primal variables $\boldsymbol{x} \in R^n$. The Complementarity Slackness Theorem (Luenberger 1984) states that if:

1. x is feasible to the primal problem (5)
2. y is feasible to the dual problem (6)

then a necessary and sufficient condition for (x, y) to be optimal solutions to their respective problems is that complementary slackness is satisfied, namely: $(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})_j \cdot y_j = 0, j = 1, \ldots, m$ and $\left(\boldsymbol{c} - \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{y}\right)_i \cdot x_i = 0, i = 1, \ldots, n.$

However, the feasibility and complementary conditions amount to:

1. $\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \geq 0, \boldsymbol{x} \geq 0$
2. $\boldsymbol{c} - \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{y} \geq 0, \boldsymbol{y} \geq 0$
3. $(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})_j \cdot y_j = 0, j = 1, \ldots, m$ and $\left(\boldsymbol{c} - \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{y}\right)_i \cdot x_i = 0, \quad i = 1, \ldots, n.$

After realizing that complementarity slackness can be re-expressed as $(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T \boldsymbol{y} = 0$ and $\left(\boldsymbol{c} - \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{y}\right)^T \boldsymbol{x} = 0$ given the nonnegativity of the quantities involved, these three sets of optimality conditions can be expressed succinctly as the linear complementarity problem with only inequalities (i.e., "pure complementarity problem") with $F : R^{n+m} \rightarrow R^{n+m}$ given as follows:

$$F(\boldsymbol{x}, \boldsymbol{y}) = \begin{pmatrix} \boldsymbol{c} - \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{y} \\ \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b} \end{pmatrix} = \begin{pmatrix} 0 & -\boldsymbol{A}^{\boldsymbol{T}} \\ \boldsymbol{A} & 0 \end{pmatrix} \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} + \begin{pmatrix} \boldsymbol{c} \\ -\boldsymbol{b} \end{pmatrix}$$

Moreover, if the original primal LP had equalities, via a similar line of reasoning, the result would be a mixed (as opposed to pure) complementarity problem. This shows that every linear program is an instance of an MCP. A key distinction to be made here between optimization and complementarity problems is that the latter's formulation involves both primal and dual variables whereas the former's formulation is only in terms of primal (or just dual) variables. As will be shown in some of the examples below, the complementarity approach can lead to richer models that manipulate the dual variables (i.e., prices) while

also considering the primal (usually physical) variables in many infrastructure applications. First, the next section shows that any nonlinear program via its Karush-Kuhn-Tucker (KKT) conditions is also an instance of an MCP generalizing the results for linear programs.

Consider a standard nonlinear program as follows:

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \tag{7a}$$

$$s.t.\ g_i(\boldsymbol{x}) \leq 0, i = 1, \ldots, m \tag{7b}$$

$$h_j(\boldsymbol{x}) = 0, j = 1, \ldots, p \tag{7c}$$

where $f, g_i(\boldsymbol{x}), h_j(\boldsymbol{x}) : R^n \rightarrow R$ are respectively, the objective function and the constraint functions. The KKT conditions (Bazaraa et al. 1993) are to find $\boldsymbol{x} \in R^n$, Lagrange multipliers $\boldsymbol{\lambda} \in R^m$ (for the inequality constraints) and $\boldsymbol{\gamma} \in R^p$ (for the equality constraints) such that the following conditions hold:

$$\nabla f(\boldsymbol{x}) + \sum_{i=1}^{m} \lambda_i \cdot \nabla g_i(\boldsymbol{x}) + \sum_{j=1}^{p} \gamma_j \cdot \nabla h_j(\boldsymbol{x}) = 0, \boldsymbol{x} \text{ free} \tag{8a}$$

$$-g_i(\boldsymbol{x}) \geq 0, \lambda_i \geq 0, g_i(\boldsymbol{x}) \cdot \lambda_i = 0, i = 1, \ldots, m \tag{8b}$$

$$h_j(\boldsymbol{x}) = 0, \gamma_j \text{ free}, j = 1, \ldots, p. \tag{8c}$$

Clearly the KKT conditions are just a set of equations with corresponding free variables and inequalities with associated nonnegative variables and complementarity conditions. As such, the KKT conditions are also an instance of an MCP (Gabriel 2008; Gabriel et al. (2013)) with $F : R^{n+m+p} \rightarrow R^{n+m+p}$ given as:

$$F\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{\lambda} \\ \boldsymbol{\gamma} \end{pmatrix} = \begin{pmatrix} \nabla f(x) + \sum_{i=1}^{m} \lambda_i \cdot \nabla g_i(\boldsymbol{x}) + \sum_{j=1}^{p} \gamma_j \cdot \nabla h_j(\boldsymbol{x}) \\ -g_i(\boldsymbol{x}), i = 1, \ldots, m \\ h_j(\boldsymbol{x}), j = 1, \ldots, p \end{pmatrix}$$

with the first and third sets of constraints being equations and the second set inequalities (≥ 0). Optimization problems that do not have valid KKT conditions are not directly special cases of MCPs.

Since KKT conditions for integer programs (IPs) are not generally valid, there is not a direct connection between MCPs and IPs. However, there is an indirect association between these two classes of problems. In particular, consider the following mixed, linear complementarity problem as an example. This problem in general form, is to find vectors $\boldsymbol{x}, \boldsymbol{y}$ such that:

$$0 \leq \boldsymbol{q}_1 + (\boldsymbol{M}_{11}\boldsymbol{x} + \boldsymbol{M}_{12}\boldsymbol{y}) \perp \boldsymbol{x} \geq 0 \tag{9a}$$

$$0 = \boldsymbol{q}_2 + (\boldsymbol{M}_{21}\boldsymbol{x} + \boldsymbol{M}_{22}\boldsymbol{y}), \boldsymbol{y} \text{ free} \tag{9b}$$

where the matrices $\boldsymbol{M}_{11}, \boldsymbol{M}_{12}, \boldsymbol{M}_{21}, \boldsymbol{M}_{22}$ are of order $r_1 \times n_1, r_1 \times n_2, r_2 \times n_1, r_2 \times n_2$, respectively, coinciding with $\boldsymbol{x} \in R^{n_1}, \boldsymbol{y} \in R^{n_2}$. Also, the constant vectors $\boldsymbol{q}_1, \boldsymbol{q}_2$ are of size r_1 and r_2, respectively. This system can be re-expressed as the following set of polyhedral constraints with additional binary variables $b \in \{0, 1\}^{r_1}$:

$$0 \leq \boldsymbol{q}_1 + (\boldsymbol{M}_{11}\boldsymbol{x} + \boldsymbol{M}_{12}\boldsymbol{y}) \leq K\boldsymbol{b} \tag{10a}$$

$$0 \leq \boldsymbol{x} \leq K(1 - \boldsymbol{b}) \tag{10b}$$

$$0 = \boldsymbol{q}_2 + (\boldsymbol{M}_{21}\boldsymbol{x} + \boldsymbol{M}_{22}\boldsymbol{y}), \boldsymbol{y} \text{ free} \tag{10c}$$

with K a suitably chosen positive constant (could vary for each of the r_1 constraints). To see why this works it suffices to realize that complementarity conditions are either-or type restrictions. Either one term equals zero or the other does (or both). This disjunction is equivalently expressed in (10) via the binary variables $\boldsymbol{b}$ and the constant K (Fortuny-Amat and McCarl 1981). In principle then, one could replace a linear complementarity problem's conditions by a set of linear conditions with binary variables as shown above. The problem arises in determining an appropriate constant K. Too small a value will unnecessarily restrict the problem and result in (3) being infeasible. Too large a value may result in numerical ill-conditioning that could make the problem hard to solve. See Gabriel and Leuthold (2010) for an example of disjunctive constraints and some guidance on computing the constant K relative to energy modeling in the context of a two-level optimization problem of which the bottom level is an MCP.

Nash-Cournot Production Game

The next example concerns a classical Nash-Cournot production game (Shy 1995) with two producers. Such a model is applicable in a variety of areas such as energy, manufacturing, as well as many others. Also, the model can easily be extended to more than two producers with additional producer-level constraints included or marketing-clearing conditions as depicted in the network example shown below. The particular instance of this duopoly is from Gabriel (2008); Gabriel et al. (2013).

Unlike a perfectly competitive production environment, in the current setting each producer can affect market prices by adjusting its own production level. This market power feature is encoded in the objective function of each of the players $i = 1, 2$. More specifically, each player must decide on their own production level $q_i, i = 1, 2$ given that they have knowledge of the (inverse) market demand function $p(q_1 + q_2) = \alpha - \beta(q_1 + q_2)$ where $\alpha, \beta > 0$. This function gives the price of the produced good but takes into account both producers' production levels. If only one player decided to increase production, the total price for the market would go down (since $\beta > 0$) but that producer's profit might go up due to a more favorable market share. If both producers are considering production levels under these circumstances, it is not immediately clear what might be an equilibrium solution (i.e., one in which neither player has an incentive to deviate). The Nash concept is to have each player optimally solve for their production level which maximizes net profit (or other suitable function), given that the other player's level is fixed at its own optimal level.

Assuming for ease of presentation that both players have a linear production cost function given by $c_i(q_i) = \delta_i q_i$ for $\delta_i > 0, i = 1, 2$, then the resulting profit-maximization problem that player i solves is:

$$\max_{q_i} p(q_1 + q_2) \cdot q_i - c_i(q_i) \tag{11a}$$

$$s.t.\ q_i \geq 0 \tag{11b}$$

or

$$\max_{q_i} (\alpha - \beta(q_1 + q_2)) \cdot q_i - \delta_i q_i \tag{12a}$$

$$s.t.\ q_i \geq 0 \tag{12b}$$

The KKT conditions are both necessary and sufficient for this problem. Necessity follows by the linearity of the constraints and sufficiency is because the objective function is a (strictly) concave function of q_i (in addition to the linear constraints). To see the concavity result, note that the second derivative of the objective function relative to q_i is just $-2\beta < 0$ (Bazaraa et al. 1993). The resulting KKT conditions for each player taken together form a Nash-Cournot equilibrium and are given as:

$$0 \leq 2\beta q_1 + \beta q_2 - \alpha + \delta_1, q_1 \geq 0,$$

$$(2\beta q_1 + \beta q_2 - \alpha + \delta_1) \cdot q_1 = 0 \text{ (producer 1)} \tag{13a}$$

$$0 \leq \beta q_1 + 2\beta q_2 - \alpha + \delta_2, q_2 \geq 0,$$

$$(\beta q_1 + 2\beta q_2 - \alpha + \delta_2) \cdot q_2 = 0 \text{ (producer 2)} \tag{13b}$$

These conditions taken together constitute the following pure linear complementarity problem with function F:

$$F\begin{pmatrix} q_1 \\ q_2 \end{pmatrix} = \begin{pmatrix} 2\beta q_1 + \beta q_2 - \alpha + \delta_1 \\ \beta q_1 + 2\beta q_2 - \alpha + \delta_2 \end{pmatrix} = \begin{pmatrix} 2\beta & \beta \\ \beta & 2\beta \end{pmatrix}\begin{pmatrix} q_1 \\ q_2 \end{pmatrix} + \begin{pmatrix} -\alpha + \delta_1 \\ -\alpha + \delta_2 \end{pmatrix}.$$

If one can assume that both quantities $q_i > 0$ in an equilibrium solution, then the above conditions become a set of two unknowns (production) in two equations, namely:

$$0 = 2\beta q_1 + \beta q_2 - \alpha + \delta_1, \text{ (producer 1)} \tag{14a}$$

$$0 = \beta q_1 + 2\beta q_2 - \alpha + \delta_2, \text{ (producer 2)} \tag{14b}$$

Solving for the positive production levels in these two equations amounts to using a best reaction or best response function (Osborne and Rubinstein 1994), essentially closed-form expressions for q_i as a function of the other production quantity. Due to limitations on assuming positive production for all producers, or for example, the need for considering more challenging constraints (apart from nonnegativity), it is much more efficient to use the complementarity approach. Indeed, for more realistic models, it is not always possible to use this best response approach. Some MCP examples with

realism that have been presented (in energy markets) include Hobbs (2001), Gabriel et al. (2005a), Gabriel et al. (2005b), Hobbs et al. (2008).

PIES Energy Equilibrium

While the above models have considered either one or more optimization problems or nonlinear equations as instances of MCPs, the PIES (Project Independence Evaluation System) energy planning model is an important example that combines both these two approaches into a complementarity problem (Hogan 1975; Josephy 1979; Ahn 1979; Ahn and Hogan 1982). A stylized version of a much later and more complicated generation of PIES is the National Energy Modeling Systems (NEMS). NEMS has also been shown to be an instance of an MCP (Gabriel et al. 2001) and is currently used by the U.S. Department of Energy for a variety of energy market studies and reports.

The PIES model considers the supply and demand sides of the energy market separately with x denoting the vector of decision variables in energy production. The supply side is modeled as a production cost minimization given as follows with $\boldsymbol{c}$ a vector of costs conformal with $\boldsymbol{x}$:

$$\min_{\boldsymbol{x}} \boldsymbol{c}^T\boldsymbol{x} \tag{15a}$$

$$s.t.\ \boldsymbol{A}\boldsymbol{x} \geq \boldsymbol{q} \qquad (\pi) \tag{15b}$$

$$\boldsymbol{B}\boldsymbol{x} \geq \boldsymbol{b} \qquad (\gamma) \tag{15c}$$

$$\boldsymbol{x} \geq 0 \tag{15d}$$

In this linear-programming problem, besides the nonnegativity restrictions, there are two sets of constraints: meeting demand q by a combination of energy production types Ax (15b) and other, non-demand related conditions to be met (15c). The dual prices are π and γ in (15b) and (15c), respectively. As opposed to a straightforward linear program, these two prices will enter directly into another part of the MCP formulation for this problem. In particular, the demand side of the energy market is given by econometric equations of the following form where p is price:

$$q_i(p) = q_i^0 \prod \left(\frac{p_i}{p_i^0}\right)^{e_{ij}} \tag{16}$$

Here q_i^0, p_i^0 are, respectively, reference demands and prices for energy product i, and e_{ij} is an elasticity between energy products i and j. The supply and demand sides of the energy market are combined by the equilibration condition:

$$\pi = p \tag{17}$$

In a nutshell, this condition states that the price used in the demand equation (16) should reflect the value of the resources involved, i.e., be the dual price to the demand equation from (15b).

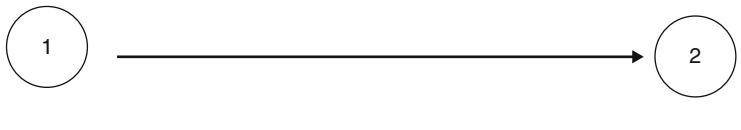
One way to join these two parts of the energy market is to substitute $\pi = p$ into (16) and then restrict π to be the dual vector to (15b). This can be done by considering the (necessary and sufficient) optimality conditions to (15) taking q as a fixed quantity. Then, the formula for q from (16) is used. The resulting MCP function for PIES (Cottle et al. 1992) is thus the following:

$$F\begin{pmatrix} x \\ \pi \\ \gamma \end{pmatrix} = \begin{pmatrix} c - A^T\pi - B^T\gamma \\ Ax - q(\pi) \\ Bx - b \end{pmatrix}$$

To try to directly incorporate both sides of the market as described above with just an optimization problem is not possible. In particular, to compute the vector of dual prices π, the LP (15) must first be solved. But to solve it, the optimal demand quantity $q = q(\pi)$ is required which in turn depends on the optimal π via (16) and (17). The MCP formulation gets around this computational issue by simultaneously determining both primal and dual variables. This feature is a strength of complementarity problems and thus easily allows combining both equations and optimization problems in one formulation.

Market Equilibrium with Underlying Network

The Nash-Cournot model described above as well as it's n-player counterpart assumes that each player has some ability to manipulate market prices by adjusting their own production levels. In fact, it is only in the objective functions of the players' problems that their separate decision variables interact. Two interesting variations on this paradigm are: Generalized Nash



Complementarity Applications, Fig. 2 Sample Two-Node Network

equilibrium problems (Facchinei and Pang 2003) and network equilibria. Generalized Nash problems allow for other players' variables to enter into the constraint set of a player. Using the two-player example shown above, in a Generalized Nash version, there might be a common constraint for each player of the form:

$$q_1 + q_2 + Inv \geq Dem \tag{18}$$

where *Inv* is the amount of inventory (for a region) and *Dem* its demand level. This common constraint would then say that the total supply (production + inventory) must at least meet the demand level. These sorts of problems belong to a class of mathematical programs that are generally harder to solve than MCPs but under certain conditions on these common constraints are expressable as complementarity problems or the related variational inequality problem. See Harker (1991) for a discussion of this result, as well as Facchinei and Pang (2003) for a theoretical treatment of variational inequalities and extensions that relate to Generalized Nash problems.

Another variation on the previous Nash equilibrium is to have multiple players each optimizing their objectives but without the ability to directly influence prices, i.e., they are price-takers. Rather, there is a market-clearing equation (or equations) whose dual variables are the associated market prices. Since the decision variables for each player contribute to the market-clearing conditions, these players have some indirect influence on the prices. As compared to the Generalized Nash model, these conditions do not appear in the constraint set of the players though. Often such problems have an underlying infrastructure network related perhaps to distribution of energy, water, or other products to transport.

Consider the following sample network equilibrium problem from Gabriel et al. (2013). There are two nodes in the network as depicted in Fig. 2. These nodes can represent cities, countries, regions, or just a market for a particular product. Production can occur at either node but only node 2 can receive additional product from the other node as indicated by the uni-directional arc. The product in question could be energy (e.g., electricity), fuels (natural gas, oil, coal), treated water, manufactured goods (e.g., televisions) or raw materials to name a few choices. There are a number of key questions that such an equilibrium model should answer.

For example, in meeting the demand at node 2, how much should be locally produced at node 2 and how much should be imported from the other node? What will be the equilibrium prices at each node if all players are acting in their own interests to maximize profits?

The production aspects can be modeled by the following optimization problem (shown here for producer A) in which net profit is to be maximized subject to production, balance, and nonnegativity constraints:

$$\max_{s_1^A, q_1^A, f_{12}^A} \pi_1 s_1^A + \pi_2 f_{12}^A - c_1^A\left(q_1^A\right) - \left(\tau_{12}^{\text{Reg}} + \tau_{12}\right) f_{12}^A \tag{19a}$$

$$s.t.\ q_1^A \leq \overline{q}_1^A \ \left(\lambda_1^A\right) \tag{19b}$$

$$s_1^A = q_1^A - f_{12}^A \ \left(\delta_1^A\right) \tag{19c}$$

$$s_1^A, q_1^A, f_{12}^A \geq 0 \tag{19d}$$

where

- $p \in \{A, B, C, D\}$ is the index for the producers
- $i \in \{1, 2\}$ is the index for the nodes
- q_n^p is the production quantity for producer p at node n
- $\overline{q}_n^p$ is the maximum production capacity for producer p at node n
- s_n^p is the amount sold by producer p at node n
- f_{12}^A, f_{12}^B are respectively, the amount of exports from node 1 to 2 by producers A and B (the other two producers do not have that option)
- π_n is the price at node n determined by market-clearing conditions
- $\tau_{12}^{Reg}, \tau_{12}$ are respectively, the exogenous, regulated export tariff when sending product from node 1 to 2 and the endogenously-determined congestion tariff between the two nodes (but exogenous from the perspective of the producer's optimization problem)

- $c_n^p\left(q_n^p\right)$ is the (marginal) production cost function for producer p at node n. For simplicity, this function is assumed linear and of the form $c_n^p\left(q_n^p\right) = \gamma_n^p q_n^p$ with $\gamma_n^p > 0$
- λ_n^p, δ_n^p are Lagrange multipliers (e.g., dual variables) for the associated constraints

Producer B's problem is similar to the one for A but the other two producers do not have any export-related terms. Since each producer is solving a linear program, the KKT conditions are both necessary and sufficient (Bazaraa et al. 1993). These KKT conditions for each of the four producers are as follows.

Producer A, node 1

$$0 \le -\pi_1 + \delta_1^A \perp s_1^A \ge 0 \tag{20a}$$

$$0 \le \gamma_1^A + \lambda_1^A - \delta_1^A \perp q_1^A \ge 0 \tag{20b}$$

$$0 \le -\pi_2 + \left(\tau_{12}^{Reg} + \tau_{12}\right) + \delta_1^A \perp f_{12}^A \ge 0 \tag{20c}$$

$$0 \le \overline{q}_1^A - q_1^A \perp \lambda_1^A \ge 0 \tag{20d}$$

$$0 = s_1^A - q_1^A + f_{12}^A, \delta_1^A \text{ free} \tag{20e}$$

Producer B, node 1

$$0 \le -\pi_1 + \delta_1^B \perp s_1^B \ge 0 \tag{21a}$$

$$0 \le \gamma_1^B + \lambda_1^B - \delta_1^B \perp q_1^B \ge 0 \tag{21b}$$

$$0 \le -\pi_2 + \left(\tau_{12}^{Reg} + \tau_{12}\right) + \delta_1^B \perp f_{12}^B \ge 0 \tag{21c}$$

$$0 \le \overline{q}_1^B - q_1^B \perp \lambda_1^B \ge 0 \tag{21d}$$

$$0 = s_1^B - q_1^B + f_{12}^B, \delta_1^B \text{ free} \tag{21e}$$

Producer C, node 2

$$0 \le -\pi_2 + \delta_2^C \perp s_2^C \ge 0 \tag{22a}$$

$$0 \le \gamma_2^C + \lambda_2^C - \delta_2^C \perp q_2^C \ge 0 \tag{22b}$$

$$0 \le \overline{q}_2^C - q_2^C \perp \lambda_2^C \ge 0 \tag{22c}$$

$$0 = s_2^C - q_2^C, \delta_2^C \text{ free} \tag{22d}$$

Producer D, node 2

$$0 \le -\pi_2 + \delta_2^D \perp s_2^D \ge 0 \tag{23a}$$

$$0 \le \gamma_2^D + \lambda_2^D - \delta_2^D \perp q_2^D \ge 0 \tag{23b}$$

$$0 \le \overline{q}_2^D - q_2^D \perp \lambda_2^D \ge 0 \tag{23c}$$

$$0 = s_2^D - q_2^D, \delta_2^D \text{ free} \tag{23d}$$

The market-clearing conditions forcing supply to equal demand are:

$$0 = \left[s_1^A + s_1^B\right] - D_1(\pi_1),\ \pi_1 \text{ free} \tag{24a}$$

$$0 = \left[s_2^C + s_2^D + f_{12}^A + f_{12}^B\right] - D_2(\pi_2),\ \pi_2 \text{ free} \tag{24b}$$

where $\left[s_1^A + s_1^B\right]$, $\left[s_2^C + s_2^D + f_{12}^A + f_{12}^B\right]$ are the supply amounts for nodes 1 and 2, respectively and $D_n(\pi_n), n = 1, 2$ are the demands at each node taking into account the nodal price π_n.

Besides production and market-clearing, in some applications (e.g., energy, water) there is a player that makes sure the network is running smoothly. This network system operator (NSO) also solves an optimization problem which can take on a variety of forms maximizing for example, social welfare or net profit to name two. Using net profit, a stylized network operator problem is as follows:

$$\max_{g_{12}} \left(\tau_{12}^{\text{Reg}} + \tau_{12}\right) g_{12} - c^{NSO}(g_{12}) \tag{25a}$$

$$s.t.\ \ g_{12} \le \overline{g}_{12}\ (\varepsilon_{12}) \tag{25b}$$

$$g_{12} \ge 0 \tag{25c}$$

where g_{12} represents the flow from node 1 to node 2 that the NSO manages, $c^{NSO}(g_{12})$ is a network operations cost function (assume linear for simplicity, i.e., $c^{NSO}(g_{12}) = \gamma^{NSO} g_{12}$ where $\gamma^{NSO} > 0$) and ε_{12} is the dual variable associated with the capacity constraint involving the flow upper bound $\overline{g}_{12}$. Like the producers' problems, this is a linear program so that the KKT conditions are both necessary and sufficient and are the following:

$$0 \le -\tau_{12}^{\text{Reg}} - \tau_{12} + \gamma^{NSO} + \varepsilon_{12} \perp g_{12} \ge 0 \tag{26a}$$

$$0 \le \overline{g}_{12} - g_{12} \perp \varepsilon_{12} \ge 0 \tag{26b}$$

To determine the congestion tariff τ_{12}, the following market-clearing conditions can be used:

$$0 = g_{12} - \left[f_{12}^A + f_{12}^B\right],\ \tau_{12}\ \text{free} \tag{27}$$

The overall market equilibrium on this network can be expressed as an MCP by collecting the KKT conditions of the producers: (20)–(23) the supply–demand market-clearing conditions (24), the KKT conditions of the NSO (26) and the market-clearing conditions of the network flows (27).

As discussed in Gabriel et al. (2013), suppose the following input data are used.

$$\begin{aligned}
\tau_{12}^{\text{Reg}} &= 0.5\\
\gamma_1^A = 10, \gamma_1^B &= 12, \gamma_2^C = 15, \gamma_2^D = 18\\
a_1 = 20, b_1 &= 1, a_2 = 40, b_2 = 2\\
\overline{q}_1^A = 10, \overline{q}_1^B &= 10, \overline{q}_2^C = 5, \overline{q}_2^D = 5\\
\overline{g}_{12} &= 5\\
\gamma^{NSO} &= 1\\
D_1(\pi_i) &= a_i - b_i\pi_i
\end{aligned}$$

Then, an MCP solution in terms of production quantities, flows, prices, and tariffs is as follows:

$$\begin{aligned}
&q_1^A = 10, q_1^B = 3, q_2^C = 5, q_2^D = 0\\
&f_{12}^A = 2.561, f_{12}^B = 2.439\\
&\quad (\text{the sum is } 5 = \text{the capacity of the link})\\
&\pi_1 = 12, \pi_2 = 15\\
&\tau_{12} = 2.5\\
&\quad (\tau_{12}^{\text{Reg}} + \tau_{12} = 3, \text{ the difference in the nodal prices})
\end{aligned}$$

Traffic Equilibrium

One of the classical problems in complementarity modeling is that of predicting steady state flows of cars (or other vehicles) along a congested road. Consider as a simple example, an origin (node 1) and two destinations (nodes 4 and 5) as well as intermediate nodes 2 and 3 as shown in Fig. 3 Wardrop 1952; Aashtiani and Magnanti 1981; Magnanti 1984; Florian 1986, 1989). These nodes can relate to intersection points, cities, regions, etc. The idea is to try to predict how many drivers will be using the individual paths in the network if for example, the price (e.g., time, disutility) of a particular path is taken into account in the decision-making process. That is, if the flow is price-based.

In this simple example, there are two origin–destination (OD) pairs: (1, 4) and (1, 5) which represent where drivers begin and end their trip. In going from node 1 to node 4, drivers can choose either to travel along path 1-2-4 or 1-3-4; for the OD pair (1, 5) there is only one path: 1-2-5. Wardrop (1952) stated an equilibrium where no driver had an incentive to deviate from a particular chosen path resulting in:

- Paths with positive flow serving the same OD pair all having equal costs (otherwise drivers would deviate to the less costly ones)
- Paths with costs higher than the minimum having no flow

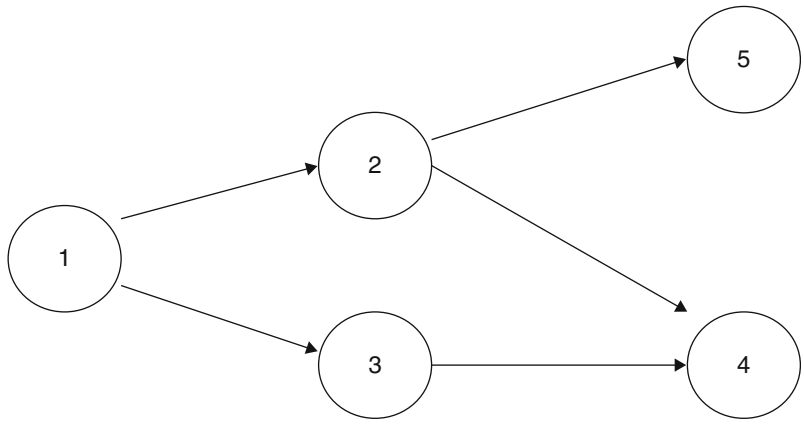
Essentially, such an equilibrium needs to take into account that all drivers are doing what is in their own best interests but that there should be no incentive for any one driver to deviate from a path they pick on their own. As compared to the previous examples of MCPs, in this case there is no explicit optimization problem(s), just an indirect acknowledgement that drivers want to minimize the time or cost of the path chosen.

To present the associated complementarity problem, it is necessary to define some related terms. First, path flows on a path p will be denoted by h_p, e.g., the flow on path 1-2-4 is h_{1-2-4}. The vector of all path flows is given by h which for this example is lexicographically given as:

$$h = \begin{pmatrix} h_{1-2-4} \\ h_{1-2-5} \\ h_{1-3-4} \end{pmatrix}.$$

Flows on an arc a are given by f_a with the vector of all such flows denoted as f. For the sample network shown above:

$$f = \begin{pmatrix} f_{1,2} \\ f_{1,3} \\ f_{2,4} \\ f_{2,5} \\ f_{3,4} \end{pmatrix}.$$



Complementarity Applications, Fig. 3 Sample Network for Traffic Equilibrium Problem

Both h and f are related by the equation:

$$f = \Delta h \tag{28}$$

where $\Delta = \left[\delta_{ap}\right]$ is the arc-path incidence matrix with $\delta_{ap} = 1$ if arc a is on path p and is equal to zero otherwise. Thus, for the network shown above:

$$\Delta = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Parallel to the flow vectors h and f are the cost vectors for paths and arcs, given respectively as $C(h)$ and $c(f)$ which for the sample network are as follows:

$$C(h) = \begin{pmatrix} C_{1-2-4}(h) \\ C_{1-2-5}(h) \\ C_{1-3-4}(h) \end{pmatrix}, c(f) = \begin{pmatrix} c_{1,2}(f) \\ c_{1,3}(f) \\ c_{2,4}(f) \\ c_{2,5}(f) \\ c_{3,4}(f) \end{pmatrix}.$$

The relationship between these two vectors is:

$$C(h) = \Delta^T c(f) \tag{29}$$

indicating that the paths costs are the sum of the arc costs for those arcs on the path. This is the standard additive model which is well-studied but not always realistic when one considers for example tolls. In that case, nonadditive approaches such as those given in Gabriel and Bernstein (1997) and Bernstein and Gabriel (1997) may be more appropriate. An important point to note is that the path costs for a particular path are a function of all the path flows in the network. Likewise, the arc costs for a given arc are a function of all the arc flows. This is a very realistic representation of the network indicating the interaction effects. A more restrictive version is to assume that path p (arc a) only affects the costs for that path (arc) in essence a separability argument. This was an initial assumption used early on in part because it led to solving an equivalent optimization problem (Magnanti 1984; Florian 1989) which was easier to solve before the large growth in complementarity problem algorithms in the 1990s.

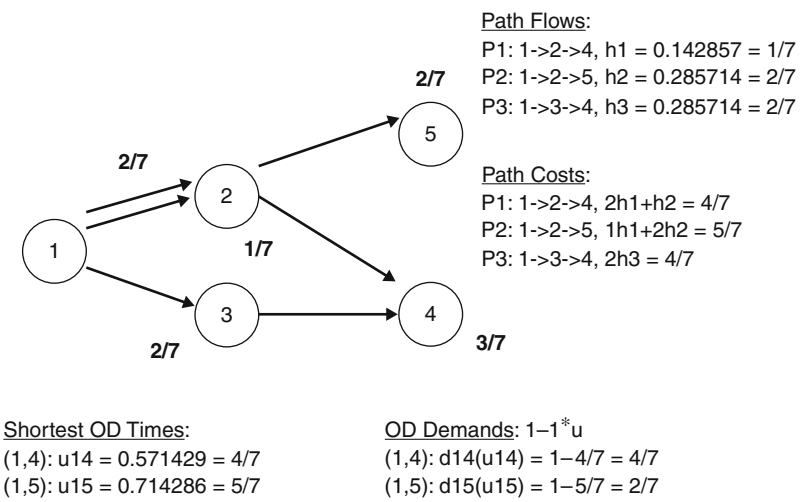
Besides the flow, the complementarity problem associated with a traffic equilibrium also needs to account for meeting the OD demand. For each OD pair i, such demand is denoted by D_i which itself is a function of the shortest time u_i (or least disutility) between the origin and destination i. In the network from Fig. 3, the vector versions of these quantities are thus:

$$D(u) = \begin{pmatrix} D_{(1,4)}(u) \\ D_{(1,5)}(u) \end{pmatrix}, \text{ with } u = \begin{pmatrix} u_{(1,4)} \\ u_{(1,5)} \end{pmatrix}.$$

There is one last notational element to define: the path-OD pair incidence matrix $\Gamma = \left[\gamma_{pi}\right]$ where $\gamma_{pi} = 1$ if path p serves OD pair i and is equal to zero otherwise. For the example above,

$$\Gamma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}$$

for paths 1-2-4, 1-2-5, 1-3-4, and OD pairs, (1,4) and (1,5), respectively.



Complementarity Applications, Fig. 4 Solution to Sample Traffic Equilibrium Problem

The formal statement of the (additive) traffic equilibrium problem is thus to find path flows h and shortest times u such that:

$$0 \leq C_p(h) - u_i \text{ for all } p \in P_i, i \in I, \; h \geq 0 \tag{30a}$$

$$0 = \left(C_p(h) - u_i\right) \cdot h_p \text{ for all } p \in P_i, i \in I \tag{30b}$$

$$0 = \sum_{p \in P_i} h_p - D_i(u) \text{ for all } i \in I, u \geq 0 \tag{30c}$$

where P_i is the set of paths that serve OD pair i and I is the set of OD pairs. Equation (30a) simply states that the path cost $C_p(h)$ must be by definition greater than or equal to the shortest time u_i for all paths serving that OD pair i; also only nonnegative path flows are allowed. Equation (30b) is a translation of the Wardrop statement that appeared above. Namely, if the path p has any positive flow then the path cost must be equal to the shortest time and this must be true for all paths serving that OD pair. Also, if the path cost is strictly greater than the shortest time, there should be no flow along that path. Lastly, (30c) indicates that the total path flow across all paths that serve OD pair i must equal the demand; also, only nonnegative shortest times u_i are allowed. As stated, (30) is not an MCP since the equations in (3) do not match up with free variables u. As shown in Aashtiani and Magnanti (1981), (30c) can be relaxed to $0 \leq \sum_{p \in P_i} h_p - D_i(u)$ for all $i \in I, u \geq 0$ as long as some mild conditions on the path cost and demand functions hold and the corresponding MCP will have a solution that matches up with (30). Another important result is that if the demand function is invertible (or just fixed demand), an arc formulation instead of the more cumbersome path version can be used (Magnanti 1984; Florian 1989). In that case, taking into account (28) and (29), the resulting MCP is as follows:

$$0 \leq \left(\Delta^T c(\Delta h) - \Gamma u\right) \perp h \geq 0 \tag{31a}$$

$$0 \leq \Gamma^T h - D(u) \perp u \geq 0 \tag{31b}$$

assuming the mild restrictions on C and D are also in effect.

To make this formulation (31) concrete, consider the following specific choice for costs and demand functions for the sample network shown above:

$$c_a(f) = f_a \tag{32a}$$

$$D_i(u) = 1 - 1u_i \tag{32b}$$

The resulting solution is shown in Fig. 4. Note that both paths 1-2-4 and 1-3-4 serve OD pair (1,4) and since they both have positive flow ($h_{1-2-4} = \frac{1}{7}$, $h_{1-3-4} = \frac{2}{7}$), by Wardrop's principle they should both

have the same costs and equal to the lowest cost (shortest time) $u_{(1,4)}$ as is shown below. But from (28):

$$f = \begin{pmatrix} f_{1,2} \\ f_{1,3} \\ f_{2,4} \\ f_{2,5} \\ f_{3,4} \end{pmatrix} = \Delta h = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1/7 \\ 2/7 \\ 2/7 \end{pmatrix} = \begin{pmatrix} 3/7 \\ 2/7 \\ 1/7 \\ 2/7 \\ 2/7 \end{pmatrix}$$

$$C_{1-2-4}(h) = \frac{3}{7} + \frac{1}{7} = \frac{4}{7}$$
$$C_{1-3-4}(h) = \frac{2}{7} + \frac{2}{7} = \frac{4}{7}$$
$$u_{(1,4)} = \frac{4}{7}$$
$$C_{1-2-5} = \frac{3}{7} + \frac{2}{7} = \frac{5}{7}$$
$$u_{(1,4)} = \frac{5}{7}$$

Concluding Remarks

In this article, complementary problems have been defined and their relevance to certain infrastructure models has been emphasized. Complementarity problems generalize optimization, game theory, and a host of other interesting problems in engineering and economics. As such, this flexible class of mathematical programs has great relevance to many important operations research problems.

See

- Bilevel Linear Programming
- Complementarity Problems
- Complementary Slackness Theorem
- Constrained Optimization Problem
- Constraint Qualification
- Convex Optimization
- Dual Linear-Programming Problem
- Duality Theorem
- Game Theory
- Integer and Combinatorial Optimization
- Karush-Kuhn-Tucker (KKT) Conditions
- Lagrange Multipliers
- Linear Programming
- Network Optimization
- Nonlinear Programming

References

Aashtiani, H. Z., & Magnanti, T. L. (1981). Equilibria on a congested transportation network. *SIAM Journal on Algebraic Discrete Methods, 2*, 213–226.

Ahn, B. H. (1979). *Computation of market equilibria for policy analysis*. New York: Garland Publishing, Inc.

Ahn, B. H., & Hogan, W. W. (1982). On convergence of the PIES algorithm for computing equilibria. *Operations Research, 30*, 281–300.

Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (1993). *Nonlinear programming theory and algorithms*. New York: John Wiley & Sons, Inc.

Bernstein, D., & Gabriel, S. A. (1997). Solving the nonadditive traffic equilibrium problem. In P. M. Pardalos, D. W. Hearn, & W. W. Hager (Eds.), *Lecture notes in economics and mathematical systems network optimization* (pp. 72–102). Berlin/New York: Springer.

Cottle, R. W., Pang, J.-S., & Stone, R. E. (1992). *The linear complementarity problem*. San Diego: Academic Press.

Facchinei, F., & Pang, J. S. (2003). *Finite-dimensional variational inequalities and complementarity problems volumes I and II*. New York: Springer.

Ferris, M. C., & Pang, J. S. (Eds.). (1997). *Complementarity and variational problems state of the art*. Baltimore: SIAM.

Ferris, M. C., Mangasarian, O. L., & Pang, J. S. (Eds.). (2001). *Complementarity: Applications, algorithms and extensions*. Dordrecht: Kluwer Academic Publishers.

Florian, M. (1986). Nonlinear cost network models in transportation analysis. *Mathematical Programming Studies, 26*, 167–196.

Florian, M. (1989). Mathematical programming applications in national, regional and urban planning. In M. Iri & K. Tanabe (Eds.), *Mathematical programming: Recent developments and applications* (pp. 57–81). Tokyo: Kluwer Academic Publishers.

Fortuny-Amat, J., & McCarl, B. (1981). A representation and economic interpretation of a two-level programming problem. *Journal of the Operational Research Society, 32*(9), 783–792.

Gabriel, S. A. (2008). Optimization and equilibrium models in energy. College Park, MD: Department of Civil and Environmental Engineering, University of Maryland. Manuscript, December 12, 2008.

Gabriel, S. A., & Bernstein, D. (1997). The traffic equilibrium problem with nonadditive path costs. *Transportation Science, 31*(4), 337–348.

Gabriel, S. A., & Leuthold, F. U. (2010). Solving discretely-constrained MPEC problems with applications in electric power markets. *Energy Economics, 32*, 3–14.

Gabriel, S. A., Kydes, A. S., & Whitman, P. (2001). The national energy modeling system: A large-scale energy-economic equilibrium model. *Operations Research, 49*(1), 14–25.

Gabriel, S. A., Kiet, S., & Zhuang, J. (2005a). A mixed complementarity-based equilibrium model of natural gas markets. *Operations Research, 53*(5), 799–818.
Gabriel, S. A., Zhuang, J., & Kiet, S. (2005b). A large-scale complementarity model of the north American natural gas market. *Energy Economics, 27*, 639–665.
Gabriel, S. A., Conejo, A. J., Fuller, J. D., Hobbs, B. F., & Ruiz, C. (2013). *Complementarity modeling in energy markets*. New York: Springer. Chapter 1.
Harker, P. T. (1991). Generalized Nash games and quasi-variational inequalities. *European Journal of Operational Research, 54*(1), 81–94.
Harker, P. T. (1993). *Lectures on computation of equilibria with equation-based methods* (CORE lecture series). Louvain-La-Neuve: CORE Foundation.
Harker, P. T., & Pang, J.-S. (1990). Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory. *Algorithms and Applications, Mathematical Programming, 48*, 161–220.
Hobbs, B. F. (2001). Linear complementarity models of nash-cournot competition in bilateral and POOLCO power markets. *IEEE Transactions on Power Systems, 16*(2), 194–202.
Hobbs, B. F., Drayton, G., Fisher, E. B., & Lise, W. (2008). Improved transmission representations in oligopolistic market models: Quadratic losses, phase shifters, and DC lines. *IEEE Transactions on Power Systems, 23*(3), 1018–1029.
Hogan, W. W. (1975). Energy policy models for project independence. *Computers & Operations Research, 2*, 251–271.
Josephy, N. H. (1979). *Hogan's PIES example and Lemke's algorithm*. University of Wisconsin-Madison: Mathematics Research Center.
Luenberger, D. G. (1984). *Linear and nonlinear programming* (2nd ed.). Reading, MA: Addison-Wesley.
Luo, Z. Q., Pang, J.-S., & Ralph, D. (1996). *Mathematical programs with equilibrium constraints*. Cambridge, UK: Cambridge University Press.
Magnanti, T. L. (1984). Models and algorithms for predicting urban traffic equilibria. In M. Florian (Ed.), *Transportation planning models* (pp. 153–186). (North-Holland), Amsterdam: Elsevier Science Publishers B.V.
Osborne, M. J., & Rubinstein, A. (1994). *A course in game theory*. Cambridge, MA: The MIT Press.
Shy, O. (1995). *Industrial organization theory and applications*. Cambridge, MA: The MIT Press.
Wardrop, J. G. (1952). Some theoretical aspects of road traffic research. *Proceedings of the ICE Part II, 1*, 325–378.

Complementarity Condition

A relation between two nonnegative vectors in which, whenever a given component of one of the vectors is positive, the corresponding component of the other vector must be zero. For example, two nonnegative n-dimensional vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ satisfy a complementarity condition if their ith components are such that $x_i\, y_i = 0$, $i = 1, \ldots, n$.

See

- ▶ Complementarity Applications
- ▶ Complementarity Problems
- ▶ Complementary Slackness Theorem

Complementarity Problems

Richard W. Cottle
Stanford University, Stanford, CA, USA

Introduction

In its most elementary form, a complementarity problem CP($\boldsymbol{f}$) is an inequality system stated in terms of a mapping $\boldsymbol{f} : R^n \to R^n$. Given $\boldsymbol{f}$, one seeks a vector $\boldsymbol{x} \in R^n$ such that

$$\boldsymbol{x}_i \geq 0, \quad f_i(\boldsymbol{x}) \geq 0, \quad \text{and} \quad x_i f_i(\boldsymbol{x}) = 0 \quad i = 1, \ldots, n. \tag{1}$$

When the mapping $\boldsymbol{f}$ is affine, say of the form $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{q} + \boldsymbol{M}\boldsymbol{x}$, problem (1) is called a linear complementarity problem, denoted LCP$(\boldsymbol{q}, \boldsymbol{M})$ or sometimes just $(\boldsymbol{q}, \boldsymbol{M})$. Otherwise, it is called a nonlinear complementarity problem and is denoted NCP($\boldsymbol{f}$).

If $\bar{\boldsymbol{x}}$ is a solution to (1) satisfying the additional nondegeneracy condition $\bar{x}_i + f_i(\bar{\boldsymbol{x}}) > 0, i = 1, \ldots, n$, the indices i for which $\bar{x}_i > 0$ or $f_i(\bar{\boldsymbol{x}}) > 0$ form complementary subsets of $\{1, \ldots, n\}$. This is believed to be the origin of the term complementary slackness as used in linear and nonlinear programming. It was this terminology that inspired the name complementarity problem.

Sources of Complementarity Problems

The complementarity problem is intimately linked to the Karush-Kuhn-Tucker necessary conditions of local optimality found in mathematical programming

theory. This connection was brought out in Cottle (1964, 1966) and again in Cottle and Dantzig (1968). Finding solutions to such systems was one of the original motivations for studying the subject. Another was the finding of equilibrium points in bimatrix and polymatrix games. This kind of application was emphasized by Howson (1963) and Lemke and Howson (1964). These early contributions also included essentially the first algorithms for these types of problems. There are numerous applications of the linear and nonlinear complementarity problems in computer science, economics, various engineering disciplines, finance, game theory, statistics, and mathematics. Descriptions of—and references to—these applications can be found in the books by Murty (1988), Cottle et al. (1992, 2009), Isac (1992), Isac et al. (2002), and Facchinei and Pang (2003). The survey article by Ferris and Pang (1997) is the richest compendium yet published on engineering and economic applications of complementarity problems.

Equivalent Formulations

The problem CP($\boldsymbol{f}$) can be formulated in several equivalent ways. An obvious one calls for a solution $(\boldsymbol{x}, \boldsymbol{y})$ to the system

$$\boldsymbol{y} - \boldsymbol{f}(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \geq 0, \quad \boldsymbol{y} \geq 0, \quad \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = 0. \tag{2}$$

Another is to find a zero $\boldsymbol{x}$ of the mapping

$$\boldsymbol{g}(\boldsymbol{x}) = \min\{\boldsymbol{x}, \boldsymbol{f}(\boldsymbol{x})\} \tag{3}$$

where the symbol $\min\{\boldsymbol{a}, \boldsymbol{b}\}$ denotes the componentwise minimum of the two n-vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. Yet another equivalent formulation asks for a fixed point of the mapping

$$\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{x} - \boldsymbol{g}(\boldsymbol{x}),$$

i.e., a vector $\boldsymbol{x} \in R^n$ such that $\boldsymbol{x} = \boldsymbol{h}(\boldsymbol{x})$.

The formulation of CP($\boldsymbol{f}$) given in (3) is related to the (often nonconvex) optimization problem:

$$\begin{aligned} &\text{minimize } \boldsymbol{x}^{\mathrm{T}} f(\boldsymbol{x}) \\ &\text{subject to } \boldsymbol{f}(\boldsymbol{x}) \geq 0 \\ &\qquad \boldsymbol{x} \geq 0 \end{aligned} \tag{4}$$

In such a problem, the objective is bounded below by zero, thus any feasible solution of (4) for which the objective function $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{f}(\boldsymbol{x}) = 0$ must be a global minimum as well as a solution of CP($\boldsymbol{f}$). As it happens, there are circumstances (for instance, the monotonicity of the mapping $\boldsymbol{f}$) under which all the local minima for the mathematical programming problem (4) must in fact be solutions of (3).

Also noteworthy is a result of Eaves and Lemke (1981) showing that the LCP is equivalent to solving a system of equations $\boldsymbol{y} = \varphi(\boldsymbol{x})$ where the mapping $\varphi : R^n \to R^n$ is piecewise linear. In particular, LCP($\boldsymbol{q}$, $\boldsymbol{M}$) is equivalent to finding a vector $\boldsymbol{u}$ such that

$$\boldsymbol{q} + \boldsymbol{M}\boldsymbol{u}^{+} - \boldsymbol{u}^{-} = \boldsymbol{0}$$

where (for $i = 1, \ldots, n$) $u_i^{+} = \max\{0, u_i\}$ and $u_i^{-} = -\min\{0, u_i\}$.

The Linear Complementarity Problem

The LCP has quite an extensive literature, far more so than the NCP. This is most likely attributable to the LCP's relatively greater accessibility. Within this field of study, there are several main directions: the existence and uniqueness (or number of) solutions, mathematical properties of the problem, generalizations of the problem, algorithms, applications, and implementations.

Much of the theory of the linear complementarity problem is strongly linked in various ways to matrix classes. For instance, one of the earliest theorems on the existence of solutions is due to Samelson et al. (1958). Motivated by a problem in structural mechanics, they showed that the LCP($\boldsymbol{q}$, $\boldsymbol{M}$) has a unique solution for every $\boldsymbol{q} \in R^n$ if and only if the matrix $\boldsymbol{M}$ has positive principal minors. (That is, the determinant of every principal submatrix of $\boldsymbol{M}$ is positive.) The class of such matrices has come to be known as **P**, and its members are called **P**-matrices. (It is significant that the Samelson-Thrall-Wesler theorem characterizes a class of matrices in terms of the LCP.) The class **P** includes all positive definite (**PD**) matrices, i.e., those square matrices $\boldsymbol{M}$ for which $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x} > 0$ for all $\boldsymbol{x} \neq 0$. In the context of the LCP, the term "positive definite" does not

require symmetry. An analogous definition (and usage) holds for positive semi-definite (**PSD**) matrices, namely, $\boldsymbol{M}$ is **PSD** if $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{M}\boldsymbol{x} > 0$ for all $\boldsymbol{x}$. Some authors refer to such matrices as monotone because of their connection with monotone mappings. **PSD**-matrices have the property that the associated LCPs ($\boldsymbol{q}$, $\boldsymbol{M}$) are solvable whenever they are feasible, whereas LCPs ($\boldsymbol{q}$, $\boldsymbol{M}$) in which $\mathbf{M} \in \mathbf{PD}$ are always feasible and (since $\mathbf{PD} \subset \mathbf{PSD}$) are always solvable. Murty (1968, 1972) gave this distinction a more general matrix form. He defined $\mathbf{Q}$ as the class of all square matrices for which LCP($\boldsymbol{q}$, $\boldsymbol{M}$) has a solution for all $\boldsymbol{q}$ and $\mathbf{Q}_0$ as the class of all square matrices for which LCP($\boldsymbol{q}$, $\boldsymbol{M}$) has a solution whenever it is feasible. Although the goal of usefully characterizing the classes $\mathbf{Q}$ and $\mathbf{Q}_0$ has not yet been realized, much is known about some of their special subclasses. Indeed, there are now literally dozens of matrix classes for which LCP existence theorems have been established. See Murty (1988), Cottle et al. (1992, 2009), Cottle (2010) and Isac (1992) for an abundance of information on this subject.

Algorithms for Solving LCPs

The algorithms for solving linear complementarity problems are of two major types: pivoting (or, direct) and iterative (or, indirect). Algorithms of the former type are finite procedures that attempt to transform the problem ($\boldsymbol{q}$, $\boldsymbol{M}$) to an equivalent system of the form $(\boldsymbol{q}', \boldsymbol{M}')$ in which $\boldsymbol{q}' \geq 0$. Doing this is not always possible; it depends on the problem data, usually on the matrix class (such as **P**, **PSD**, etc.) to which $\boldsymbol{M}$ belongs. When this approach works, it amounts to carrying out a principal pivotal transformation on the system of equations

$$\boldsymbol{y} = \boldsymbol{q} + \boldsymbol{M}\boldsymbol{x}.$$

To such a transformation there corresponds an index set α (with complementary index set $\bar{\alpha} = \{1, \ldots, n\} \backslash \alpha$) such that the principal submatrix $\boldsymbol{M}_{\alpha\alpha}$ is nonsingular. When this (block pivot) operation is carried out, the system

$$\begin{aligned} \boldsymbol{y}_\alpha &= \boldsymbol{q}_\alpha + \boldsymbol{M}_{\alpha\alpha}\boldsymbol{x}_\alpha + \boldsymbol{M}_{\alpha\bar{\alpha}}\boldsymbol{x}_{\bar{\alpha}} \\ \boldsymbol{y}_{\bar{\alpha}} &= \boldsymbol{q}_{\bar{\alpha}} + \boldsymbol{M}_{\bar{\alpha}\alpha}\boldsymbol{x}_\alpha + \boldsymbol{M}_{\bar{\alpha}\bar{\alpha}}\boldsymbol{x}_{\bar{\alpha}} \end{aligned}$$

becomes

$$\begin{aligned} \boldsymbol{x}_\alpha &= \boldsymbol{q}'_\alpha + \boldsymbol{M}'_{\alpha\alpha}\boldsymbol{y}_\alpha + \boldsymbol{M}'_{\alpha\bar{\alpha}}\boldsymbol{x}_{\bar{\alpha}} \\ \boldsymbol{y}_{\bar{\alpha}} &= \boldsymbol{q}'_{\bar{\alpha}} + \boldsymbol{M}'_{\bar{\alpha}\alpha}\boldsymbol{y}_\alpha + \boldsymbol{M}'_{\bar{\alpha}\bar{\alpha}}\boldsymbol{x}_{\bar{\alpha}} \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{q}'_\alpha &= -\boldsymbol{M}_{\alpha\alpha}^{-1}\boldsymbol{q}_\alpha \\ \boldsymbol{M}'_{\alpha\bar{\alpha}} &= -\boldsymbol{M}_{\alpha\alpha}^{-1} \\ \boldsymbol{M}'_{\alpha\bar{\alpha}} &= -\boldsymbol{M}_{\alpha\alpha}^{-1}\boldsymbol{M}_{\alpha\bar{\alpha}} \\ \boldsymbol{q}'_{\bar{\alpha}} &= \boldsymbol{q}_{\bar{\alpha}} - \boldsymbol{M}_{\bar{\alpha}\alpha}\boldsymbol{M}_{\alpha\alpha}^{-1}\boldsymbol{q}_\alpha \\ \boldsymbol{M}'_{\bar{\alpha}\alpha} &= \boldsymbol{M}_{\bar{\alpha}\alpha}\boldsymbol{M}_{\alpha\alpha}^{-1} \\ \boldsymbol{M}'_{\bar{\alpha}\bar{\alpha}} &= \boldsymbol{M}_{\bar{\alpha}\bar{\alpha}} - \boldsymbol{M}_{\bar{\alpha}\alpha}\boldsymbol{M}_{\alpha\alpha}^{-1}\boldsymbol{M}_{\alpha\bar{\alpha}}. \end{aligned}$$

There are two main pivoting algorithms used in processing LCPs. The more robust of the two is due to Lemke (1965). Lemke's method embeds the LCP($\boldsymbol{q}$, $\boldsymbol{M}$) in a problem having an extra "artificial" nonbasic (independent) variable x_0 with coefficients specially chosen so that when x_0 is sufficiently large, all the basic variables become nonnegative. At the least positive value of x_0 for which this is so, there will (in the nondegenerate case) be (exactly) one basic variable whose value is zero. That variable is exchanged with x_0. Thereafter the method executes a sequence of (almost complementary) simple pivots. In each case, the variable becoming basic is the complement of the variable that became nonbasic in the previous exchange. The method terminates if either x_0 decreases to zero—in which case the problem is solved—or else there is no basic variable whose value decreases as the incoming nonbasic variable is increased. The latter outcome is called termination on a secondary ray. For certain matrix classes, termination on a secondary ray is an indication that the given LCP has no solution. Eaves (1971) was among the first to study Lemke's method from this point of view.

The other pivoting algorithm for the LCP is called the Principal Pivoting Method (see Cottle and Dantzig (1968)). The algorithm has two versions: symmetric and asymmetric. The former executes a sequence of principal (block) pivots or order 1 or 2, whereas the latter does sequences of almost complementary pivots, each of which results in a block principal pivot or order potentially larger than 2. The class of problems to

which the Principal Pivoting Method applies is more restrictive. (See Cottle et al. (1992, 2009) for a treatment of this algorithm.)

Iterative methods are often favored for the solution of very large linear complementarity problems. In such problems, the matrix $\boldsymbol{M}$ tends to be sparse (i.e., to have a small percentage of nonzero elements) and frequently structured. Since iterative methods do not modify the problem data, these features of large-scale problems can be used to advantage. Ordinarily, however, an iterative method does not terminate finitely; instead, it generates a convergent sequence of trial solutions. As is to be expected, the applicability of algorithms in this family depends on the matrix class to which $\boldsymbol{M}$ belongs. Details on several algorithms of this type are presented in the books by Kojima et al. (1991) as well as the one by Cottle et al. (1992, 2009).

Some Generalizations

The linear and nonlinear complementarity problems have been generalized in numerous ways. One of the earliest generalizations was given by Habetler and Price (1971) and Karamardian (1971) who defined the problem $\mathrm{CP}(K, \boldsymbol{f})$ as that of finding a vector $\boldsymbol{x}$ in the closed convex cone K such that $\boldsymbol{f}(x) \in K^*$ (the dual cone) and $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{f}(\boldsymbol{x}) = 0$. Through this formulation, a connection can be made between complementarity problems and variational inequality problems, that is, problems $\mathrm{VI}(X, \boldsymbol{f})$ wherein one seeks a vector $\boldsymbol{x}^* \in X$ (a nonempty subset of R^n) such that

$$\boldsymbol{f}(\boldsymbol{x}^*)^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}^*) \geq 0 \quad \text{for all } \boldsymbol{y} \in X.$$

Karamardian (1971) established that when X is a closed convex cone, say K, with dual cone K^*, then $\mathrm{CP}(K, \boldsymbol{f})$ and $\mathrm{VI}(X, \boldsymbol{f})$ have exactly the same solutions (if any).

Robinson (1979) has considered the generalized complementarity problem $\mathrm{CP}(K, \boldsymbol{f})$ defined above as an instance of a generalized equation, namely to find a vector $\boldsymbol{x} \in R^n$ such that

$$0 \in \boldsymbol{f}(\boldsymbol{x}) + \partial\psi_K(\boldsymbol{x})$$

where ψ_K is the indicator function of the closed convex cone K and ∂ denotes the subdifferential operator as used in convex analysis.

Among the diverse generalizations of the linear complementarity problem, the earliest appears in Samelson et al. (1958). There—for given $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ and n-vector $\boldsymbol{c}$—the authors considered the problem of the finding n-vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ such that

$$\boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} = \boldsymbol{c}, \quad \boldsymbol{x} \geq 0, \quad \boldsymbol{y} \geq 0 \quad \text{and} \quad \boldsymbol{x}^{\mathrm{T}}\boldsymbol{y} = 0.$$

A different generalization was introduced by Cottle and Dantzig (1970). In this sort of problem, one has an affine mapping $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{q} + \boldsymbol{N}\boldsymbol{x}$ where $\boldsymbol{N}$ is of order $\sum_{j=1}^{k} p_j \times n$ partitioned into k blocks; the vectors $\boldsymbol{q}$ and $\boldsymbol{y} = \boldsymbol{f}(\boldsymbol{x})$ are partitioned conformably. Thus,

$$\boldsymbol{y}^j = \boldsymbol{q}^j + \boldsymbol{N}^j\,\boldsymbol{x} \text{ for } j = 1, \ldots, k.$$

The problem is to find a solution of the system

$$\boldsymbol{y} = \boldsymbol{q} + \boldsymbol{N}\boldsymbol{x}, \quad \boldsymbol{x} \geq 0, \quad \boldsymbol{y} \geq 0, \quad \text{and} \quad x_j \prod_{i=1}^{p_j} y_i^j = 0$$

$$(j = 1, \ldots, k).$$

Several authors have further investigated this vertical generalization while others have studied some analogous horizontal generalizations. For representative papers on the vertical LCP, see Ebiefung (1995) and Mohan and Neogy (1997). For the horizontal generalization, Tütüncü and Todd (1995) and Zhang (1994). A further generalization called extended linear complementarity problem (ELCP) was introduced by Mangasarian and Pang (1995) and subsequently developed in Gowda (1995, 1996) and Sznajder and Gowda (1995). Also called the extended linear complementarity problem is another variant expounded by De Schutter and De Moor (1996) that captures the previously mentioned HLCP, VLCP and ELCP.

The Nonlinear Complementarity Problem

The NCP($\boldsymbol{f}$) as an identified problem first appeared in Cottle (1964, 1966). There—under very strong assumptions on f—an existence theorem and an

analogue of the principal pivoting method for the LCP were presented. As described in Pang (1995), contemporary iterative NCP algorithms tend to fall into three categories: (i) the basic Newton method, (ii) nonsmooth-equations approaches, and (iii) interior-point methods. Some algorithms are inspired by the equivalence between the NCP(f) and the variational inequality problem $\mathrm{VI}(X, f)$ in the case where $X = R^n_+$. Some seek zeros of a function such as g defined in (3) whereas others attack the nonlinear program (4) or a variant thereof. Despite the existence of several fine collections of research articles on nonlinear complementarity problems, the authoritative surveys of Harker and Pang (1990) and Pang (1995) came as close as anything then available to a monograph on this topic. The field now benefits from the publication of the masterful two-volume work on variational inequalities and complementarity problems by Facchinei and Pang (2003).

Software for Complementarity Problems

Information about available software for (mixed) complementarity problems can be found by searching the World Wide Web.

See

- ▶ Complementarity Applications
- ▶ Complementary Slackness Theorem
- ▶ Game Theory
- ▶ Matrices and Matrix Algebra
- ▶ Nonlinear Programming
- ▶ Quadratic Programming

References

Cottle, R. W. (1964). *Nonlinear programs with positively bounded Jacobians*, Ph.D. Thesis, Department of Mathematics, University of California, Berkeley. (See also, Technical Report ORC 64–12 (RR), Operations Research Center, University of California, Berkeley.)

Cottle, R. W. (1966). Nonlinear programs with positively bounded Jacobians. *SIAM Journal on Applied Mathematics, 14*, 147–158.

Cottle, R. W. (2010). A field guide to the matrix classes found in the literature of the linear complementarity problem. *Journal of Global Optimization, 46*, 571–580.

Cottle, R. W., & Dantzig, G. B. (1968). Complementary pivot theory of mathematical programming. *Linear Algebra and Its Applications, 1*(1968), 103–125.

Cottle, R. W., & Dantzig, G. B. (1970). A generalization of the linear complementarity problem. *Journal of Combinatorial Theory, 8*, 79–90.

Cottle, R. W., Pang, J. S., & Stone, R. E. (1992). *The linear complementarity problem*. Boston: Academic Press.

Cottle, R. W., Pang, J. S., & Stone, R. E. (2009). *The linear complementarity problem. Classics in applied mathematics*. Philadelphia: SIAM.

De Schutter, B., & De Moor, B. (1996). The extended linear complementarity problem. *Mathematical Programming, 71*, 289–326.

Eaves, B. C. (1971). The linear complementarity problem. *Management Science, 17*, 612–634.

Eaves, B. C., & Lemke, C. E. (1981). Equivalence of LCP and PLS. *Mathematics of Operations Research, 6*, 475–484.

Ebiefung, A. A. (1995). Existence theory and Q-matrix characterization for the generalized linear complementarity problem. *Linear Algebra and Its Applications, 223*(224), 155–169.

Facchinei, F., & Pang, J. S. (2003). *Finite-dimensional variational inequalities and complementarity problems*. New York: Springer.

Ferris, M. C., & Pang, J. S. (1997). Engineering and economic applications of complementarity problems. *SIAM Review, 39*, 669–713.

Gowda, M. S. (1995). On reducing a monotone horizontal LCP to an LCP. *Applied Mathematics Letters, 8*, 97–100.

Gowda, M. S. (1996). On the extended linear complementarity problem. *Mathematical Programming, 72*, 33–50.

Habetler, G. J., & Price, A. J. (1971). Existence theory for generalized nonlinear complementarity problems. *Journal of Optimization Theory and Applications, 7*, 223–239.

Harker, P. T., & Pang, J. S. (1990). Finite-dimensional variational inequality and nonlinear complementarity problems: A survey of theory, algorithms and applications. *Mathematical Programming, Series B, 48*, 161–220.

Howson, J. T., Jr. (1963). *Orthogonality in linear systems*, Ph.D. Thesis, Department of Mathematics, Rensselaer Institute of Technology, Troy, New York.

Isac, G. (1992). *Complementarity problems, lecture notes in mathematics 1528*. Berlin: Springer-Verlag.

Isac, G. (2000). *Topological methods in complementarity theory*. Dordrecht: Kluwer.

Isac, G., Bulavsky, V. A., & Kalashnikov, V. V. (2002). *Complementarity, equilibrium, efficiency, and economics*. Dordrecht: Kluwer.

Karamardian, S. (1971). Generalized complementarity problem. *Journal of Optimization Theory and Applications, 8*, 161–168.

Kojima, M., et al. (1991). *A unified approach to interior point algorithms for linear complementarity problems*. Berlin: Springer-Verlag. lecture notes in computer science 538.

Lemke, C. E. (1965). Bimatrix equilibrium points and mathematical programming. *Management Science, 11*, 681–689.

Lemke, C. E., & Howson, J. T., Jr. (1964). Equilibrium points of bimatrix games. *SIAM Journal on Applied Mathematics, 12*, 413–423.

Luo, Z. Q., Pang, J. S., & Ralph, D. (1996). *Mathematical programs with equilibrium constraints*. New York: Cambridge University Press.
Mangasarian, O. L., & Pang, J. S. (1995). The extended linear complementarity problem. *SIAM Journal on Matrix Analysis and Applications, 16*, 359–368.
Mangasarian, O. L., & Solodov, M. V. (1993). Nonlinear complementarity as unconstrained and constrained minimization. *Mathematical Programming, Series B, 62*, 277–297.
Mohan, S. R., & Neogy, S. K. (1997). Vertical block hidden Z-matrices and the generalized linear complementarity problem. *SIAM Journal on Matrix Analysis and Applications, 18*, 181–190.
Murty, K. G. (1968). *On the number of solutions to the complementarity problem and spanning properties of complementary cones*. Ph.D. Thesis, Department of Industrial Engineering and Operations Research, University of California, Berkeley.
Murty, K. G. (1972). On the number of solutions to the complementarity problem and spanning properties of complementary cones. *Linear Algebra and Its Applications, 5*, 65–108.
Murty, K. G. (1988). *Linear complementarity, linear and nonlinear programming*. Berlin: Heldermann-Verlag.
Pang, J. S. (1995). Complementarity problems. In R. Horst & P. Pardalos (Eds.), *Handbook of global optimization* (pp. 271–338). Dordrecht: Kluwer.
Robinson, S. M. (1979). Generalized equations and their solutions, part I: Basic theory. *Mathematical Programming Study, 10*, 128–141.
Samelson, H., Thrall, R. M., & Wesler, O. (1958). A partition theorem for Euclidean n-space. *Proceedings of the American Mathematical Society, 9*, 805–807.
Sznajder, R., & Gowda, M. S. (1995). Generalizations of *Po*-and *P*-properties; extended vertical and horizontal LCPs. *Linear Algebra and Its Applications, 223/224*, 695–715.
Tütüncü, R. H., & Todd, M. J. (1995). Reducing horizontal linear complementarity problems. *Linear Algebra and Its Applications, 223*(224), 717–730.
Zhang, Y. (1994). On the convergence of a class of infeasible interior-point algorithm for the horizontal linear complementarity problem. *SIAM Journal on Optimization, 4*, 208–227.

Complementary Pivot Algorithm

▶ Quadratic Programming

Complementary Slackness Theorem

For the symmetric form of the primal and dual problems the following theorem holds: For optimal feasible solutions of the primal and dual (symmetric) systems, whenever inequality occurs in the *k*th relation of either system (the corresponding slack variable is positive), then the *k*th variable of its dual is zero; if the *k*th variable is positive in either system, the *k*th relation of its dual is equality (the corresponding slack variable is zero). Feasible solutions to the primal and dual problems that satisfy the complementary slackness conditions are also optimal solutions. A similar theorem holds for the unsymmetric primal-dual problems: For optimal feasible solutions of the primal and dual (unsymmetric) systems, whenever the *k*th relation of the dual is an inequality, then the *k*th variable of the primal is zero; if the *k*th variable of the primal is positive, then the *k*th relation of the dual is equality. This theorem just states the optimality conditions of the simplex method.

See

▶ Complementarity Applications
▶ Complementarity Condition
▶ Complementarity Problems
▶ Symmetric Primal-Dual Problems
▶ Unsymmetric Primal-Dual Problems

Complex Problem Analyzing Method (Compram)

Dorien J. DeTombe
International Research Society on Methodology of Societal Complexity, Amsterdam, The Netherlands

Complex societal problems are worldwide natural problems caused by viruses such as the flu pandemic, fowl plague, and HIV/AIDS; local natural disasters especially earthquakes, hurricanes, avalanches and floods; technical dangers caused by industry including pollution (CO2), traffic, and nuclear power plants; climate change and agricultural activities; man-made threats such as wars, terrorism, internet vulnerability, stock exchange manipulation, credit crises, and identity theft. The concept of societal complexity and an approach to their resolution, the Complex Problem Analyzing Method (Compram),

are discussed by DeTombe (2001). Compram is based on the idea that societal problems must be handled in a multi-disciplinary and cooperative manner by experts, stakeholders, and policy makers. Compram combines aspects of different methods into a structured interactive approach for policy making to find possible transitions of the situation that can be mutually accepted and implemented (DeTombe 1994).

The related difficult and complicated group processes are guided and structured by a facilitator. Those involved discuss the content and possible solutions based on a cooperative (simulation) model of the problem. The methodology emphasizes facilitating the exchange of knowledge, and understanding and communication between the participants. Compram has been used on a theoretical basis for handling over sixty real-life cases in the field of societal policymaking. The Organisation for Economic Cooperation and Development (OECD) suggests that the analysis of a complex societal problem be supported by the application of Compram (OECD 2006). (Further information on Compram and Societal Complexity can be found on Web sites maintained by the author).

See

- ▶ Community OR
- ▶ Soft Systems Methodology
- ▶ Wicked Problems

References

DeTombe, D. J. (1994). *Defining complex interdisciplinary societal problems*. A theoretical study for constructing a co-operative problem analyzing method: the method Compram. Thesis publishers Amsterdam, ISBN 90 5170 302-3.

DeTombe, D. J. (2001). Compram, a method for handling complex societal problems. *European Journal of Operational Research, 128*(2), 266–281.

OECD. (2006). *Report on global safety*. Report on the workshop on science and technology for a safer society, July 2006, Paris.

Compram

- ▶ Complex Problem Analyzing Method (Compram)

Computational Biology

Harvey J. Greenberg[1] and Allen G. Holder[2]
[1]University of Colorado-Denver, Denver, CO, USA
[2]Rose-Hulman Institute of Technology, Terre Haute, IN, USA

Computational biology is an interdisciplinary field that applies the techniques of computer science, applied mathematics, and statistics to address biological questions. OR is also interdisciplinary and applies the same mathematical and computational sciences, but to decision-making problems. Both focus on developing mathematical models and designing algorithms to solve them. Models in computational biology vary in their biological domain and can range from the interactions of genes and proteins to the relationships among organisms and species.

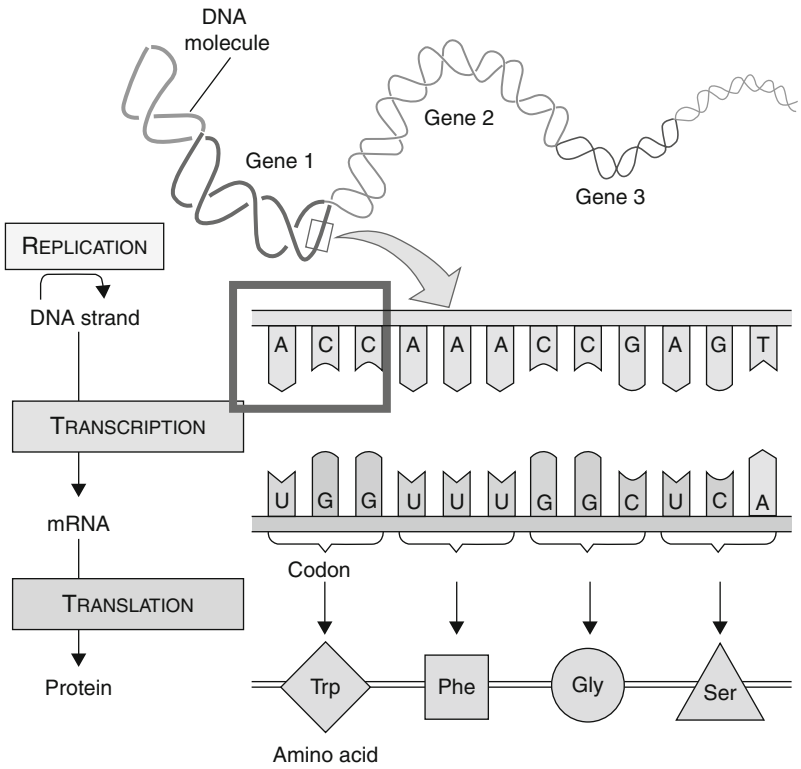
Genes are stretches of deoxyribonucleic acid (DNA), which is sometimes called the user manual for life and is a double-stranded helix of nucleic acids bonded by base-pairs of complements (`a-t`, `c-g`). The central dogma of molecular biology asserts that information in a cell flows from DNA to ribonucleic acid (RNA) to protein (note, Francis Crick used dogma when he introduced this in 1958 to mean without foundation because there was no experimental evidence at that time). Proteins are the workers of the cell, and there is much focus on recognizing, predicting, and comparing their properties (Fig. 1).

Proteins interact either directly by modifying each other's properties through direct contact or indirectly by participating in the production and modification of cellular metabolites. Collectively, the biochemical reactions and the possible intermediates that produce a metabolite comprise a metabolic pathway, and a metabolic network is a collection of these pathways. The study of complex networks like that of the metabolism is called systems biology.

Linear Programming: A linear program (LP) is an optimization problem in which the variables are in $\mathbb{R}^N$, and the constraints and the objective are linear.

Flux Balance Analysis (FBA) – A biochemical process is defined by n reactions that convert m compounds:

$$a_{1j}x_1 + \cdots + a_{mj}x_m \underset{k_-^j}{\overset{k_+^j}{\rightleftharpoons}} b_{1j}x_1 + \cdots + b_{mj}x_m,$$



Computational Biology, Fig. 1 Central dogma of molecular biology

where x_i is the concentration of the ith compound, and $k^j_{\pm}$ is the jth reaction rate (for a 2-way reaction the reverse rate need not equal the forward rate). The corresponding ODE is:

$$\frac{\mathrm{d}x_i(t)}{\mathrm{d}t} = \sum_{j=1}^{n} (b_{ij} - a_{ij}) \left(k^j_+ x_1^{a_{1j}} \cdots x_m^{a_{mj}} - k^j_- x_1^{a_{1j}} \cdots x_m^{a_{mj}} \right)$$
$$= \sum_{j=1}^{n} S_{ij} v_j(x),$$

where v is the flux (production or consumption of mass per unit area per unit time), and S_{ij} is defined as a stoichiometric (pronounced `stoy-kee-uh-me'-trik`) coefficient. These coefficients are interpreted as:

$S_{ij} > 0 \Rightarrow$ rate of compound i produced in reaction j;
$S_{ij} < 0 \Rightarrow$ rate of compound i consumed in reaction j.

The following holds asymptotically provided that the system approaches a steady state toward equilibrium concentrations $\bar{x}$:

$$\lim_{t \to \infty} \frac{\mathrm{d}x(t)}{\mathrm{d}t} = Sv(\bar{x}) = 0. \tag{1}$$

Dropping the dependence of the flux on $\bar{x}$, the flux cone is defined by this homogeneous system plus non-negativity for one-way reactions, indexed by J:

$$\mathcal{F} = \{v : Sv = 0, \;\; v_J \geq 0\}. \tag{2}$$

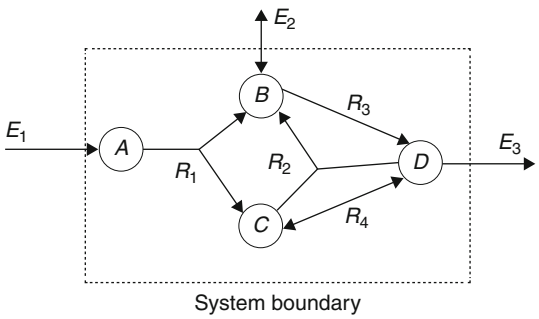
In a metabolic network, reactions are distinguished between external and internal. The flux associated with an external reaction is an exchange between the network of interest and the cell's environment (Fig. 2).

The stoichiometric matrix for the internal reactions is extended to include external reactions, each being a singleton column with ± 1:

$$S = \begin{array}{c} \\ A \\ B \\ C \\ D \end{array} \begin{array}{c} \begin{array}{cccc|ccc} R_1 & R_2 & R_3 & R_4 & E_1 & E_2 & E_3 \end{array} \\ \left[\begin{array}{cccc|ccc} -1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & -2 & 0 & 0 & -1 & 0 \\ 2 & -1 & 0 & -2 & 0 & 0 & 0 \\ 0 & -2 & 1 & 3 & 0 & 0 & -1 \end{array} \right] \end{array}$$

Computational Biology, Fig. 2 Example metabolic network with four internal and three external reactions

$R_1 : A \rightharpoonup B + 2C$ (multiple output)
$R_2 : C + 2D \rightharpoonup B$ (multiple input)
$R_3 : 2B \rightharpoonup D$ (simple)
$R_4 : 2C \leftrightharpoons 3D$ (simple, 2-way)
$E_1 : \rightharpoonup A$ (supply)
$E_2 : B \leftrightharpoons$ (2-way exchange)
$E_3 : D \rightharpoonup$ (demand)



All reactions are one-way reactions, except R_4 and E_2, so $J = \{1, 2, 3, 5, 7\}$, leaving v_4 and v_6 without sign restriction in the flux cone.

Strictly speaking, a metabolic network is usually not a network in the OR sense because some internal reactions have multiple inputs or outputs (sometimes called a process network in chemical engineering). Hence, LP is used, rather than specialized network algorithms, to find fluxes. The FBA LP model has the form:

$$\max \quad c^T v : \quad v \in \mathcal{F} \cap \mathcal{B}, \tag{3}$$

where $\mathcal{B}$ is a bounding set so that the linear program has an optimal solution if it is feasible. A common objective is to maximize the rate of growth defined in terms of metabolites, where the objective coefficients (c) depend on the organism. Other objectives include maximizing some metabolite production, minimizing by-product production, minimizing substrate requirements, and minimizing mass nutrient uptake (Palsson 2006).

An optimal basis depends on the definition of $\mathcal{B}$. Three possibilities, which may be combined, are:

$$\begin{aligned} &\text{simple bound :} && L_K \le v_K \le U_K \\ &\text{fixing inputs and/or outputs :} && v_K = \bar{v}_K \\ &\text{normalization :} && \sum_{j \in K} v_j = b, \end{aligned}$$

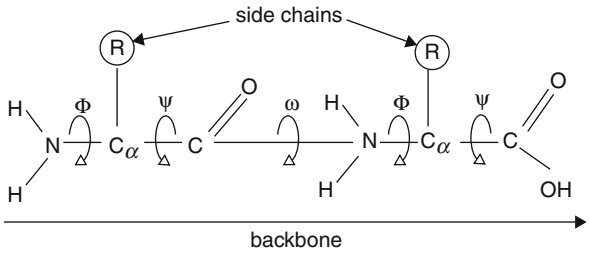
where K is a subset of reactions. Inputs and outputs are generally a subset of the exchanges. Normalization applies to one-way reactions – i.e., $K \subseteq J$. Each extreme ray of the flux cone corresponds to an extreme point of the polytope. The converse is generally not true – viz., fixing the flux of a reaction that transports metabolites in or out of the cell can introduce extreme points with no extreme ray of the flux cone passing through them.

Pathways are subnetworks with a single biological effect. In an ordinary network, where each internal reaction has a single input and output, this is a path. A cut set is defined as a set of reactions whose removal renders the stoichiometric (1) infeasible for a specified output. For an ordinary network, the OR terminology is a disconnecting set. A minimal cut set for a specified output is, in OR terminology, simply a cut set. For the example, a cut set that separates D from the rest of the network is $\{R_1, R_3, R_4, E_1\}$. Finding a (minimal) cut set in the general case becomes an IP, using binary variables to block pathways to some specified output.

Nonlinear Programming: A nonlinear program (NLP) is defined by having the objective or some constraint function be nonlinear in the decision variables.

Protein Folding – Most proteins go through a process that twists and turns the molecules from their primary state of a linear progression of amino acids to a native three-dimensional state in which it remains. That process is called folding, and it is theoretically possible to predict a protein's native state, or structure, by knowing its primary state. This determines a protein's function, and some diseases (e.g., Alzheimer's, Huntington's, and cystic fibrosis) are associated with protein misfolding.

Predictive models became possible following the work of Christian B. Anfinsen, who in 1961 published experimental results supporting the Thermodynamic Hypothesis: A protein's native state is uniquely determined by its primary sequence; it transitions to a state of minimum free energy. This leads to a nonlinear program with the decision space defined as the spatial coordinates of atoms, constrained by the biochemistry of a protein's defining amino acid sequence. The objective function is a free energy determined by potential energies from atomic bonds and non-bond interactions.



Computational Biology, Fig. 3 Covalent bonds along the backbone result in a residue for each of the amino acids. The torsion angles are denoted by Ψ and Φ; ω is the dihedral angle

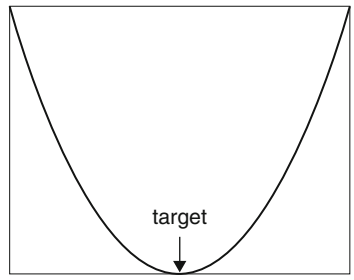
The bonds for the sequence of amino acids shown in Fig. 3 are covalent, meaning that they share electrons, and these strong bonds hold the backbone together. Objective terms for the ith covalent bond include the energies required to stretch, bend, and twist the bond.

Action	Energy	
Stretching	E^{stretch}	$= \sum_i K_i^L \left(L_i - L_i^0\right)^2$
Bending	E^{bend}	$= \sum_i K_i^\theta \left(\theta_i - \theta_i^0\right)^2$
Twisting	E^{twist}	$= \sum_i K_i^\phi (1 - \cos(\omega_i))$

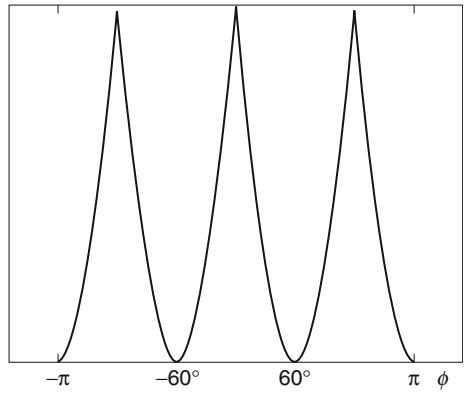
The variables are the bond length (L) and the bond angles, ω and $\theta = (\Psi, \Phi)$, which are determined by atomic coordinates. Parameters include target values (L^0, θ^0). Weight parameters (K) are scale factors that put the energy terms in the same unit; those values can be measured or derived. For example, if it requires 100 kcal/mole to break a bond, and two positive charges within 3.3 Å (Angstrom) have at least 100 kcal/mole, then the total energy is reduced by breaking a bond to keep positive charges distant. Estimating these values to determine weight parameters is not an exact science, so even these basic energy functions are inexact, and there are other energy functions for non-covalent bonds and among non-bonding atoms.

Two common energy functions estimate the electrostatic and Van der Waals interactions:

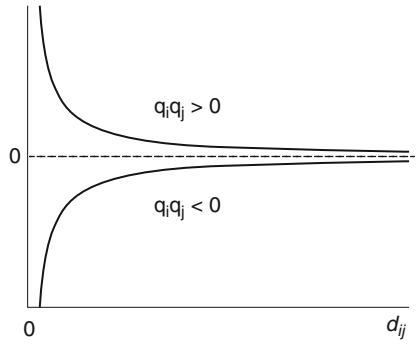
Action	Energy	
Electrostatic	E^{elec}	$= \sum_{i<j} K_{ij}^{\text{elec}} \frac{q_i q_j}{d_{ij}}$
Van der Waals	E^{vdw}	$= \sum_{i<j} K_{ij}^{\text{vdw}} \left(\left(\frac{d_{ij}^*}{d_{ij}}\right)^{12} - \alpha_{ij} \left(\frac{d_{ij}^*}{d_{ij}}\right)^{6} \right)$



Computational Biology, Fig. 4 The squared deviation of E^{stretch} and E^{bend} is convex



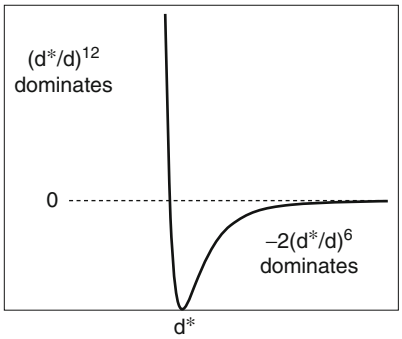
Computational Biology, Fig. 5 E^{twist} with $\omega = 3/2(\phi - \pi)$



Computational Biology, Fig. 6 E^{elec} depends on the sign of $q_i q_j$. Oppositely signed atoms attract, so the energy is negative and favors them being close

The variables are the pairwise distances (d), which are determined by the atomic coordinates. Parameters are the atomic charges (q) and equilibrium distances (d^*) (Figs. 4–7).

The NLP approach (Floudas and Pardalos 2000) uses energy principles that underlie molecular



Computational Biology, Fig. 7 Lennard-Jones approximation of E^{vdw} for $\alpha = 2$

dynamics, and these methods attempt to find the native state and a pathway to it. In practice, not all parameters are grounded in some physical law. An energy function could include contributions from non-bonded and uncharged pairs, based on their distance and radii. Alternatively, known structures can be used to predict an unknown structure, based on their evolutionary similarity. This is called homology, and it is focused on determining the native state and not on discerning the dynamic pathways to reach it.

The multi-modal shape of the energy landscape leads to the Levinthal Paradox: Many proteins reach their native state within milliseconds, yet the number of stable conformations grows exponentially in the number of amino acids. One explanation is that proteins fold into a nearby local minimum of the free energy instead of the global minimum. Global optimization methods based on this principle are called funneling methods. Another explanation is that the dimension of the problem is not the length of the amino acid sequence but is instead the number of chains that obey patterns not fully understood. Combinatorial optimization methods based on this principle are called chain growth and zipping and assembly algorithms.

Comparing Protein Function – A protein's function is determined by its 3D native state, of which many confirmations are known. Comparing protein structures relates protein function and collects proteins into functionally similar families that help identify a protein's functions.

Proteins typically have multiple functional domains, each of which would act as an independent protein if its amino acid subsequence had folded independently. Two proteins are considered to be functionally similar if they share a (nearly) common domain. Each domain is composed of secondary structures, notably α-helices and β-sheets, illustrated in Fig. 8. In structure alignment the goal is to best align the secondary structures between two proteins' domains. The input to the alignment problem is a set of coordinates for the C_α atoms for each domain – i.e., the spatial coordinates for the carbon atoms linked to the side chains (c.f., Fig. 3).

To remove a dependency on rigid body motion, structures are often aligned with respect to pairwise distances, d_{ij}, which is a measure between the ith and jth C_α atoms. Let d'_{ij} and d''_{kr} be the intra-distance measures for the two domains, and consider the binary variable:

$$x_{ik} = \begin{cases} 1 & \text{if the } i^{th}\ C_\alpha \text{ atom of the first domain is paired with} \\ & \text{the } k^{th}\ C_\alpha \text{ atom of the second domain;} \\ 0 & \text{otherwise.} \end{cases}$$

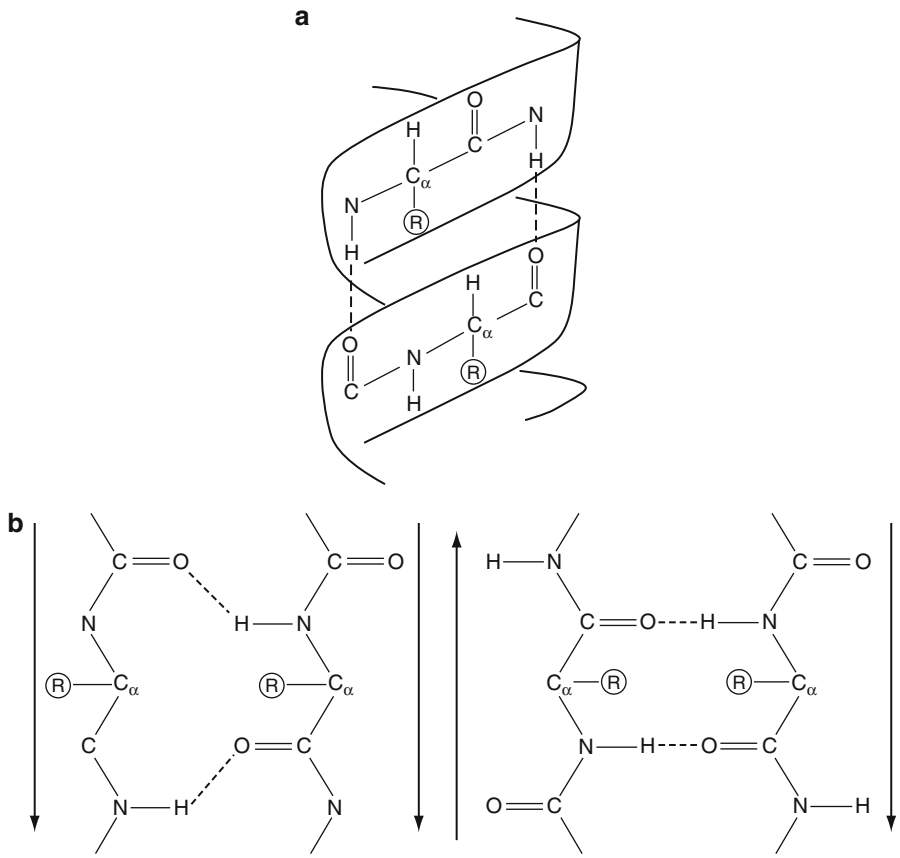
An optimal pairing between the two domains can be calculated by solving a quadratic integer program:

$$\max \sum_{i,k,j,r} x_{ik} x_{jr} d'_{ij} d'_{kr} : \sum_k x_{ik} \le 1,$$
$$\sum_i x_{ik} \le 1,\ x_{ik} = 0,\ (i,k) \in \mathcal{S},$$

where $(i, k) \in \mathcal{S}$ if the ith and kth C_α atoms are in different types of secondary structures.

Besides the choice of metric, a variation is to allow pairings between C_α atoms whose secondary structures are different. This is accommodated by removing the restriction that $x_{ik} = 0$ for $(i, k) \in \mathcal{S}$ and adding penalty terms in the objective: $-\sum_{(i,k)\in\mathcal{S}} p_{ik} x_{ik}$. The problem as stated includes the possibility of a non-sequential alignment, i.e., one in which the C_α atoms can be paired independent of the amino acid sequence. A combinatorial optimization model of alignments that requires the same ordering of the amino acid residues is called contact map optimization (Burkowski 2009; Glodzik and Skolnick 1994; Goldman et al. 1999).

Integer Programming: An integer program (IP) is an optimization problem in which some or all of the variables are restricted to be integer valued. For combinatorial optimization, the integer values are simply $\{0, 1\}$.



Computational Biology, Fig. 8 Secondary structures formed along the backbone define a protein's shape. *Dotted lines* represent hydrogen bonds; ® represents a side chain. (**a**) α-Helix, most closely packed arrangement of residues, defined by three parameters: pitch, rise, and turn. (**b**) *β*-Sheets form if the backbone is loosely packed, almost fully extended; they can be parallel (*left*), antiparallel (*right*), or a mixture

Pathway Analysis – Consider the FBA model (3) with added binary variables associated with each process with finite bounds (given or derived), $L_j \le v_j \le U_j$:

$$y_j = \begin{cases} 1 & \text{if } v_j \neq 0; \\ 0 & \text{otherwise.} \end{cases}$$

Replacing the bound constraints with $L_j y_j \le v_j \le U_j y_j$ forces $v_j = 0$ if $y_j = 0$. This corresponds to excluding reaction j, which is called a knock-out. Drug side effects are caused by unintended knock-outs, which, if cannot be avoided, can at least be identified and minimized. In drug design, one may want to block all pathways to some final output. If P is a pathway leading to the targeted output, then adding the constraint

$$\sum_{j \in P} y_j \le |P| - 1$$

removes the pathway, where $j \in P$ if pathway P contains reaction j.

A cut set can be computed with successive pathway-generation for a specified output and adding its pathway-elimination constraint. For the example in Fig. 2, pathways to produce D can be generated by fixing $v_7 = 1$ (and not have y_7). The first basic optimal solution uses reactions R_1, R_3, R_4, E_1, E_3. This leads to the addition of the constraint:

$$y_1 + y_3 + y_4 + y_5 \le 3.$$

The next pathway generated is R_3, E_1, and $y_3 = 0$ satisfies both pathway constraints. After eliminating R_3, the solution is R_1, R_4, E_1, E_3.

Other logical constraints include process conflict, $y_j + y_{j'} \le 1$ (i.e., inclusion of j requires exclusion of j'), and process dependence, $y_j \ge y_{j'}$ (i.e., exclusion of j requires exclusion of j'), for $j \neq j'$.

Rotamer Assignment – Part of the protein folding problem is knowing the side-chain conformations – i.e., knowing the torsion angles of the bonds (c.f., Fig. 3). The rotation about a bond is called a rotamer, and there are libraries that give configuration likelihoods, for each amino acid (from which energy values can be derived). The rotamer

assignment (RoA) problem is to find an assignment of rotamers to sites that minimizes the total energy of the molecule. For the protein folding problem, the amino acid at each site is known. There are about 10–50 rotamers per amino acid, depending on what else is known (such as knowing that the amino acid is located in a helix), so there are about 10^n to 50^n rotamer assignments for a protein of length n.

Let r be in the set of rotamers that can be assigned to site i, denoted by $\mathcal{R}_i$, and let

$$x_{ir} = \begin{cases} 1 & \text{if rotamer } r \text{ is assigned to site } i; \\ 0 & \text{otherwise.} \end{cases}$$

An optimal assignment is a solution to the quadratic semi-assignment problem:

$$\min \sum_i \sum_{r \in R_i} \left(\mathcal{E}_{ir} x_{ir} + \sum_{j>i} \sum_{t \in R_j} E_{irjt} x_{ir} x_{jt} \right):$$
$$\sum\nolimits_{r \in R_j} x_{ir} = 1 \, \forall i, \quad x \in \{0, 1\}.$$

The objective function includes two types of energy: (1) within a site, E_{ir}, and (2) between rotamers of two different sites, $\mathcal{E}_{irjt}$ for $i \neq j$. The summation condition $j > i$ avoids double counting, where $E_{irjt} = E_{jtir}$.

Besides its role in determining a protein's structure, the RoA problem is useful in drug design. Specifically, the RoA problem can be used to determine a minimum-energy docking site for a ligand, which is a small molecule such as a hormone or neurotransmitter that binds to a protein and modifies its function. The ligand-protein docking problem is characterized by only a few sites, and if the protein is known, the dimensions are small enough that the RoA problem can be solved exactly. However, if the protein is to be engineered, then there can be about 500 rotamers per site (20 acids @ 25 rotamers each), in which case solutions are computed with metaheuristics or approximation algorithms. There are other bioengineering problems associated with the RoA problem, such as determining protein-protein interactions. While the mathematical structure is the same, the applications have different energy data, which can affect algorithm performance (Forrester and Greenberg 2008).

Also see Clote and Backofen (2000), Jones and Pevzner (2004), and Lancia (2006).

Dynamic Programming: This is a computational approach to sequential decision making. Two fundamental biological sequences are taken from the alphabet of nucleic acids, {a, c, g, t}, and from the alphabet of amino acids, {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V}. The former is a segment of DNA (or RNA if u replaces t – i.e., uracil instead of thymine); the latter is a protein segment.

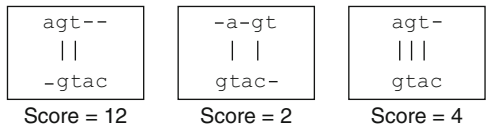
Sequence Alignment – Two sequences can be optimally aligned by dynamic programming, where optimal is one that maximizes an objective that has two parts:

1. A *scoring function*, given in the form of an $m \times m$ matrix S, where m is the size of the alphabet. The value of S_{ij} measures a propensity for the i^{th} alphabet-character in one sequence to align with the j^{th} alphabet-character in some position of the other sequence.
   Example: Let $s =$ agt and $t =$ gtac. If the first character of s is aligned with the first character of t, then the score is S_{ag}, which is the propensity for a to be aligned with g.
2. A *gap penalty function*, expressed in two parts: a fixed cost of beginning a gap, denoted G_{open}, and a cost to extend the gap, denoted G_{ext}.
   Example: Let $s =$ agt and $t =$ gtac. One alignment is $\begin{matrix} \mathtt{agt-} \\ \mathtt{gtac} \end{matrix}$, which puts a gap at the end of the first sequence.

A gap is called an indel because it can be either an insertion into one sequence or a deletion from the other sequence: insert $\downarrow$ $\begin{bmatrix} - \\ \mathtt{a} \end{bmatrix}$ delete $\uparrow$ If one sequence evolved directly from the other, the evolutionary operation is determined by their time-order. If they have a common ancestor, they evolved along different paths, resulting in the indel when comparing them. The evolutionary biology explains why sequences can be more similar than a simple alignment (without gaps) may suggest.

Figure 9 shows three different alignments for the two nucleic acid sequences, agt and gtac. Scores are shown for the following scoring matrix and do not account for gapping:

$$S = \begin{matrix} \mathtt{a} & \mathtt{c} & \mathtt{g} & \mathtt{t} \end{matrix}$$
$$\begin{bmatrix} 6 & 1 & 2 & 1 \\ 1 & 6 & 1 & 2 \\ 2 & 1 & 6 & 1 \\ 1 & 2 & 1 & 6 \end{bmatrix} \begin{matrix} \mathtt{a} \\ \mathtt{c} \\ \mathtt{g} \\ \mathtt{t} \end{matrix}$$



Computational Biology, Fig. 9 Three alignments for two sequences

If the objective is a linear affine function of gap lengths, the total objective function for the 2-sequence alignment problem is:

$$\sum_{i,j} S_{s_i t_j} - G_{\text{open}}(N_s + N_t) - G_{\text{ext}}(M_s + M_t),$$

where the sum is over aligned characters, s_i from sequence s with t_j from sequence t. The number of gaps opened is N_s in sequence s and N_t in sequence t; the number of gap characters (–) is M_s in sequence s and M_t in sequence t. In the examples of Fig. 9, if $G_{\text{open}} = 2$ and $G_{\text{ext}} = 1$, the gap penalties are 7, 9, and 3, respectively.

The alphabet is extended to include the gap character, with S extended to include gap extension, as $S_{a-} = S_{-a} = G_{\text{ext}}$ for all a in the alphabet. (So, G_{ext} includes the penalty for the first alignment with -.) Let s^i denote the subsequence $(s_1, \ldots, s_i)$, with $s^0 = \varnothing$. Here is the DP recursion for $G_{\text{open}} = 0$:

$$F(s^i, t^j) = \max \begin{cases} F(s^{i-1}, t^{j-1}) + S_{s_i t_j} & \text{match;} \\ F(s^{i-1}, t^j) + S_{s_i -} & \text{insert} - \text{into } t; \\ F(s^i, t^{j-1}) + S_{-t_j} & \text{insert} - \text{into } s. \end{cases} \tag{4}$$

The initial conditions are:

$$\begin{aligned} &F(\emptyset, \emptyset) = 0; \\ &F(s^i, \emptyset) = F(s^{i-1}, \emptyset) + S_{s_i -}, \quad i = 1, \cdots |s|; \\ &F(\emptyset, t^j) = F(\emptyset, t^{j-1}) + S_{-t_j}, \quad j = 1, \cdots |t|. \end{aligned}$$

The DP recursion (4) is for global alignment, and it has been extended to allow $G_{\text{open}} > 0$ and to not penalize leading or trailing gaps (allowing a short sequence to be aligned with a large one meaningfully). Local alignment is finding maximal substrings (contiguous subsequences) with an optimal global alignment having maximum score (Gusfield 1997; Waterman 1995).

Sequences from many species can be compared simultaneously in a Multiple Sequence Alignment (MSA). One way to evaluate an MSA is by summing pairwise scores. Figure 10 shows an example. The sum-of-pairs score, based on the scoring matrix S, is shown for each column. For example, column 1 has $3S_{\texttt{aa}} + 3S_{\texttt{ac}} = 3$. The sum-of-pairwise scores for column 2 is zero because gap scores are not shown by columns; they are penalized for each sequence (rows of alignment) with $G_{\text{open}} = 2$ and $G_{\text{ext}} = 1$. The total objective value is $152 - 37 = 115$.

MSA is a computational challenge to exact DP due to the combinatorial explosion of the state space, but one could use approximate DP or formulate MSA as an IP.

Phylogenetic Tree Construction – Phylogeny is the evolutionary history of some biological entity. A phylogenetic tree (PT) is a graphical presentation of a phylogeny. A leaf represents an Operational Taxonomic Unit (OTU), which can be various levels – e.g., species, genes, pathways, and enzymes. Each edge, or branch, is a relation between a pair of OTUs. Each internal node is constructed so that the resulting PT is consistent with the OTU data, and the root represents a common ancestor of the OTUs.

Example. Consider five OTUs and an MSA of DNA sites with six base-pairs (Fig. 11):

	site					
OTU	1	2	3	4	5	6
A	c	a	g	a	c	a
B	c	a	g	g	t	a
C	c	g	g	g	t	a
D	t	g	c	g	t	a
E	t	g	c	a	c	t

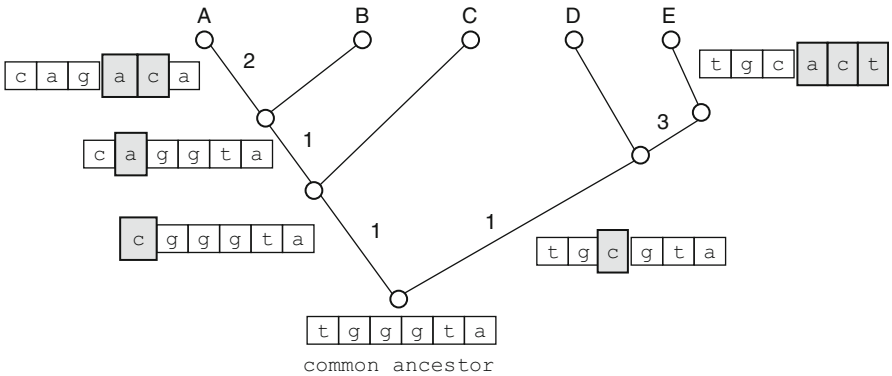
If the number of mutations is the distance between two sequences, then the distance between OTUs is the length of the unique path between them in the PT. The example has the distance matrix:

$$D = \begin{array}{c} \begin{matrix} A & B & C & D & E \end{matrix} \\ \begin{bmatrix} 0 & & & & \\ 2 & 0 & & & \\ 3 & 1 & 0 & & \\ 5 & 3 & 2 & 0 & \\ 8 & 6 & 5 & 3 & 0 \end{bmatrix} \end{array} \begin{matrix} A \\ B \\ C \\ D \\ E \end{matrix}$$

Computational Biology, Fig. 10 A multiple alignment of four sequences

													Gap penalty
	a	-	g	a	g	t	-	a	c	t	-	-	-	11
	a	a	g	t	a	t	-	-	a	t	-	-	-	9
	a	-	-	t	a	t	a	a	-	-	-	-	t	10
	c	-	g	t	a	-	-	a	c	t	c	c	t	7
score:	21	0	18	21	24	18	0	18	8	18	0	0	6	37

Total = 152



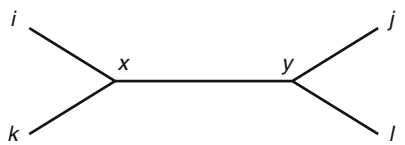
Computational Biology, Fig. 11 The example maximum-parsimony PT has eight mutations, shown on the branches. (All other PTs have more than 8.)

This is not the same as the MSA distance. For example, $D(A, E) = 8$ in the PT but is only 4 in the MSA.

Regardless of how the distance matrix is derived (MSA or not), there may not exist a PT that satisfies specified distances. For that to be true, it is necessary and sufficient that the metric be additive – i.e., for any four leaves, there exist labels i, j, k, ℓ such that

$$D(i,j) + D(k,\ell) = D(i,\ell) + D(j,k) \geq D(i,k) + D(j,\ell).$$

The reason for this is that there must be some splitting i, k from j, ℓ with an internal branch:



Additivity does not usually hold, so the problem is to construct a PT whose associated leaf-distance matrix, D, minimizes some function of nearness to the given D^0, such as $\|D - D^0\|$. This problem is NP-hard. Heuristics include sequential clustering: Un-weighted/Weighted Pair Group Method with Arithmetic Mean (UPGMA/WPGMA) and neighbor-joining algorithms.

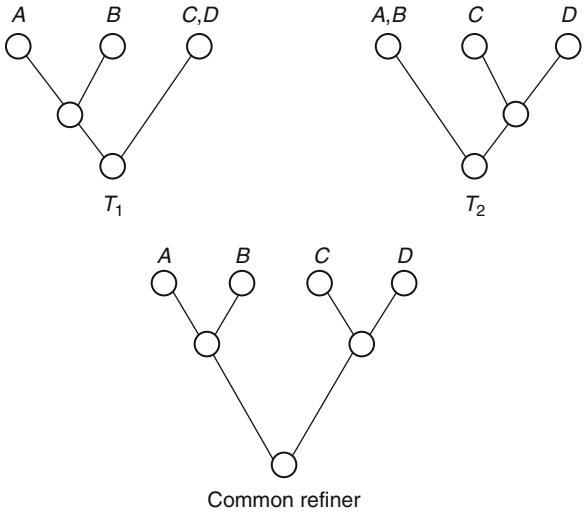
There may be multiple PTs, which generally come from different data – e.g., one from an MSA of a DNA segment, another from the maximum likelihood of some property. If a series of edge-contractions is applied to a PT, the resulting PT is called a refinement and the original is called a refiner. Two trees are compatible if they have a common refiner. One problem is to determine whether two PTs are compatible, and if so, what is their common refiner? If incompatible, how is a PT constructed that has some agreement with the given PTs?

A Matrix Representation with Parsimony (MRP) of a PT with k internal nodes is a binary matrix defined as:

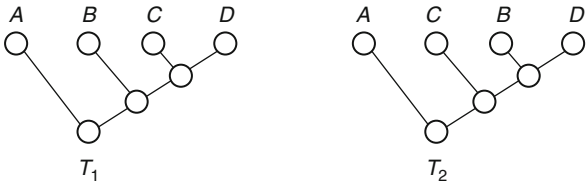
$$M_{ij} = \begin{cases} 1 & \text{if internal node } j \text{ is in the (unique) path} \\ & \text{from the root to OTU } i; \\ 0 & \text{otherwise.} \end{cases}$$

Conversely, given a binary matrix, if it has an associated PT, it is called a perfect phylogeny.

Given two PTs for the same OTUs with MRPs, M^1, M^2, their column-union is $[M^1\ M^2]$.



Computational Biology, Fig. 12 PTs T_1, T_2 are compatible



Computational Biology, Fig. 13 PTs T_1, T_2 are incompatible

Theorem. Two PTs are compatible if, and only if, their MRP column-union represents a perfect phylogeny.

The trees in Fig. 12 have the MRP column-union:

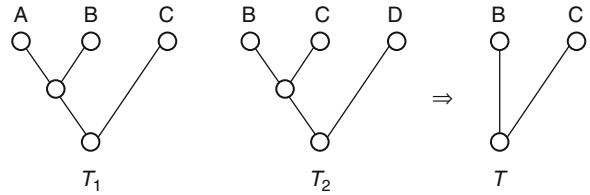
$$M = \begin{array}{c} \begin{matrix} M^1 & M^2 \end{matrix} \\ \left[\begin{array}{c|c} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{array}\right] \end{array} \begin{matrix} A \\ B \\ C \\ D \end{matrix}$$

This is the MRP of the common refiner in Fig. 12 and represents a perfect phylogeny.

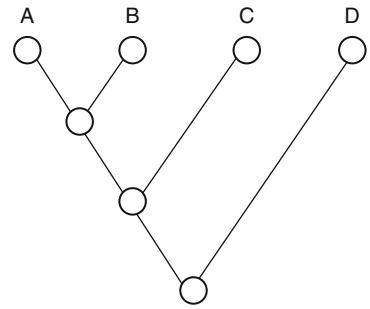
The MRP column-union of the PTs in Fig. 13 is:

$$M = \begin{array}{c} \begin{matrix} M^1 & \quad M^2 \end{matrix} \\ \left[\begin{array}{cc|cc} 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{array}\right] \end{array} \begin{matrix} A \\ B \\ C \\ D \end{matrix}$$

M does not correspond to any PT. (After drawing A, C, D with four internal nodes as the path to D, OTU B cannot be drawn with the path 0-1-3-4 without introducing the cycle, 1-2-3-1.)



Computational Biology, Fig. 14 A maximum agreement subtree with 2 of the 4 OTUs



Computational Biology, Fig. 15 An agreement supertree of the trees in Fig. 14

Suppose the trees are incompatible. A Maximum Agreement Subtree (MAST) is a refined subtree with the greatest number of leaves (Fig. 14).

The DP recursion for two subtrees (Steel and Warnow 1993) is nontrivial. The state is a pair of subtrees with specified roots, $\left(T_1^r, T_2^s\right)$. Each tree has an inclusion-ordered sequence of such subtrees, which is computed during the recursion. The decision space to compute $MAST\left(T_1^r, T_2^s\right)$, given $MAST\left(T_1^{r'}, T_2^{s'}\right)$ for $\left(T_1^{r'}, T_2^{s'}\right) \prec \left(T_1^r, T_2^s\right)$, requires the computation of a maximum weighted-matching on the complete r-s bipartite graph, weighted with $\{MAST(r', s')\}$.

Whereas MAST uses an intersection of PT information, a supertree uses their union. Construction methods vary, and some of the criteria address common order preservation. An agreement supertree, T, is a minimal tree such that each T_i is a refined subtree of T (Fig. 15).

Markov Chains and Processes: A stochastic process has the Markov property if the transition from one state to the next depends on only the current state. Classical models include the evolution of some biological states over time (Allen 2003; Wilkinson 2006). Molecular applications of Markov models also

consider ordered sequences of nucleotides (viz., DNA and RNA) and amino acids (viz., proteins).

CpG Island Recognition – In the human genome the appearance of the dinucleotide CG is rare because it causes the cytosine (C) to be chemically modified by methylation, which causes it to mutate into thymine (T). Methylation is suppressed around the promoters, or start regions, of many genes, and there are more CG dinucleotides than elsewhere. Such regions are called CpG islands, and they are typically a few hundred bases long. (CpG is used instead of CG to avoid confusion with a C–G base pair; the p is silent.) The recognition problem is: Given a short segment of a genomic sequence, decide if it is part of a CpG island.

Two Markov chains are defined: P^+ is the state-transition matrix within a CpG island; P^- is the state-transition matrix outside a CpG island. Each is applied to the given sequence and the log-odds ratio determines which is more likely.

Example. Consider a first-order Markov chain model with transition matrices determined by the frequencies in a database having more than 60,000 human DNA sequences:

$$P^+ = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{bmatrix} 0.18 & 0.27 & 0.43 & 0.12 \\ 0.17 & 0.37 & 0.27 & 0.19 \\ 0.16 & 0.34 & 0.38 & 0.12 \\ 0.08 & 0.36 & 0.38 & 0.18 \end{bmatrix} \end{array}$$

$$P^- = \begin{array}{c} \\ \mathtt{A} \\ \mathtt{C} \\ \mathtt{G} \\ \mathtt{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \end{array} \\ \begin{bmatrix} 0.30 & 0.20 & 0.29 & 0.21 \\ 0.32 & 0.30 & 0.08 & 0.30 \\ 0.25 & 0.25 & 0.30 & 0.20 \\ 0.18 & 0.24 & 0.29 & 0.29 \end{bmatrix} \end{array}$$

Given the sequence AACTTCG, its total log-odds ratio is

$$\sum_{i=1}^{6} \log_2\left(P^+_{s_i s_{i+1}} / P^-_{s_i s_{i+1}}\right) = -0.737 + 0.433 - 0.659 - 0.688 + 0.585 + 1.755 = 0.6888.$$

The conclusion is that the DNA segment is in a CpG island.

There is enough data to support the use of the more accurate 5th-order Markov chain, whose six-tuples correspond to two coding regions. At least 4^5 six-tuples are required in the database to estimate the conditional probabilities, $\Pr(x_6|x_1x_2x_3x_4x_5)$, which directly yield the state-transition probabilities:

$$\Pr(y_1y_2y_3y_4y_5|x_1x_2x_3x_4x_5) = \begin{cases} \Pr(x_6|x_1x_2x_3x_4x_5) & \text{if } y = (x_2x_3x_4x_5x_6); \\ 0 & \text{otherwise.} \end{cases}$$

For the particular example, there are only two state transitions, and the same database gives the transition probabilities:

$$P^+(\mathtt{C}|\mathtt{AACTT}) = 0.4 \qquad P^-(\mathtt{C}|\mathtt{AACTT}) = 0.2$$
$$P^+(\mathtt{G}|\mathtt{ACTTC}) = 0.1 \qquad P^-(\mathtt{G}|\mathtt{ACTTC}) = 0.3$$

In this case, the more accurate 5th-order chain yields the log-odds ratio $\log_2 0.4/0.2 + \log_2 0.1/0.3 = -0.585$, and the conclusion is that the DNA segment is not in a CpG island.

A host of related problems use the same Markov model. For example, transcription splices the DNA into coding regions, called exons, removing the remainder, called introns (misnamed junk DNA). A structure recognition problem is to identify exons versus introns.

Many of the structure recognition, comparison, and prediction problems have hidden states, but emissions are observed according to a known probability. These are Hidden Markov Models (HMMs) and are central in modern biology (Durbin et al. 1998).

Queueing Theory: A queue in a system is any set of objects awaiting service, and service is some process (es) involving the object.

T-Cell Signaling – A T-cell is a type of white blood cell distinguished by having a receptor – i.e., an ability to bind to other molecules. The receptor interacts with intracellular pathway components, starting a cascade of protein interactions called signal transduction. A way to view this process is that a T-cell receptor (TCR) enters a queue upon activation and goes through a series of processes, such as phosphorylation (Wedagedera and Burroughs 2006). Service completion is defined by the deactivation of the TCR, returning it to the inactive pool; however, it is possible that the T-cell's service is aborted before it completes service. Of interest is the probability of activation – i.e., in service for some threshold of

time. If it completes service and detects infection, the T-cell signals cell death (called apoptosis; the second p is silent).

Other queueing models apply to genetic networks, allowing signals that affect the population to enter and leave the system (Arazi et al. 2004; Jamalyaria et al. 2005). This applies queueing to a broad range of self-assembly systems – i.e., form an arrangement without external guidance.

Simulation: Dynamical state evolution is fundamental in both classical mathematical biology and modern systems biology. Evolution and biochemical pathways are prime examples; the underlying state-transition structure and the sheer size are sufficient to need simulation.

The kinetic laws of a biosystem depend upon the objects, particularly their scale (viz., molecules vs. cells). The deterministic rate equations have the form:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = f_i(x;k) \qquad \text{for } i = 1,\ldots,m,$$

where x is the system state (e.g., concentrations of m metabolites) and k is a vector of parameters, called rate constants.

Sources of randomness can be intrinsic – e.g., errors in parameter estimation, or extrinsic – e.g., protein production in random pulses (Meng et al. 2004). To deal with reaction uncertainty, Gillespie (2008, 1977) introduced the probability equation:

$$\Pr(x;t+\mathrm{d}t) = \sum\nolimits_r \Pr(x-v_r;t)a_r(x-v_r)\,\mathrm{d}t + \Pr(x;t) \times \left(1 - \sum\nolimits_r a_r(x)\mathrm{d}t\right),$$

where $a_r(x)\,\mathrm{d}t$ is the probability that reaction r occurs in the time interval $(t, t+\mathrm{d}t)$, changing the state from x to $x + v_r$. The first summation represents being one reaction removed from the state x; the last term represents having no reaction during the interval.

Auto-regulatory Network – Puchalka and Kierzek (2004) consider a metabolic network with regulatory processes and random fluctuations in gene expression. Using Gillespie's equation, given the state x at time t, the probability that the next reaction, r, occurs during $(t+\tau, t+\tau+\mathrm{d}t)$ is given by:

$$\Pr(\tau, r|x,t) = a_r(x)\; e^{-\sum_j a_j(x)\tau}.$$

The simulation is run by generating (τ, r) using this joint density function. The simulation also allows for pulse production – a receptor site may be on or off to regulate gene expression (restricting the choice of r).

Other models use rare-event simulation, such as for tumor development (Abbott 2002). Simulation is used in systems biology to understand how non-dominant pathways affect assembly kinetics (Zhang and Schwartz 2006).

Game Theory: The central idea of game theory is that each player has its own objective to optimize. Historically, evolutionary biologists used game theory to model natural selection (Maynard Smith 1982; Perc and Szolnoki 2010). In OR, game theory is used to model competition for economic resources, and this extends to modeling species-invasion into an existing ecosystem. The same game model applies to the propagation of tumor cells that can mutate to create a cancer population that overwhelms normal cells (Tomlinson 1997). New applications are at the molecular scale, such as the following example.

Protein Binding – There are two sets of players: protein classes (including drugs) and DNA binding sites. Their joint strategies result in allocation of proteins to sites. Sites seek to maximize their occupancy; proteins seek to minimize excess binding. Sites compete for nearby proteins; proteins choose target sites to which they transport. (Mechanisms to achieve these choices are not well understood.) The affinity for protein i to bind to site j is denoted by the constant K_{ij}, but this applies only if the protein is in the proximity of the site.

Let $i = 1,\ldots, N_p$ index proteins and $j = 1,\ldots, N_s$ index sites, and consider the parameters:

$$\begin{aligned} v_i &= \text{nuclear concentration},\\ E_{ij} &= \text{transport affinity},\\ K_{ij} &= \text{binding affinity}. \end{aligned}$$

A protein's decision variable is its fractional transported amounts, $p^i = \left(p^i_0,\ldots,p^i_{N_s}\right)$, where $p^i_0 = 1 - \sum_{j=1}^{N_s} p^i_j$ is the portion of protein i not allocated to a site. A site's decision variable is its choice of binding frequency, $s^j = \left(s^j_0,\ldots,s^j_{N_p}\right)$, where $s^j_0 = 1 - \sum_{i=1}^{N_p} s^i_j$ is the portion of time that site j is unoccupied. There are resource constraints on joint strategies, notably $s^j_i \le p^i_j v_i$ for $i > 0$ – i.e., binding cannot exceed allocated concentration.

A solution is a joint strategy $(\bar{p}, \bar{s})$ that satisfies the optimality criteria:

$$\bar{p}^i \in \underset{p^i \in P(\bar{s})}{\operatorname{argmax}} \left\{ f_p^i(p^i, \bar{s}) \right\} \quad \bar{s}^j \in \underset{s^j \in S(\bar{p})}{\operatorname{argmin}} \left\{ f_s^j(\bar{p}, s^j) \right\},$$

where f_p, f_s denote objective functions for each protein and site, and $P \subseteq \mathbb{R}_+^{N_s+1}$, $S \subseteq \mathbb{R}_+^{N_p+1}$ denote feasible regions, each dependent on the other decisions. An example of objective functions are maximizing total binding affinity and minimizing the amount of protein not assigned:

$$f_p^i(p^i, s) = \sum_{j=1}^{N_s} E_{ij} p_j^i \left(1 - s_0^j\right)$$

$$f_s^j(s^j, p) = s_0^j \sum_{i=1}^{N_p} K_{ij} \left(p_j^i v_i - s_i^j\right).$$

With mild modifications, a solution exists and there is a simple algorithm to find it (Pérez-Breva et al. 2006).

This game model is a simplification of a broader biology, where sites can coordinate, not just compete, and proteins can form complexes to bind to the same site. There are also promoters that bind to a protein in order to send it to another site. Although current thinking is that proteins roam randomly until they bump into an unoccupied site for which they have affinity, the game model attributes a purposeful behavior to proteins, suggesting that they choose to transport to some site. While this rational behavior is not due to intelligence, it could be due to an environmental context that is not yet understood and whose net effect makes proteins behave as if they are rational players.

See

- Dynamic Programming
- Game Theory
- Integer and Combinatorial Optimization
- Linear Programming
- Markov Chains
- Markov Processes
- Network Optimization
- Nonlinear Programming
- Queueing Theory
- Simulation of Stochastic Discrete-Event Systems

References

Abbott, R. (2002). *CancerSim: A computer-based simulation of Hanahan and Weinberg's Hallmarks of Cancer*. Master's thesis, The University of New Mexico, Albuquerque, NM.

Allen, L. J. S. (2003). *An introduction to stochastic processes with applications to biology*. Upper Saddle River, NJ: Pearson Education.

Arazi, A., Ben-Jacob, E., & Yechiali, U. (2004). Bridging genetic networks and queueing theory. *Physica A: Statistical Mechanics and its Applications, 332*, 585–616.

Burkowski, F. (2009). *Structural bioinformatics: An algorithmic approach* (Mathematical and computational biology). Boca Raton, FL: Chapman & Hall/CRC.

Clote, P., & Backofen, R. (2000). *Computational molecular biology*. New York: John Wiley & Sons.

Durbin, R., Eddy, S., Krogh, A., & Mitchison, G. (1998). *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge, UK: Cambridge University Press.

Floudas, C. A., & Pardalos, P. M. (Eds.). (2000). *Local and global approaches. Optimization in computational chemistry and molecular biology*. Dordrecht: Kluwer Academic.

Forrester, R. J., & Greenberg, H. J. (2008). Quadratic binary programming models in computational biology. *Algorithmic Operations Research, 3*(2), 110–129.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical reactions. *The Journal of Physical Chemistry, 81*(25), 2340–2361.

Gillespie, D. T. (2008). Simulation methods in systems biology. In M. Bernardo, P. Degano, & C. Zavattaro (Eds.), *Formal methods for computational systems biology* (LNCS, Vol. 5016, pp. 125–167). Berlin: Springer.

Glodzik, A., & Skolnick, J. (1994). Flexible algorithm for direct multiple alignment of protein structures and sequences. *Bioinformatics, 10*(6), 587–596.

Goldman, D., Istrail, S., Papadimitriou, C. H. (1999). Algorithmic aspects of protein structure similarity. In *40th Annual Symposium on Foundations of Computer Science (FOCS)* (pp 512–521). IEEE Computer Society Press.

Gusfield, D. (1997). *Algorithms on strings, trees, and sequences: Computer science and computational biology*. Cambridge, UK: Cambridge University Press.

Jamalyaria, F., Rohlfs, R., & Schwartz, R. (2005). Queue-based method for efficient simulation of biological self-assembly systems. *Journal of Computational Physics, 204*(1), 100–120.

Jones, N. C., & Pevzner, P. A. (2004). *An introduction to bioinformatics algorithms*. Cambridge, MA: MIT Press.

Lancia, G. (2006). Applications to computational molecular biology. In G. Appa, P. Williams, P. Leonidas, & H. Paul (Eds.), *Handbook on modeling for discrete optimization* (International series in operations research and management science, Vol. 88, pp. 270–304). Berlin: Springer.

Maynard Smith, J. (1982). *The theory of games and the evolution of animal conflicts*. Cambridge, UK: Cambridge University Press.

Meng, T. C., Somani, S., & Dhar, P. (2004). Modeling and simulation of biological systems with stochasticity. *In Silico Biology, 4*(3), 293–309.

Palsson, B. Ø. (2006). *Systems biology: Properties of reconstructed networks*. New York: Cambridge University Press.
Perc, M., & Szolnoki, A. (2010). Coevolutionary games – a mini review. *BioSystems, 99*(2), 109–125.
Pérez-Breva, L., Ortiz, L. E., Yeang, C.-H., & Jaakkola, T. (2006). Game theoretic algorithms for protein-DNA binding. In *Proceedings of the 12th Annual Conference on Neural Information Processing (NIPS)*, Vancouver, Canada.
Puchalka, J., & Kierzek, A. M. (2004). Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks. *Biophysical Journal, 86*(3), 1357–1372.
Steel, M., & Warnow, T. (1993). Kaikoura tree theorems: Computing the maximum agreement subtree. *Information Processing Letters, 48*(3), 77–82.
Tomlinson, I. P. M. (1997). Game-theory models of interactions between tumour cells. *European Journal of Cancer, 33*(9), 1495–1500.
Waterman, M. S. (1995). *Introduction to computational biology: Maps, sequences, and genomes (interdisciplinary statistics)*. Boca Raton, FL: Chapman & Hall/CRC.
Wedagedera, J. R., & Burroughs, N. J. (2006). T-cell activation: A queueing theory analysis at low agonist density. *Biophysical Journal, 91*, 1604–1618.
Wilkinson, D. J. (2006). *Stochastic modelling for systems biology*. Boca Raton, FL: Chapman & Hall/CRC.
Zhang, T., & Schwartz, R. (2006). Simulation study of the contribution of oligomer/oligomer binding to capsid assembly kinetics. *Biophysical Journal, 90*, 57–64.

Computational Complexity

Leslie Hall
The Johns Hopkins University, Baltimore, MD, USA

Introduction

The term computational complexity has two usages which must be distinguished. On the one hand, it refers to an algorithm for solving instances of a problem: broadly stated, the computational complexity of an algorithm is a measure of how many steps the algorithm will require in the worst case for an instance or input of a given size. The number of steps is measured as a function of that size.

The term's second, more important use is in reference to a problem itself. The theory of computational complexity involves classifying problems according to their inherent tractability or intractability — that is, whether they are easy or hard to solve. This classification scheme includes the well-known classes *P* and *NP*; the terms *NP*-complete and *NP*-hard are related to the class *NP*.

Algorithms and Complexity

To understand what is meant by the complexity of an algorithm, algorithms, problems, and problem instances must be defined. Moreover, one must understand how one measures the size of a problem instance and what constitutes a step in an algorithm. A problem is an abstract description coupled with a question requiring an answer; for example, the Traveling Salesman Problem (TSP) is: "Given a graph with nodes and edges and costs associated with the edges, what is a least-cost closed walk (or *tour*) containing each of the nodes exactly once?" An instance of a problem, on the other hand, includes an exact specification of the data: for example, "The graph contains nodes 1, 2, 3, 4, 5, and 6, and edges (1, 2) with cost 10, (1, 3) with cost 14, . . ." and so on. Stated more mathematically, a problem can be thought of as a function p that maps an instance x to an output $p(x)$ (an answer).

An algorithm for a problem is a set of instructions guaranteed to find the correct solution to any instance in a finite number of steps. In other words, for a problem p, an algorithm is a finite procedure for computing $p(x)$ for any given input x. Computer scientists model algorithms by a mathematical construct called a Turing machine, but a more concrete model will be considered here. In a simple model of a computing device, a "step" consists of one of the following operations: addition, subtraction, multiplication, finite-precision division, and comparison of two numbers. Thus if an algorithm requires one hundred additions and 220 comparisons for some instance, then the algorithm requires 320 steps on that instance. In order to make this number meaningful, it should be expressed as a function of the size of the corresponding instance, but determining the exact function would be impractical. Instead, since the main concern is how long the algorithm takes (in the worst case) asymptotically as the size of an instance gets large, one formulates a simple function of the input size that is a reasonably tight upper bound on the actual number of steps. Such a function is called the complexity or running time of the algorithm.

Technically, the *size* of an instance is the number of bits required to encode it. It is measured in terms of the inherent dimensions of the instance (such as the number of nodes and edges in a graph), plus the number of bits required to encode the numerical information in the instance (such as the edge costs). Since numerical data are encoded in binary, an integer C requires about $\log_2 |C|$ bits to encode and so contributes logarithmically to the size of the instance. The running time of the algorithm is then expressed as a function of these parameters, rather than the precise input size. For example, for the TSP, an algorithm's running time might be expressed as a function of the number of nodes, the number of edges, and the maximum number of bits required to encode any edge cost. As was seen, the complexity of an algorithm is only a rough estimate of the number of steps that will be required on an instance. In general — and particularly in analyzing the inherent tractability of a problem — an asymptotic analysis is the main interest: how does the running time grow as the size of the instance gets very large? For these reasons, it is useful to introduce Big-O notation. For two functions $f(t)$ and $g(t)$ of a nonnegative parameter t, $f(t) = O(g(t))$ if there is a constant $c > 0$ such that, for all sufficiently large t, $f(t) \leq cg(t)$. The function $cg(t)$ is thus an asymptotic upper bound on f. For example, $100(t^2 + t) = O(t^2)$, since by taking $c = 101$ the relation follows for $t \geq 100$; however, $0.0001\ t^3$ is not $O(t^2)$. Notice that it is possible for $f(t) = O(g(t))$ and $g(t) = O(f(t))$ simultaneously.

An algorithm is said to run in polynomial time (is a polynomial-time algorithm) if the running time $f(t) = O(P(t))$, where $P(t)$ is a polynomial function of the input size. Polynomial-time algorithms are generally (and formally) considered efficient, and problems for which polynomial time algorithms exist are considered easy. For the remainder of this article, the term polynomial will mean as a function of the input size.

The Classes P and NP

In order to establish a formal setting for discussing the relative tractability of problems, computer scientists first define a large class of problems called recognition (or decision) problems. This class comprises precisely those problems whose associated question requires the answer yes or no. For example, consider the problem of determining whether an undirected graph is connected (that is, whether there a path between every pair of nodes in the graph). This problem's input is a graph G consisting of nodes and edges, and its question is, "Is G connected?" Notice that most optimization problems are not recognition problems, but most have recognition counterparts. For example, a recognition version of the TSP has as input both a graph G, with costs on the edges, *and* a number K. The associated question is, "Does G contain a traveling salesman tour of length less than or equal to K?" In general, an optimization problem is not much harder to solve than its recognition counterpart. One can usually embed the recognition algorithm in a binary search over the possible objective function values to solve the optimization problem with a polynomial number of calls to the embedded algorithm.

The class P is defined as the set of recognition problems for which there exists a polynomial-time algorithm, where P stands for polynomial time. Thus, P comprises those problems that are formally considered easy. The larger problem class NP contains the class P. The term NP stands for nondeterministic polynomial and refers to a different, hypothetical model of computation, which can solve the problems in NP in polynomial time (for further explanation, see references).

The class NP consists of all recognition problems with the following property: for any yes-instance of the problem there exists a polynomial-length certificate or proof of this fact that can be verified in polynomial time. The easiest way to understand this idea is by considering the position of an omniscient being (say, the wizard Merlin) who is trying to convince a mere mortal that some instance is a yes-instance. Suppose the problem is the recognition version of the TSP, and the instance is a graph G and the number $K = 100$. Merlin knows that the instance does contain a tour with length at most 100. To convince the mortal of this fact, he simply hands her a list of the edges of this tour. This list is the certificate: it is polynomial in length, and the mortal can easily verify, in polynomial time, that the edges do in fact form a tour with length at most 100.

There is an inherent asymmetry between yes and no in the definition of NP. For example, there is no obvious, succinct way for Merlin to convince

a mortal that a particular instance does NOT contain a tour with length at most 100. In fact, by reversing the roles played by yes and no leads to a problem class known as *Co-NP*. In particular, for every recognition problem in *NP* there is an associated recognition problem in *Co-NP* obtained by framing the *NP* question in the negative (e.g., "Do *all* traveling salesman tours in *G* have length *greater* than *K*?"). Many recognition problems are believed to lie outside both of the classes *NP* and *Co-NP*, because they seem to possess no appropriate certificate. An example would be the problem consisting of a graph *G* and two numbers *K* and *L*, with the question, "Is the number of distinct traveling salesman tours in *G* with length at most *K* exactly equal to *L*?"

NP-Complete Problems

To date, no one has found a polynomial-time algorithm for the TSP. On the other hand, no one has been able to prove that no polynomial-time algorithm exists for the TSP. How, then, can one argue persuasively that the TSP and many problems in *NP* are intractable? Instead, an argument is presented that is slightly weaker but also compelling. It is shown that the recognition version of the TSP, and scores of other *NP* problems, are the *hardest* problems in the class *NP* in the following sense: if there is a polynomial-time algorithm for any one of these problems, then there is a polynomial-time algorithm for every problem in *NP*. Observe that this is a very strong statement, since *NP* includes a large number of problems (such as integer programming) that appear to be extremely difficult to solve, both in theory and in practice! Problems in *NP* with this property are called *NP*-complete. Otherwise stated, it seems highly unlikely that a polynomial algorithm will be found for any *NP*-complete problem, since such an algorithm would actually provide polynomial time algorithms for *every* problem in *NP*!

The class *NP* and the notion of complete problems for *NP* were first introduced by Cook (1971). In that paper, he demonstrated that a particular recognition problem from logic, SATISFIABILITY, was *NP*-complete, by showing directly how every other problem in *NP* could be encoded as an appropriate special case of SATISFIABILITY. Once the first *NP*-complete problem had been established, however, it became easy to show that others were *NP*-complete. To do so requires simply providing a polynomial transformation from a known *NP*-complete problem to the candidate problem. Essentially, one needs to show that the known hard problem, such as SATISFIABILITY, is a special case of the new problem. Thus, if the new problem has a polynomial-time algorithm, then the known hard problem has one as well.

Related Terms

The term *NP*-hard refers to any problem that is at least as hard as any problem in *NP*. Thus, the *NP*-complete problems are precisely the intersection of the class of *NP*-hard problems with the class *NP*. In particular, optimization problems whose recognition versions are *NP*-complete (such as the TSP) are *NP*-hard, since solving the optimization version is at least as hard as solving the recognition version.

The polynomial hierarchy refers to a vast array of problem classes both beyond *NP* and *Co-NP* and within. There is an analogous set of definitions which focuses on the space required by an algorithm rather than the time, and these time and space definitions roughly correspond in a natural way. There are complexity classes for parallel processing, based on allowing a polynomial number of processors. There are classes corresponding to randomized algorithms, those that allow certain decisions in the algorithm to be made based on the outcome of a coin toss. There are also complexity classes that capture the notions of optimization and approximability. The most famous open question concerning the polynomial hierarchy is whether the classes *P* and *NP* are the same, i.e., $P \stackrel{?}{=} NP$. If a polynomial algorithm were discovered for any *NP*-complete problem, then all of *NP* would collapse to *P*; indeed, most of the polynomial hierarchy would disappear.

In algorithmic complexity, two other terms are heard frequently: strongly polynomial and pseudo-polynomial. A strongly polynomial-time algorithm is one whose running time is bounded polynomially by a function *only* of the inherent dimensions of the problem and independent of the sizes of the numerical data. For example, most sorting algorithms are strongly polynomial, since they normally require a number of comparisons polynomial in the number of entries and do not depend on the actual values being

sorted; an algorithm for a network problem would be strongly polynomial if its running time depended only on the numbers of nodes and arcs in the network, and not on the sizes of the costs or capacities.

A pseudo-polynomial-time algorithm is one that runs in time polynomial in the dimension of the problem and the magnitudes of the data involved (provided these are given as integers), rather than the base-two logarithms of their magnitudes. Such algorithms are technically exponential functions of their input size and are therefore not considered polynomial. Indeed, some *NP*-complete and *NP*-hard problems are pseudo-polynomially solvable (sometimes these are called weakly *NP*-hard or-complete, or *NP*-complete in the ordinary sense). For example, the *NP*-hard knapsack problem can be solved by a dynamic programming algorithm requiring a number of steps polynomial in the size of the knapsack and the number of items (assuming that all data are scaled to be integers). This algorithm is exponential-time since the input sizes of the objects and knapsack are logarithmic in their magnitudes. However, as Garey and Johnson (1979) observe, "A pseudo-polynomial-time algorithm. . .will display 'exponential behavior' only when confronted with instances containing 'exponentially large' numbers, [which] might be rare for the application we are interested in. If so, this type of algorithm might serve our purposes almost as well as a polynomial time algorithm." The related term strongly *NP*-complete (or unary *NP*-complete) refers to those problems that remain *NP*-complete even if the data are encoded in unary (that is, if the data are small relative to the overall input size). Consequently, if a problem is strongly *NP*-complete then it cannot have a pseudo-polynomial-time algorithm unless $P = NP$.

For textbook introductions to the subject, see Papadimitriou (1993) and Sipser (1997). The most important reference on the subject, Garey and Johnson (1979), contains an outstanding, relatively compact introduction to complexity. Further references, including surveys and full textbooks, are given below.

See

- ▶ Combinatorics
- ▶ Graph Theory
- ▶ Integer and Combinatorial Optimization

References

Arora, S., & Barak, B. (2009). *Computational complexity: A modern approach*. Cambridge, UK: Cambridge University Press.

Bovet, D. P., & Crescenzi, P. (1994). *Introduction to the theory of complexity*. Englewood Cliffs, NJ: Prentice-Hall.

Cook, S. A. (1971). The complexity of theorem-proving procedures. *Proceedings of the 3rd Annual ACM Symposium on Theory of Computing*, 151–158.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. New York: W.H. Freeman.

Karp, R. M. (1975). On the computational complexity of combinatorial problems. *Networks, 5*, 45–68.

Lewis, H. R., & Papadimitriou, C. H. (1997). *Elements of the theory of computation* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Papadimitriou, C. H. (1985). Computational complexity. In E. L. Lawler, J. K. Lenstra, A. H. G. Rinnooy Kan, & D. B. Shmoys (Eds.), *The traveling salesman problem: A guided tour of combinatorial optimization*. Chichester, UK: Wiley.

Papadimitriou, C. H. (1993). *Computational complexity*. Redwood City, CA: Addison-Wesley.

Papadimitriou, C. H., & Steiglitz, K. (1982). *Combinatorial optimization: Algorithms and complexity*. Englewood Cliffs, NJ: Prentice-Hall.

Shmoys, D. B., & Tardos, E. (1989). Computational complexity of combinatorial problems. In L. Lovász, R. L. Graham, & M. Groetschel (Eds.), *Handbook of combinatorics*. Amsterdam: North-Holland.

Sipser, M. (1997). *Introduction to the theory of computation*. Belmont, CA: PWS-Kent.

Stockmeyer, L. J. (1990). Complexity theory. In E. G. Coffman Jr., J. K. Lenstra, & A. H. G. Rinnooy Kan (Eds.), *Handbooks in operations research and management science* (Computation, Chapter 8, Vol. 3). Amsterdam: North Holland.

Computational Geometry

Isabel M. Beichl[1], Javier Bernal[1], Christoph Witzgall[1] and Francis Sullivan[2]

[1]National Institute of Standards & Technology, Gaithersburg, MD, USA

[2]Supercomputing Research Center, Bowie, MD, USA

Introduction

Computational geometry is the discipline of exploring algorithms and data structures for computing geometric objects and their often extremal attributes. The objects are predominantly finite collections

of points, flats, hyperplanes, arrangements, or polyhedra, all in finite dimensions. The algorithms are typically finite, their complexity playing a central role. Emphasis is on problems in low dimensions, exploiting special properties of the plane and 3-space.

A relatively young field; its name coined in the early 1970s. It has since witnessed explosive growth, stimulated in part by the largely parallel development of computer graphics, pattern recognition, cluster analysis, and modern industry's reliance on computer-aided design (CAD) and robotics (Forrest 1971; Graham and Yao 1990; Lee and Preparata 1984). It plays a key role in the emerging fields of automated cartography and computational metrology.

The *Handbook of Discrete and Computational Geometry*, edited by Goodman and O'Rourke (1997), provides overviews of key topics. For general texts, see Preparata and Shamos (1985), O'Rourke (1987), Edelsbrunner (1987), and, de Berg et al. (2008). Pertinent concepts of discrete geometry are presented in Grünbaum (1967).

There are strong connections to operations research, whose classical problems such as finding a minimum spanning tree, a maximum-length matching, or a Steiner tree become problems in computational geometry when posed in Euclidean or related normed linear spaces. The Euclidean traveling salesman problem remains *NP*-complete (Papadimitriou 1977). Facility location, and shortest paths in the presence of obstacles, are other examples. Polyhedra and their extremal properties, typical topics of computational geometry, also lie at the foundation of linear programming. Its complexity, particularly in lower dimensions, attracted early computational geometric research, heralding the achievement of linear complexity for arbitrary fixed dimension (Megiddo 1982, 1984; Clarkson 1986).

Problems

A fundamental problem is to determine the convex hull conv(S) of a set S of n points in d-dimensional Cartesian space $\mathfrak{R}^d$. This problem has a weak and a strong formulation. Its weak formulation requires only the identification of the extreme points of conv(S). In operations research terms, that problem is well known as (the dual of) identifying redundant constraints in a system of linear inequalities. The strong formulation requires, in addition, characterization of the facets of the polytope conv(S). For dimension $d > 3$, the optimal complexity of the strong convex hull problem in $\mathfrak{R}^d$ is $O(n^{\lceil d/2 \rfloor})$ (Chazelle 1991).

Early $O(n \log n)$ methods for delineating convex hulls in the plane — vertices and edges of the convex hull of a simple polygon can be found in linear time — were based on divide-and-conquer (Graham and Yao 1983) and (Preparata and Hong 1977). In this widely used recursive strategy, a problem is divided into subproblems whose solutions, having been obtained by further subdivision, are then combined to yield the solution to the original problem. Divide-and-conquer heuristics find applications in Euclidean optimization problems such as optimum-length matching (Reingold and Supowit 1983).

The following bridge problem is, in fact, a linear program: given two sets S_1 and S_2 of planar points separated by a line, find two points $p_1 \varepsilon S_1$ and $p_2 \varepsilon S_2$ such that the line segment $[p_1, p_2]$ is an upper edge of the convex hull conv($S_1 \cup S_2$), bridging the gap between the two sets. Or, through which edge does a given directed line leave the — not yet delineated — convex hull of n points in the plane? As a linear program of fixed dimension 2, the bridge problem can be solved in linear time. Kirkpatrick and Seidel (1986) have used it along with a divide-and-conquer paradigm to devise an $O(n \log m)$ algorithm for the planar convex hull of n points, m of which are extreme.

When implementing a divide-and-conquer strategy, one typically wishes to divide a set of points $S \subset \mathfrak{R}^d$ by a straight line into two parts of essentially equal cardinality, that is, to execute a ham-sandwich cut. This can be achieved by finding the median of, say, the first coordinates of the points in S. It is a fundamental result of the theory of algorithms that the median of a finite set of numbers can be found in linear time. The bridge problem is equivalent to a double ham-sandwich cut of a planar set: given a first cut, find a second line quartering the set. Threeway cuts in three dimensions and results about higher dimensions were reported in Dobkin and Edelsbrunner (1984).

The Euclidean post office problem is a prototype for a class of proximity search problems encountered, for instance, in the implementation of expert systems. Sites p_i of n post offices in $\mathfrak{R}^d$ are given, and the task is

to provide suitable preprocessing for efficiently identifying a post office closest to any client location.

Associated with this problem is the division of space into postal regions, that is, sets of locations $V_i \subset \Re^d$ closer to postal site p_i than to any other site p_j. Each such region V_i around site p_i is a convex polyhedron, whose facets are determined by perpendicular bisectors, that is, (hyper)planes or lines of equal distance from two distinct sites. Those polyhedra form a polyhedral complex covering $\Re^d$ known as a Voronoi diagram. The Voronoi diagram and its dual, the Delaunay triangulation, are important related concepts in computational geometry.

Once a Delaunay triangulation of a planar set of n sites has been established, an $O(n \log n)$ procedure, a pair of nearest points among these sites can be found in linear time. The use of Delaunay triangulations for computational geometric problems was pioneered by Shamos and Hoey (1975).

The problem of efficiently finding a Voronoi cell V_i for an arbitrary query point p is an example of point location in subdivisions. Practical algorithms for locating a given point in a subdivision of the plane generated by n line segments in time $O(\log n)$ requiring preprocessing of order $O(n \log n)$ and storage of size $O(n \log n)$ or $O(n)$, respectively, have been proposed (Preparata 1990). For point location in planar Voronoi diagrams, Edelsbrunner and Maurer (1985) utilized acyclic graphs and packing. A probabilistic approach to the post office problem is given in Clarkson (1985).

Whether a given point lies in a certain simple polygon can be decided by an $O(n)$ process of examining the boundary intersections of an arbitrary ray emanating from the point in question. For convex polygons, an $O(n)$ preprocessing procedure permits subsequent point inclusion queries to be answered in $O(\log n)$ time (Bentley and Carruthers 1980).

An important concept with operations research implications is the medial axis of a polygon (Lee 1982), the locus of interior points with equal distance from the boundary; more precisely, those interior points with more than one closest boundary point. Such medial axes may be obtained in $O(n)$ time (Chin et al. 1995).

Let $h_e(x)$ be the truth function expressing point inclusion in the half plane to the left of a directed line segment e. Muhidinov and Nazirov (1978) have shown that a polygonal set can be characterized by a Boolean expression of n such functions, one for each edge e of the polygonal set, where each such function occurs only once in the expression. This Boolean expression transforms readily to an algebraic expression for the characteristic function of the polygon. For 3-dimensional polyhedral bodies, Dobkin, Guibas, Hershberger, and Snoeyink (1988) investigated the existence and determination of analogous constructive solid geometry (CSG) representations (they may require repeats of half space truth functions). In general, CSG representations use Boolean operations to combine primitive shapes, and are at the root of some commercial CAD/CAM and display systems. For a survey of methods for representing solid objects see Requicha (1980).

Given a family of polygons, a natural generalization of point inclusion is to ask how many of those polygons include a query point. This and similar intersection-related problems are subsumed under the term stabbing. The classical 1-dimensional stabbing problem involves n intervals. Here the stabbing number can be found in $O(\log n)$ time and $O(n)$ space after suitable preprocessing. Similar results hold for special classes of polygons such as rectangles (Edelsbrunner 1983).

Sweep-techniques rival divide-and-conquer in popularity. Plane-sweep or line-sweep, for instance, conceptually moves a vertical line from left to right over the plane, registering objects as it passes them. Plane-sweep permits one to decide in $O(n \log n)$ time (optimal complexity) whether n line segments in the plane have at least one intersection (Shamos and Hoey 1976).

Important special cases of the above intersection problem are testing for (self-)intersection of paths and polygons. Polygon simplicity can be tested for in linear time by trying to triangulate the polygon.

Polygon triangulation, more precisely, decomposing the interior of a simple polygon into triangles whose vertices are also vertices of the polygon, is a celebrated problem of computational geometry. In a seminal paper, Garey, Johnson, Preparata, and Tarjan (1978) proposed an $O(n \log n)$ algorithm for triangulating a simple polygon of n vertices. They used a plane sweep approach for decomposing the polygon into monotone polygons, which can each be triangulated in linear time. A related idea is to provide a trapezoidization of the polygon, from which a triangulation can be obtained in linear time. Chazelle (1990) introduced the concept of

a visibility map, a tree structure which might be considered a local trapezoidization of the polygon, and based on it an $O(n)$ triangulation algorithm for simple polygons. In 3-space, an analogous tetrahedralization (without additional Steiner points for vertices) for nonconvex polyhedral bodies may not exist. Moreover, the problem of deciding such existence is *NP*-complete (Ruppert and Seidel 1989).

For algorithms that depend on sequential examination of objects, bucketing or binning may improve performance by providing advantageous sequencing (Devroye 1986). The idea is to partition an area into a regular pattern of simple shapes such as rectangles to be traversed in a specified sequence. The problem at hand is then addressed locally within buckets or bins followed by adjustments between subsequent or neighboring buckets. Bucketing-based algorithms have provided practical solutions to Euclidean optimization problems, such as shortest paths, optimum-length matching, and a Euclidean version of the Chinese Postman Problem: minimizing the pen movement of a plotter (Asano et al. 1985). The techniques of quadtrees and octrees might be considered as hierarchical approaches to bucketing, and are often the methods of choice for image processing and spatial data analysis including surface representation (Samet 1990a, b).

The position of bodies and parts of bodies, relative to each other in space, determines visibility from given vantage points, shadows cast upon each other, and impediments to motion. Hidden line and hidden surface algorithms are essential in computer graphics, as are procedures for shadow generation and shading (Sutherland et al. 1974; Atherton et al. 1978). Franklin (1980) used bucketing techniques for an exact hidden surface algorithm.

Lozano-Pérez and Wesley (1979) used the concept of a visibility graph for planning collision-free paths: given a collection of mutually disjoint polyhedral objects, the node set of the above graph is the set of all vertices of those polyhedral objects, and two such nodes are connected if the two corresponding vertices are visible from each other.

The piano movers problem captures the essence of motion planning (Schwartz and Sharir 1983, 1989). Here a 2-dimensional polygonal figure, or a line segment (ladder), is to be moved, both translating and rotating, amidst polygonal barriers.

Geometric objects encountered in many areas such as Computer-Aided Design (CAD) are fundamentally nonlinear (Dobkin and Souvaine 1990). The major thrust is generation of classes of curves and surfaces with which to interpolate, approximate, or generally speaking, represent data sets and object boundaries (Barnhill 1977; Bartels et al. 1987; Farin 1988). A classical approach, building on the concepts of splines and finite elements, has been to use piecewise polynomial functions over polyhedral tilings such as triangulations. Examples are the TIN (triangulated irregular network) approach popular in terrain modeling, C^1 functions over triangulations, and the arduous solution of the corresponding C^2 problem (Heller 1990; Lawson 1977; Alfeld and Barnhill 1984).

Bézier curves and surfaces involve an elegant concept: the use of control points to define elements of curves and surfaces, permitting intuition-guided manipulation important in CAD (Forrest 1972). In general, polynomials are increasingly supplanted by rational functions, which suffer fewer oscillations per numbers of coefficients (Tiller 1983). All these techniques culminate in NURBS (non-uniform rational B-splines) which are recommended for curve and surface representation in most industrial applications.

In geometric calculations, round-off errors due to floating-point arithmetic may cause major problems (Fortune and Milenkovic 1991). When testing, for instance, whether given points are collinear, a tolerance level, eps, is often specified, below which deviations from a collinearity criterion are ignored. Points p_1, p_2, p_3 and p_2, p_3, p_4, but not p_1, p_2, p_4 may thus be found collinear. Such and similar inconsistencies may cause a computation to abort. Robust algorithms are constructed so as to avoid breakdown due to inconsistencies caused by round-off (Guibas et al. 1989; Beichl and Sullivan 1990). Alternatively, various forms of exact arithmetic are increasingly employed (Fortune and Van Wyck 1993; Yap 1993). Inconsistencies occur typically whenever an inequality criterion is satisfied as an equality. An example is the degeneracy behavior of the simplex method of linear programming. Lexicographic perturbation methods can be employed to make consistent selections of subsequent feasible bases and thus assure convergence. Similar consistent tie breaking, coupled with exact arithmetic, is the aim of the simulation of simplicity approach proposed by Edelsbrunner and Mücke (1988) in a more general computational context.

See

- Chinese Postman Problem
- Cluster Analysis
- Convex Hull
- Facility Location
- Minimum Spanning Tree Problem
- Simplex Method (Algorithm)
- Splines
- Traveling Salesman Problem
- Voronoi Constructs

References

Alfeld, P., & Barnhill, R. E. (1984). A transfinite C 2 interpolant over triangles. *Rocky Mountain Journal of Mathematics, 14*, 17–39.

Asano, T., Edahiro, M., Imai, H., & Iri, M. (1985). Practical use of bucketing techniques in computational geometry. In G. T. Toussaint (Ed.), *Computational geometry*. New York: North Holland.

Atherton, P., Weiler, K., & Greenberg, D. P. (1978). Polygon shadow generation. *Computers and Graphics, 12*, 275–281.

Barnhill, R. E. (1977). Representation and approximation of surfaces. In J. R. Rice (Ed.), *Mathematical software III*. New York: Academic Press.

Bartels, R. H., Beatty, J. C., & Barski, B. A. (1987). *An introduction to splines for use in computer graphics*. Los Altos, CA: Morgan Kaufmann.

Beichl, I., & Sullivan, F. (1990). A robust parallel triangulation and shelling algorithm. *Proceedings of 2nd Canadian Conference on Computational Geometry*, 107–111.

Bentley, J. L., & Carruthers, W. (1980). Algorithms for testing the inclusion of points in polygons. *Proceedings of 18th Allerton Conference on Communication, Control and Computing*, 11–19.

Bentley, J. L., Weide, B. W., & Yao, A. C. (1980). Optimal expected-time algorithms for closest point problems. *ACM Transactions on Mathematical Software, 6*, 563–580.

Chazelle, B. (1990). Triangulating the simple polygon in linear time. *Proceedings of 31st Annual IEEE Symposium on the Foundations of Computer Science*, 220–230.

Chazelle, B. (1991). An optimal convex hull algorithm and new results on cuttings. *Proceedings of 32nd Annual IEEE Symposium on the Foundations of Computer Science*, 29–38.

Chin, F., Snoeyink, J., & Wang, C. A. (1995). Finding the medial axis of a simple polygon in linear time. *Proceedings 6th Annual International Symposium on Algorithms and Computation*. Lecture notes in computer science (Vol. 1004, pp. 382–391). New York: Springer.

Clarkson, K. L. (1985). A probabilistic algorithm for the post office problem. *Proceedings of the 17th Annual ACM Symposium on Theory Computation*, 175–184.

Clarkson, K. L. (1986). Linear programming in$_{0(n3}$ ^{d2}) time. *Information Processing Letters, 22*, 21–24.

de Berg, M., Cheong, O., van Kreveld, M., & Overmars, M. (2008). *Computational geometry: Algorithms and applications* (3rd ed.). New York: Springer.

Devroye, L. (1986). *Lecture notes on bucket algorithms*. Boston: Birkhäuser.

Dobkin, D. P., & Edelsbrunner, H. (1984). Ham-sandwich theorems applied to intersection problems. *Proceedings of 10th International Workshop Graph-Theoretic Concepts in Computer Science (WG 84)*, 88–99.

Dobkin, D., Guibas, L., Hershberger, J., & Snoeyink, J. (1988). An efficient algorithm for finding the CSG representation of a simple polygon. *Computer Graphics, 22*, 31–40.

Dobkin, D. P., & Souvaine, D. L. (1990). Computational geometry in a curved world. *Algorithmica, 5*, 421–457.

Edelsbrunner, H. (1983). A new approach to rectangle intersections, parts I and II. *International Journal of Computer Mathematics, 13*(209–219), 221–229.

Edelsbrunner, H. (1987). *Algorithms in combinatorial geometry*. New York: Springer.

Edelsbrunner, H., & Maurer, H. A. (1985). Finding extreme points in three dimensions and solving the post-office problem in the plane. *Information Processing Letters, 21*, 39–47.

Edelsbrunner, H., & Mücke, E. P. (1988). Simulation of simplicity: a technique to cope with degenerate algorithms. *Proceedings of the 4th Annual ACM Symposium on Computational Geometry*, 118–133.

Farin, G. (1988). *Curves and surfaces for computer aided geometric design*. New York: Academic Press.

Forrest, A. R. (1971). Computational geometry. *Proceedings of the Royal Society of London Series A, 321*, 187–195.

Forrest, A. R. (1972). Interactive interpolation and approximation by bézier polynomials. *The Computer Journal, 15*, 71–79.

Fortune, S., & Milenkovic, V. (1991). Numerical stability of algorithms for line arrangements. *Proceedings of the 7th Annual ACM Symposium on Computational Geometric*, 3342–341.

Fortune, S., & Van Wyck, C. (1993). Efficient exact arithmetic for computational geometry. *ACM Symposium on Computational Geometry*, Vol. 9, 163–172.

Franklin, W. R. (1980). A linear time exact hidden surface algorithm. *Proceedings of the SIGGRAPH '80, Computer Graphics*, Vol. 14, pp. 117–123.

Garey, M. R., Johnson, D. S., Preparata, F. P., & Tarjan, R. E. (1978). Triangulating a simple polygon. *Information Processing Letters, 7*, 175–179.

Graham, R. L., & Yao, F. F. (1983). Finding the convex hull of a simple polygon. *Journal of Algorithms, 4*, 324–331.

Graham, R., & Yao, F. (1990). A whirlwind tour of computational geometry. *The American Mathematical Monthly, 97*, 687–701.

Grünbaum, B. (1967). *Convex polytopes*. New York: Wiley Interscience.

Guibas, L. J., Salesin, D., & Stolfi, J. (1989). Epsilon geometry: Building robust algorithms from imprecise computations. *Proceedings 5th Annual ACM Symposium on Computational Geometry*, 208–217.

Heller, M. (1990). Triangulation algorithms for adaptive terrain modeling. *4th Symposium on Spatial Data Handling*, 163–174.

Kirkpatrick, D. (1983). Optimal search in planar subdivisions. *SIAM Journal on Computing, 12*, 28–35.
Kirkpatrick, D. G., & Seidel, R. (1986). The ultimate planar convex hull algorithm? *SIAM Journal on Computing, 15*, 287–299.
Lawson, C. L. (1977). Software for C^1 surface interpolation. In J. R. Rice (Ed.), *Mathematical software III*. New York: Academic.
Lee, D. T. (1982). Medial axis transformation of a planar shape. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 4*, 363–369.
Lee, D. T., & Preparata, F. P. (1984). Computational geometry–A survey. *IEEE Transactions on Computers, c-33*, 1072–1101.
Lozano-Pérez, T., & Wesley, M. A. (1979). An algorithm for planning collision-free paths among polyhedral obstacles. *Communications of the ACM, 22*, 560–570.
Megiddo, N. (1982). Linear-time algorithms for linear programming in R^3 and related problems. *Proceedings of the 23rd Annual IEEE Symposium on the Foundations of Computer Science*, 329–338.
Megiddo, N. (1984). Linear programming in linear time when the dimension is fixed. *Journal of the ACM, 31*, 114–127.
Muhidinov, N., & Nazirov, S. (1978). Computerized recognition of closed plane domains. *Voprosy Vychislitel'noj i Prikladnoj Matematiki (Tashkent)*, *53*, 96–107, 182.
O'Rourke, J. (1987). *Art gallery theorems and algorithms*. New York: Oxford University Press.
Papadimitriou, C. H. (1977). The euclidean traveling salesman problem is *NP*-complete. *Theoretical Computer Science, 4*, 237–244.
Preparata, F. P. (1990). Planar point location revisited. *International Journal of Foundations of Computer Science, 24*(1), 71–86.
Preparata, F. P., & Hong, S. J. (1977). Convex hulls of finite sets of points in two and three dimensions. *Communications of the ACM, 20*, 87–93.
Preparata, F. P., & Shamos, M. I. (1985). *Computational geometry: An introduction*. New York: Springer.
Reingold, E. M., & Supowit, K. J. (1983). Probabilistic analysis of divide-and-conquer heuristics for minimum weighted euclidean matching. *Networks, 13*, 49–66.
Requicha, A. A. G. (1980). Representations for rigid solids: Theory, methods, and systems. *ACM Computing Surveys, 12*, 437–464.
Ruppert, J., & Seidel, R. (1989). On the difficulty of tetrahedralizing 3-dimensional non-convex polyhedra. *Proceedings of the 5-th Annual ACM Symposium on Computational Geometry*, 380–392.
Samet, H. (1990a). *The design and analysis of spatial data structures*. Reading, PA: Addison Wesley.
Samet, H. (1990b). *Applications of spatial data structures: Computer graphics, image processing and GIS*. Reading, PA: Addison Wesley.
Schwartz, J. T., & Sharir, M. (1983). On the 'piano movers' problem, I: The case of a two-dimensional rigid polygonal body moving amidst polygonal barriers. *Communications on Pure and Applied Mathematics, 36*, 345–398.
Schwartz, J. T., & Sharir, M. (1989). A survey of motion planning and related geometric algorithms. In D. Kapur & J. Mundy (Eds.), *Geometric reasoning* (pp. 157–169). Cambridge, MA: MIT Press.
Shamos, M. I., & Hoey, D. (1975). Closest-point problems. *Proceedings of the 16th Annual IEEE Symposium on the Foundations of Computer Science*, 151–162.
Shamos, M. I., & Hoey, D. (1976). Geometric intersection problems. *Proceedings of the 17th Annual IEEE Symposium on the Foundations of Computer Science*, 208–215.
Sutherland, I. E., Sproull, R. F., & Shumacker, R. A. (1974). A characterization of ten hidden surface algorithms. *ACM Computing Surveys, 6*, 1–55.
Tiller, W. (1983). Rational B–splines for curve and surface representation. *IEEE Computer Graphics and Applications, 3*(6), 61–69.
Yap, C. (1993). Towards exact geometric computation. *Proceedings of the 5th Canadian Conference on Computational Geometry*, 405–419.

Computational Intelligence

► Artificial Intelligence

Computational Organization Theory

Terrill L. Frantz[1], Kathleen M. Carley[2] and William A. Wallace[3]
[1]Peking University, Shenzhen, Guangdong, China
[2]Carnegie Mellon University, Pittsburgh, PA, USA
[3]Rensselaer Polytechnic Institute, Troy, NY, USA

Introduction

As inexpensive and massive amounts of computing power have rapidly become more widely available, the operational aspects of computational-based organizational research have become a reality. Today, the concepts of Computational Organization Theory (COT) can be easily implemented and practiced by an ever-increasingly larger group of researchers. Some foresee such computer-science related computational thinking (Wing 2006), as the future of all scholarly research, and COT is part of this broader trend.

COT involves the theorizing about, describing, understanding, and predicting the behavior of organizations and the process of organizing, using

quantitative-based and structured approaches (computational, mathematical and logical models). This involves computational abstractions that are incorporated into organizational research and practice through COT tools, procedures, measures and knowledge.

The notion of an organization, as used here, spans the wide range of human-conceived collections of people, i.e., groups, teams, societies, corporations, industries, and governments, see Carley and Prietula, (1994); Prietula, Carley, and Gasser, (1998); and Gilbert and Doran, (1994). COT practitioners use computational models and analysis to develop a better understanding of fundamental principles for organizing and behaviors within an organization. Organizational members, i.e., people, are considered information-processing actors. They can interact with and adapt to their environment. They can learn, and they can communicate. While their behavior is certainly complex, this behavior and the underlying determinate of the behaviors can be reduced to basic mathematical equations and algorithms. With this formalization, researchers can develop complete computerized models of an organization, which enables the use of computer simulation to create virtual worlds for non-obtrusive experimentation. After running these simulations the collective outcome of these virtual interactions and behaviors can be quantified and collected for extensive analysis. Typically, the results from these experiments are then incorporated into a formalized and thoughtful comparison against findings from controlled lab experiments and real-world empirical cases studies. The history of COT is rich with academic insight, with its research and application proving fruitful to organization researchers and practitioners alike.

History

The field of COT has benefitted from several decades of research. One of the earliest works is Cyert and March's *The Behavioral Theory of the Firm*, (1963), in which a simple information-processing model of an organization is used to address issues of organization design and performance. During the past decade an explosion of interest has occurred for theory development and testing in the organizational and social sciences (Carley 1995). The use is expanding for a number of reasons: (a) there is growing recognition that social and organizational processes are complex, dynamic, adaptive, and nonlinear, and, thus, are hard to study in the real-world; (b) researchers and practitioners have come to realize that organizational and social behavior emerges from interactions within and between ecologies of entities (people, groups, technologies, agents, etc.), which is hard to reproduce and control in the laboratory and real-world; and (c) the relationships among these entities are critical constraints on individual and organizational action, which is hard to control with direct human-based research. Researchers now recognize that organizations are inherently computational since they have a need to scan and observe their environment, store facts and programs, communicate among members and with their environment, and transform information by human or automated decision making (Burton and Obel 1996).

COT has a fundamentally interdisciplinary intellectual history with contributions from social network theory, distributed artificial intelligence and the organizational information processing tradition. Within COT, researchers draw heavily on work in the information/resource processing tradition (Simon 1947; March and Simon 1958; Thompson 1967; Galbraith 1973; Cyert and March 1963; Pfeffer and Salancik 1978) and social information processing (Salancik and Pfeffer 1978), as modified by work in cognitive science (Carley and Newell 1994), institutionalism (Powell and DiMaggio 1991), population ecology (Hannan and Freeman 1977, 1989), and the contemporary contingency theory (Baligh et al. 1990). Within social network and communication/coordination theory, there has been important work done on measures of organizational design and communication (Wasserman and Faust 1994; Malone 1986), cognitive social structures (Krackhardt 1987), network effects on performance, influence, and power (Wasserman and Galaskiewicz 1994; Kaufer and Carley 1993; Granovetter 1985; Burt 1992), and research on inter-organizational networks (Baum and Oliver 1991; Stuart and Podolny 1996). Within the area of distributed artificial intelligence, researchers draw on findings regarding representation (Lesser and Corkill 1988); teams (Decker 1996); coordination (Durfee and Montgomery 1991); and strategy (Gasser and Majchrzak 1994).

Methodological Approaches

Models are both integral and integrating components of theory building in COT. No matter what their disciplinary home, researchers in this area assume that meaningful and predictive models of organizations can be built. Computational organizational theorists use models to (1) describe organizational phenomena observed in the world, including structuring real or hypothetical experiences as described or postulated by individuals or groups, (2) formalize and integrate theoretical principles from science that are relevant to organizational activities, and (3) simulate the dynamics of temporal changes in a particular organizational process, action, or policy.

Models are abstractions of reality, and modeling is the process of creating these abstractions. Because reality is near infinitely complex and all empirical data are processed with reference to that complexity, model building involves the simplification of reality as data are transformed into knowledge. The models created are, essentially, forms of codified knowledge and used to represent the reality of things not known from things that are known (Waisel et al. 1998). Modeling is the sine qua non of science. Virtually all-scientific activities require modeling in some sense, and any scientific theory requires this kind of representational system (Nersessian 1992). Models are usually thought of as being quantitative, and able to be represented mathematically. However, qualitative models are no less and arguably more common, particularly in the context of COT.

Employing a variety of methodologies has made advances in computational organization theory. To illustrate this variety, five of the most significant approaches to modeling will be discussed: (1) general intellective models, (2) distributed artificial intelligence and multi-agent models, (3) organizational engineering models, (4) social network models, and (5) mathematical and/or logic based models.

Organizational theorists are most familiar with the general intellective models. These models often represent the organization or various processes as a set of nonlinear equations and/or a set of interacting agents. In these models, the focus is on explaining and theorizing about a particular aspect of organizational behavior. Consequently, the models often abstract many of the factors in actual organizations, laying bear only the entities and relations essential to the theory. Models embody theory about how the team, group, or organization will behave. Given these models, a series of virtual experiments are run to test the effect of a change in a particular process, action, policy, etc. These models are used to illustrate the theory's story about how the organization will behave under various conditions. These models enable cumulative theory building as multiple researchers rebuild, augment, and develop variations of earlier models.

Many researchers are building organizational models using multi-agent techniques. Multi-agent techniques have grown out of the work in distributed artificial intelligence. Distributed artificial intelligence intended to perform highly specific but stylized tasks such as soccer, navigation or surveillance (Bond and Gasser 1988; Gasser and Huhns 1989; Cohen 1986). Strength of this approach is the focus on representation and knowledge. For example attention is often focused on how to represent the task and knowledge about how to do the task via the agent. Another strength of this approach is a focus on decision making as search. Models are often developed to address issues of communication, coordination, planning, or problem solving, often with the intent of using these models as the brains in artificial agents. These models can explain many organizational phenomena and test the adequacy and efficiency of various definitions or representation schemes. Today, much of this work goes under the rubric of multi-agent modeling. Work in this area is beginning to focus on the role of emotions, the development of team mental models, and coordination of large numbers of agents. From an organizational theory perspective two issues stand out. First, how scalable are these models and representation schemes? That is, do the results from systems of two to five agents performing a highly stylized task generalize to larger more complex organizations? Second, when are these cognitively simple agents adequate or valid representations of human behavior?

Organizational engineering models are characterized by the extensive detail with which they represent the formal sides of organizations or tasks (organizational chart, workflow, communication paths, and rework routines) and the attention to the specific features of particular organization. These models generally focus on predicting overall

organizational or group response rather than the actions and behaviors of individual agents. These models are sufficiently detailed that they can be used to analyze potential policy changes and address what-if questions for the particular organization for which the model has been tuned (Levitt et al. 1994; Gasser and Majchrzak 1994). Model adequacy is often demonstrated by determining whether the parameters can be adjusted so that one or more important team or organizational behaviors is described at least at a qualitative level. Importantly, simply having managers work with the research team to elicit the data on the organization needed to model it often leads the manager to gain important insights into organizational problems. As such, these models are a valuable decision aid. The same is true of system models.

Social network models are characterized by representations of teams, groups, organizations, and markets in terms of the relationships among individuals or organizations. These models emphasize the structural or relational aspect of the organization and demonstrate when and how they can affect individual or organizational behaviors. Work in this field has focused on developing models of network adaptation, evolution, and change, and on developing a better understanding of how agent knowledge affects and is affected by an agent's position in the network. Network models have successfully been used to examine issues such as power and performance, information diffusion, innovation, and turnover. The adequacy of these models is determined using techniques from non-parametric statistics.

Logic models are characterized by representations of organizations and organizational processes using the techniques and formalisms of formal logic. Such models enable researchers to focus on the generative aspects of organizational form given a specific grammar (See Salanick and Leblebici 1998) and to test the consistency of extant verbal theories. These models tend to be among the most limited in their realism and the least likely to capture dynamic aspects of organizational behavior. However, these models are the only ones from which complete proofs and an exhaustive understanding of behavior can be generated. These models provide, independent of a specific machine implementation, a way of assessing the internal validity of extant theories and generating proofs about organizing behavior.

This brief review of these methodological approaches just begins to describe the vast array of modeling techniques and tools that have been used to examine organizations. These and other approaches address a variety of questions about organizations ranging from questions of design, to questions of learning, to questions of culture. As work continues in this field, researchers are beginning to employ models, which contain intellective and emulative elements. These models, for example, draw on the work in cognitive science and contribute to the work on multi-agent systems, use network representations and measures, and use logic in developing formalizations.

Models and Applications

COT models extend from simple intellective principles of general decision-making behavior (Cohen et al. 1972; Carley 1992) to representations of the decision processes and information flow within specific real-world organizations (Levitt et al. 1994; Zweben and Fox 1994). Models may even operationally specific management-decisions, or practices and polices (Gasser and Majchrzak 1992, 1994; Majchrzak and Gasser 1991, 1992). These COT models enable the researcher to examine the potential impact of general management strategies (Gasser and Majchrzak 1994; Carley and Svoboda 1996), or enable the manager to examine the organizational implications of specific management decisions (Levitt et al. 1994).

Several multipurpose computational-models of organization have been developed including well-known models such as the Garbage Can Model (Cohen et al. 1972), Plural-Soar (Carley et al. 1992), Team-Soar (Kang et al. 1998), and ORGAHEAD (Lee and Carley 2004). In a review of the state of computational modeling (Ashworth and Carley 2004, 2007), 29 specific organization theory computer simulations were found to have been introduced between 1989 and 2003; the authors also made a point that the richness of the models has also increased over those years. More recently, the CONSTRUCT model has been used extensively for theory generation and testing, notably in realms looking at the impact of communications occurring through diverse mEdia. CONSTRUCT provides a vigorous model of organization that has its

roots in symbolic interactionism (Blumer 1969), structural interactionism (Stryker 1980), and structural differentiation theory (Blau 1970). These core-theories are combined into a computational theory called constructuralism (Carley 1991), which is embodied in the CONSTRUCT model. The model recognizes that people interact within a dynamic social-based organizational network and are characteristically information-seeking agents. They interact to exchange information and purposefully may seek out others who have information that they do not yet hold. Others seeking their information, or knowledge are also seeking them out. This interaction dynamic is played out innumerable times in any organization. When this dynamic is coupled with the organization-membership changes (hiring and firing) in an organization, this emerging micro-interaction dynamic is manifested in complex organization-level dynamics and outcomes.

Computational organizational theorists often address issues of organizational design, organizational learning, and organizational adaptation. Consider the design question: organizations, through their design, are expected to be able to overcome the cognitive, physical, temporal, and institutional limitations of individual agency. Research has shown that there is no single organizational design that yields the optimal performance under all conditions, yet it has shown, that for a particular task and under particular conditions, there is a set of optimal designs. Organizational performance itself is dynamic, even under the same design (Cohen 1986). Thus, the determination of which organizational design is best depends on a plethora of factors, which interact in complex nonlinear ways to effect performance. Such factors include the task(s) being performed; intelligence, cognitive capabilities, skills, or training; available resources; quality and quantity of information; volatility of the environment; legal or political constraints on organizational design; the type of outcome desired (e.g., efficiency, effectiveness, accuracy, or minimal costs). The organization's design is considered to be capable of being intentionally changed in order to improve its performance. Consequently, computational models focused on design should be an invaluable decision aid to managers who are interested in comparing and contrasting different types of organizations. Researchers are thus providing guidelines for when to use which design, and developing computational tools for enabling managers to do just-in-time design.

Organizational learning, adaptation and change are one of the areas where COT continues to provide invaluable knowledge and understandable promise. In most organizations, multiple types of learning appear to co-exist and interact in complex ways. Organizational learning has been characterized in terms of the search for knowledge (Levinthal and March 1981), constraint based optimization (Carley and Svoboda 1996), and aggregation of individual learning (Carley 1992). In organizational learning, one major challenge is to link multiple models of organizational learning together and to see how they inform each other. It is necessary to understand how organizational networks evolve and how an evolved organizational design can be characterized as being statistically different from an initial design. Such issues of measurement are subjects of continued research within the field of COT.

Research Opportunities

The focus of COT is evolving. Past research has focused on representations of natural or human organizations. Increasingly, researchers are using COT methods to study organizations that are also composed of artificial agents, or combinations of both human and artificial agents. Human organizations, and artificial systems in general, often show intelligence and a set of capabilities that are distinct from the intelligence and capabilities of the membership within them. These systems can exhibit organization, intentional adaptation, and can display non-random and repeated patterns and processes of action, communication, knowledge, and memory regardless of whether or not the agents are human. By improving our understanding of the behavior of artificial worlds in general, researchers may discover whether there are general principles of organizing that transcend the type of agent in the organization. Artificial or virtual organizations are appearing and being used to do certain tasks such as scheduling or robotic control. One of the issues is how to structure inter-agent coordination and communications. Should organizations of humans and artificial agents be designed in the same way? Do artificial agents need to communicate the same type of information, as do

humans to be effective? Modeling the interactivity of humans and artificial agents should enable us to answer these questions.

COT will move theories of organizations beyond empirical description to predictive modeling. By focusing on the components (such as agent, structure, task, and resources), the networks of connections among these components (such as the communication structure or the resource access structure), and the processes by which they are altered (such as routines, learning, adaptation), a more dynamic and coherent view of the organization as an embedded, complex, adaptive system of human and automated agents with greater predictive ability will emerge (Carley and Prietula 1994). Attending to these factors will necessarily increase the complexity and veridicality of the models, as well as increasing the difficulty in building and validating the models. The resulting models, however, will be capable of addressing the concerns of both the theoretician and the practitioner, and yield greater predictive ability and practical guidance. COT thus has the potential to generate a better theoretical understanding of organizations, better tools for designing and reengineering organizations in real-time, and better tools for teaching people how teams, groups, and organizations function.

See

► Organization

References

Ashworth, M., & Carley, K. M. (2004). Toward unified organization theory: Perspectives on the state of computational modeling. *Proceedings of the NAACSOS 2004 Conference*, Pittsburgh, PA.

Ashworth, M., & Carley, K. M. (2007). Can tools help unify organization theory? Perspectives on the state of computational modeling. *Computational and Mathematical Organization Theory, 13*(1), 89–111.

Baligh, H. H., Burton, R. M., & Obel, B. (1990). Devising expert systems in organization theory: The organizational consultant. In M. Masuch (Ed.), *Organization, management, and expert systems*. Berlin: Walter De Gruyer.

Baum, J., & Oliver, C. (1991). Institutional linkages and organizational mortality. *Administrative Science Quarterly, 36*, 187–218.

Blau, P. M. (1970). A formal theory of differentiation in organizations. *American Sociological Review, 35*(2), 201–218.

Blumer, H. (1969). *Symbolic interactionism: Perspective and method*. Englewood Cliffs, NJ: Prentice-Hall.

Bond, A., & Gasser, L. (Eds.). (1988). *Readings in distributed artificial intelligence*. San Mateo, CA: Kaufmann.

Burt, R. (1992). *Structural holes: The social structure of competition*. Boston: Harvard University Press.

Burton, R. M., & Obel, B. (1996). Organization. In S. I. Gass & C. M. Harris (Eds.), *Encyclopedia of operations research and management science*. Norwood, MA: Kluwer Academic Publishers.

Carley, K. M. (1991). A theory of group stability. *American Sociological Review, 56*(3), 331–354.

Carley, K. M. (1992). Organizational learning and personnel turnover. *Organization Science, 3*(1), 20–46.

Carley, K. M. (1995). Computational and mathematical organization theory: Perspective and directions. *Computational and Mathematical Organization Theory, 1*(1), 39–56.

Carley, K. M., Kjaer-Hansen, J., Prietula, M., & Newell, A. (1992). Plural-soar: A prolegomenon to artificial agents and organizational behavior. In M. Masuch & M. Warglien (Eds.), *Distributed intelligence: Applications in human organizations* (pp. 87–118). Amsterdam: Elsevier Science.

Carley, K. M., & Newell, A. (1994). The nature of the social agent. *Journal of Mathematical Sociology, 19*(4), 221–262.

Carley, K. M., & Prietula, M. J. (Eds.). (1994). *Computational organization theory*. Hillsdale, IN: Lawrence Erlbaum Associates.

Carley, K. M., & Svoboda, D. M. (1996). Modeling organizational adaptation as a simulated annealing process. *Sociological Methods and Research, 25*, 138–168.

Cohen, M. D. (1986). Artificial intelligence and the dynamic performance of organizational designs. In J. G. March & R. Weissinger-Baylon (Eds.), *Ambiguity and command: Organizational perspectives on military decision making* (pp. 53–70). Marshfield, MA: Pitman.

Cohen, M. D., March, J. G., & Olsen, J. P. (1972). A garbage can model of organizational choice. *Administrative Science Quarterly, 17*, 1–25.

Cyert, R., & March, J. G. (1963). *A behavioral theory of the firm*. Englewood Cliffs, NJ: Prentice-Hall.

Decker, K. (1996). TAEMS: A framework for environment centered analysis and design of coordination mechanisms. In G. M. P. O'Hare & N. R. Jennings (Eds.), *Foundations of distributed artificial intelligence*. New York: John Wiley.

Durfee, E. H., & Montgomery, T. A. (1991). Coordination as distributed search in a hierarchical behavior space. *IEEE Transactions on Systems, Man, and Cybernetics, 21*, 1363–1378.

Galbraith, J. (1973). *Designing complex organizations*. Reading, MA: Addison-Wesley.

Gasser, L., & Huhns, M. N. (Eds.). (1989). *Distributed artificial intelligence* (Vol. 2). New York: Morgan Kaufmann.

Gasser, L., & Majchrzak, A. (1992). HITOP-A: Coordination, infrastructure, and enterprise integration. *Proceedings of the First International Conference on Enterprise Integration* (pp. 373–378). Hilton Head, SC: MIT Press.

Gasser, L., & Majchrzak, A. (1994). ACTION integrates manufacturing strategy, design, and planning. In P. Kidd & W. Karwowski (Eds.), *Ergonomics of hybrid automated systems IV* (pp. 133–136). Amsterdam: IOS Press.

Gilbert, N., & Doran, J. (Eds.). (1994). *Simulating societies: The computer simulation of social phenomena*. London: UCL Press.
Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *The American Journal of Sociology, 91*, 481–510.
Hannan, M. T., & Freeman, J. (1977). The population ecology of organizations. *The American Journal of Sociology, 82*, 929–964.
Hannan, M. T., & Freeman, J. (1989). *Organizational ecology*. Cambridge, MA: Harvard University Press.
Kang, M., Waisel, L. B., & Wallace, W. A. (1998). Team-soar: A model for team decision making. In M. Prietula, K. Carley, & L. Glasser (Eds.), *Simulating organizations: Computational models of institutions and groups* (pp. 23–45). Menlo Park, CA: AAAI Press/The MIT Press.
Kaufer, D. S., & Carley, K. M. (1993). *Communication at a distance: The effect of print on socio-cultural organization and change*. Hillsdale, IN: Lawrence Erlbaum Associates.
Krackhardt, D. (1987). Cognitive social structures. *Social Networks, 9*, 109–134.
Lee, J-S., & Carley, K. M. (2004). *OrgAhead: A computational model of organizational learning and decision making [Version 2.1.5]* (Technical Report CMU-ISRI-04-117), Carnegie Mellon University, School of Computer Science, Institute for Software Research International.
Lesser, D. D., & Corkill, D. D. (1988). Functionally accurate, cooperative distributed systems. In A. H. Bond & L. Gasser (Eds.), *Readings in distributed artificial intelligence*. San Mateo, CA: Morgan Kaufmann.
Levinthal, D., & March, J. G. (1981). A model of adaptive organizational search. *Journal of Economic Behavior and Organization, 2*, 307–333.
Levitt, R. E., Cohen, G. P., Kunz, J. C., Nass, C. I., Christiansen, T., & Jin, Y. (1994). The Virtual Design Team: Simulating how organization structure and information processing tools affect team performance. In K. M. Carley & M. J. Prietula (Eds.), *Computational organization theory* (pp. 1–18). Hillsdale, IN: Erlbaum.
Majchrzak, A., & Gasser, L. (1991). On using artificial intelligence to integrate the design of organizational and process change in US manufacturing. *Artificial Intelligence and Society, 5*, 321–338.
Majchrzak, A., & Gasser, L. (1992). HITOP-A: A tool to facilitate interdisciplinary manufacturing systems design. *International Journal of Human Factors in Manufacturing, 2*(3), 255–276.
Malone, T. W. (1986). Modeling coordination in organizations and markets. *Management Science, 33*, 1317–1332.
March, J., & Simon, H. (1958). *Organizations*. New York: John Wiley.
Nersessian, N. J. (1992). How do scientists think? Capturing the dynamics of conceptual change in science. In R. N. Giere (Ed.), *Cognitive models of science* (Vol. 15). Minneapolis, MN: Minnesota Press.
Pfeffer, J., & Salancik, G. R. (1978). *The external control of organizations: A resource dependence perspective*. New York: Harper and Row.
Powell, W. W., & DiMaggio, P. J. (1991). *The new institutionalism in organizational analysis*. Chicago: University of Chicago Press.
Prietula, M. J., Carley, K. M., & Gasser, L. (Eds.). (1998). *Simulating organizations: Computational models of institutions and groups*. Menlo Park, CA: AAAI Press/The MIT Press.
Salancik, G. R., & Pfeffer, J. (1978). A social information professing approach to job attitudes and task design. *Administrative Science Quarterly, 23*, 224–253.
Salanick, G. R., & Leblebici, H. (1998). Variety and form in organizing transactions: A generative grammar of organization. *Research in the Sociology of Organizations, 6*, 1–31.
Simon, H. A. (1947). *Administrative behavior*. New York: Free Press.
Stryker, S. (1980). *Symbolic interactionism: A social structure version*. Menlo Park, CA: Benjamin/Cummings Publishing.
Stuart, T. E., & Podolny, J. M. (1996). Local search and the evolution of technological capabilities. *Strategic Management Journal, 17*, 21–38.
Thompson, J. D. (1967). *Organizations in action*. New York: McGraw-Hill.
Waisel, L., Wallace, W. A., & Willemain, T. (1998). Using diagrammatic reasoning in mathematical modeling: The sketches of expert modelers. *Proceedings of the AAAI 1997 Fall Symposium on Reasoning with Diagrammatic Representations II*. Menlo Park, CA: AAAI Press.
Wasserman, S., & Faust, K. (1994). *Social network analysis: Methods and applications*. New York: Cambridge University Press.
Wasserman, S., & Galaskiewicz, J. (Eds.). (1994). *Advances in social network analysis: Research in the social and behavioral sciences*. Thousand Oaks, CA: Sage.
Wing, J. M. (2006). Computational thinking. *Communications of the ACM, 49*(3), 33–35.
Zhiang, L., & Carley, K. (1995). DYCORP: A computational framework for examining organizational performance under dynamic conditions. *Journal of Mathematical Sociology, 20* (2–3), 193–218.
Zweben, M., & Fox, M. S. (Eds.). (1994). *Intelligent scheduling*. San Mateo, CA: Morgan Kaufmann.

Computational Probability

Broadly defined, computational probability is the computer-based analysis of stochastic models with a special focus on algorithmic development and computational efficacy. The computer and information revolution has made it easy for stochastic modelers to build more realistic models even if they are large and seemingly complex. Computational probability is not just concerned with questions raised by the numerical computation of existing analytic solutions and the exploitation of standard probabilistic properties. It is the additional concern of the probabilist, however, to ensure that the solutions

obtained are in the best and most natural form for numerical computation. Before the advent of modern computing, much effort was directed at obtaining insight into the behavior of formal models, while avoiding computation. On the other hand, the early difficulty of computation has allowed the development of a large number of formal solutions from which limited qualitative conclusions may be drawn, and whose appropriateness for algorithmic implementation has not been seriously considered. Ease of computation has now made it feasible to have the best of all worlds: computation is now possible for classical models heretofore not completely solved, while complex algorithms can be developed for providing often needed insights on stochastic behavior.

See

- ▶ Applied Probability
- ▶ Computer Science and Operations Research Interfaces
- ▶ Matrix-Analytic Stochastic Models
- ▶ Phase-Type Probability Distributions
- ▶ Simulation of Stochastic Discrete-Event Systems

References

Drew, J., Evans, D., Glen, A., & Leemis, L. (2008). *Computational probability: Algorithms and applications in the mathematical sciences*. New York: Springer.

Computer Science and Operations Research Interfaces

John W. Chinneck[1] and Ramesh Sharda[2]
[1]Carleton University, Ottawa, Ontario, Canada
[2]Oklahoma State University, Stillwater, OK, USA

Introduction

Operations research (OR) and computer science (CS) grew up together. George Dantzig relates how his early work on linear programming (LP), the archetypal OR method, led to the funding of the development of the first electronic computers during the 1940s (Dantzig 1988, 2002). In the early years of commercial computing, a large fraction of all computing effort was devoted to linear programming. Rapid development of applications for linear programming and then for the many other OR methods followed quickly thereafter, and such developments still continue. From that same starting point, CS and computer engineering developed on parallel tracks, with the disciplines continuing to interact while developing their own separate traditions and foci of study. There has been a renewed exploration of the many areas of overlap, with the development of much improved hybrid methods for solving difficult problems.

Evidence of ongoing interest in the OR/CS interface is easy to find. There are a number of academic journals devoted to the interface, including *The INFORMS Journal on Computing, Computers and Operations Research, Computers and Industrial Engineering, Computational Optimization and Applications*, and *Mathematical Programming Computation*. The INFORMS Computing Society, a subgroup of INFORMS, the largest OR professional group, is devoted to the study of the interfaces between OR and CS.

OR can be viewed as a collection of methodologies for solving common problems related to operating organizations and designing systems. Computers are essential in using these techniques to solve problems of industrial scale. Computers carry out the numerous calculations involved in most OR methods and provide database functions to manage the very large volumes of data that are input and output. There are several important interfaces between OR and the discipline of CS; some of the main interfaces are reviewed below.

Computer Hardware

The essential interface of OR with CS and computer engineering is the computer itself. Computer hardware has seen numerous changes since the 1940s: mainframes, supercomputers, inexpensive personal computers, with additional major changes that include grids, clouds, and inexpensive multi-core machines. This has affected OR in terms of the methods used and the scale of the problems solved. OR methods have been adapted to solve extremely challenging problems of

very large scale by taking advantage of inexpensive and massively parallel computer architectures. Capabilities such as pipelining, vectorization, and superscalar computations have been employed in implementations of the simplex method, as well as interior point methods for LP. Algorithms have also been developed to exploit multiple as well as massively parallel processors: see the summaries by Zenios (1989) and Eckstein (1993). Thain et al. (2005) address distributed computing, which makes use of massively parallel computing resources that are heterogeneous and physically distributed, and subject to interruption by other uses that have higher priority. The CONDOR software used for this purpose is primarily directed toward high-throughput computing. As described in a special cluster of papers in the *INFORMS Journal on Computing* (*Volume 21, Issue 3, 2009*) on high-throughput optimization, the CONDOR software is a major enabler of large-scale optimization because it facilitates flexible access to a large pool of computers. A second major theme in the special cluster is the parallelization of tree search of various kinds.

Computers themselves, especially computer chips composed of Very Large Scale Integrated (VLSI) circuits, are extremely complex to design. There are many difficult optimization problems to solve during the design process. Some examples: What is the best way to arrange the devices on the chip to pack the maximum number of devices into the smallest area? How should the connecting wires be routed to minimize the total length of wiring? Which technologies should be employed for each of the devices? Here, not surprisingly, OR optimization techniques find many applications. The OR techniques of queueing analysis and simulation are also widely used to investigate the behavior of the chips prior to their production and the behavior of the entire computer system. For example, buffering delays related to queueing for memory or CPU access can be estimated. The survey by Chinneck et al. (2005) describes the many applications of OR in Computer-Aided Design (CAD) of VLSI chips.

Other useful developments in computer hardware that generally contribute to speedier computations also improve the speed of OR-related computations. These developments include cache memory and superscalar computation. Since the LP matrix computations involve working with sparse matrices, use of cache memory allows faster access and manipulation of matrix elements. Similarly, superscalar architectures, as well as vectorization facilities of the new computer, allow vectored calculations. LP codes such as Gurobi and IBM's CPLEX are examples of codes that have exploited the recent developments in computer architecture quite well.

OR has benefitted greatly from advances in computer design that originate in CS and computer engineering. Larger and more complex problems can now be solved. At the same time, the advances in computer hardware would likely not have been possible without the use of OR techniques in generating the designs. The fields are mutually reinforcing.

Software: Algorithms

Perhaps the widest area of overlap between OR with CS is software, particularly algorithms. While CS has a general interest in all algorithms, OR constitutes a particularly important subset of algorithms that have immediate practical applications. Interestingly, the two disciplines have often approached problems of mutual interest in completely different ways. This is particularly apparent in the field of combinatorial optimization where OR has traditionally taken a more mathematical approach while CS has taken a purely algorithmic approach as in constraint programming. The two approaches have begun to merge into a stronger hybrid. For example, concepts from constraint programming have been incorporated into branch-and-bound-based implementations of mixed-integer linear programming solvers. Hooker (2007) presents an excellent exposition of this theme.

The OR repertoire has been considerably expanded through the adoption of optimization techniques that arise from the CS algorithmic tradition, instead of the traditional OR mathematical tradition. Many of these are interesting heuristics that may not provide solution guarantees, but which can be effective in practice for certain classes of very difficult problems. Examples include Genetic Algorithms (Goldberg 1989), ant colonies, particle swarms, and other evolutionary algorithms. Partly as a consequence of exposure to these CS-originated methods, OR now develops CS-flavored methods, for example, scatter search and path relinking (Glover et al. 2000). Powell (2010) merges AI and OR to solve high-dimensional

stochastic optimization problems. Both OR and CS are keenly interested in artificial intelligence, knowledge and data management, and machine learning. Here the mixing of the two traditions is common. For example, an AI technique for robot route planning may use the solution of a number of standard OR problems as steps in a larger planning algorithm.

Software: User Interfaces

Both OR and CS face issues related to user interaction with complex objects. In OR, the objects are typically mathematical models, which may consist of functions representing the objective and the constraints in an optimization application, or the relationships representing interacting objects and their governing probability distributions in the case of simulation. This sort of modeling is a particular strength of OR because it allows the power of various algorithms and analysis techniques to be brought to bear. The problem is making a large amount of complex information and data comprehensible to the user. New graphical user interfaces (drop-down menus, hierarchically expandable-contractable structures, buttons, etc.) have been rapidly adopted in commercial OR implementations. Many graphical user interfaces are being developed to aid in formulation (Chari and Sen 1997; Androulakis and Vrahatis 1996). Jones (1998) provides an excellent overview of the use of graphics and visualization technologies in modeling and solutions.

In addition, spreadsheets have become a ubiquitous paradigm for managing models and associated data, as well as tools for delivery and presentation of results. Many spreadsheets include linear and nonlinear programming algorithms as a part of the standard function set. The spreadsheets are changing the way OR analysts prepare, manage, and deliver the models. Lijima (1996) discusses an automatic model building approach.

Software: Data Structures and Databases

New data structures developed in CS are routinely used in OR algorithms. As any serious OR algorithm developer knows very well, learning about sparse matrix approaches such as linked lists, arrays, orthogonal lists, etc. is key in implementing an algorithm. As an example, Adler et al. (1989) focus on the data structures employed in their implementation of interior point methods.

The developments in data structures and databases have helped OR in modeling and algorithmics. But OR has been a key player in designing distributed databases. OR models and their solutions are important in designs of such databases. Information storage and retrieval research has also been the beneficiary of OR algorithms for query optimization. Kraft (1985) provides a good survey of this interface between OR and CS. OR approaches (e.g., mixed-integer programming) are also used in artificial intelligence. Specific examples include the use of mixed-integer programming (MIP) in automated theorem proving. A similar example is a graph theory–based approach for partitioning knowledge bases (Srikanth 1995).

Areas of Mutual Interest

Combinatorial optimization is an area of great mutual interest for both OR and CS. Here the many algorithmic tools in both communities are brought to bear. For example, the iconic traveling salesman problem (TSP), so easy to state but so difficult to solve, has been the subject of much research in both communities. OR has used approaches such as branch-and-bound and heuristics, while CS has attempted solutions using heuristics and tree search algorithms similar to branch-and-bound. The artificial intelligence and neural network communities have also focused on solving the TSP. Other heuristics approaches, such as genetic algorithms, simulated annealing, and tabu search, are being employed by both OR and CS specialists to solve combinatorial optimization problems. Computer scientists are using logic programming to solve routing and scheduling problems. These combinatorial problems have also been the focus of much research in the OR community. The TSP belongs to the larger class of routing problems, which continue to be of great interest to both communities. Potvin (2009) outlines the many methods and combinations thereof that have been applied to routing problems by both communities.

As noted above, computer design is the subject of research in both communities. This includes hardware

design, database design, and operating system design. An example given by Greenberg (1988) includes the use of random walk theory to analyze various storage allocation approaches, an important issue in operating systems. Telecommunications systems are also vital to modern information processing and the timely delivery of OR models and solutions. Real-time data access is a key in successful implementation of many models and that is possible only because of advances in telecommunications. As complex systems in their own right, telecommunications problems are the targets of much OR research: network design and routing, location analysis, etc. Decisions in telecommunications networks are based on OR approaches using queueing theory, Markov analysis, simulations, and MIP models.

Of course, the tremendous growth of the Internet has resulted in a complete transformation of OR/MS model development, solution, and delivery. It has also profoundly impacted the OR/MS profession in terms of education and professional communication through conferences and journals. Bhargava and Krishnan (1998) discuss this important interface.

Another example of the impact of CS on OR is the field of computational probability. Researchers continue to work on developing improved numerical techniques for solving large systems of equations appearing in stochastic models (Albin and Harris 1987). Simulation research and practice has also been a beneficiary of CS advances. One example is the use of artificial intelligence techniques in design and interpretation of simulations. Advances in parallel processing have led to active research in parallel simulation to speed up the computations (Fujimoto 1993).

Concluding Remarks

The objective was to illustrate the vibrancy of the symbiosis between OR and CS. As a final example, consider these areas covered in the *INFORMS Journal on Computing*: Computational Probability and Analysis, Constraint Programming and Optimization, Design and Analysis of Algorithms, Heuristic Search and Learning, Knowledge and Data Management, Modeling: Methods and Analysis, Simulation, Telecommunications, and Electronic Commerce. These areas are also covered in CS journals. As an article in *Computer World* (Betts 1993) noted, OR/MS needs corporate data for its algorithms and needs the algorithms used in strategic information systems to make a real impact. On the other hand, information systems (IS) groups need OR to build smart applications. Betts calls the individuals with significant OR/MS and CS/IS skills the new Efficiency Einsteins, a term that indeed appropriately describes the individuals trained in this interface.

See

- ▶ Algebraic Modeling Languages for Optimization
- ▶ Artificial Intelligence
- ▶ Combinatorics
- ▶ Computational Probability
- ▶ Constraint Programming
- ▶ Heuristics
- ▶ Information Systems and Database Design in OR/MS
- ▶ Integer and Combinatorial Optimization
- ▶ Integer-programming Problem
- ▶ Knowledge Management
- ▶ Linear Programming
- ▶ Nonlinear Programming
- ▶ Parallel Computing
- ▶ Simulation of Stochastic Discrete-Event Systems
- ▶ Simulation Optimization
- ▶ Telecommunication Networks
- ▶ Traveling Salesman Problem
- ▶ Vehicle Routing
- ▶ Visualization

References

Adler, L., Karmarkar, N., Resende, M. D. G., & Beiga, G. (1989). Data structures and programming techniques for the implementation of Karmarkar's algorithm. *ORSA Journal on Computing, 1*, 84–106.

Albin, S. L. & Harris, C. M. (1987). Statistical and computational problems in probability modeling. *Annals of Operations Research*, *8/9*.

Androulakis, G. S., & Vrahatis, M. N. (1996). OPTAC: A portable software package for analyzing and comparing

optimization methods by visualization. *Journal of Computational and Applied Mathematics, 72*(1), 41–62.
Betts, M. (1993, March 22). Efficiency Einsteins. *ComputerWorld*, pp. 63–65.
Bhargava, H., & Krishnan, R. (1998). The World Wide Web: Opportunities for operations research and management science. *INFORMS Journal on Computing, 10*(4), 359–383.
Bisschop, J. J., & Fourer, R. (1996). New constructs for the description of combinatorial optimization problems in algebraic modeling languages. *Computational Optimization and Applications, 6*(1), 83–116.
Chari, K., & Sen, T. K. (1997). An integrated modeling system for structured modeling using model graphs. *INFORMS Journal on Computing, 9*(4), 397–416.
Chinneck, J. W., Nakhla, M., & Zhang, Q. J. (2005). Computer-aided design for electrical and computer engineering. In H. J. Greenberg (Ed.), *Tutorials on emerging methodologies and applications in operations research* (pp. 6-1 to 6-44). Springer Science + Business Media.
Choobineh, J. (1991). SQLMP: A data sublanguage for representation and formulation of linear mathematical models. *ORSA Journal on Computing, 3*, 358–375.
Dantzig, G. (1988, August). Impact of linear programming on computer development. *OR/MS Today*, pp. 12–17.
Dantzig, G. (2002). Linear programming. *Operations Research, 50*, 42–47.
Eckstein, J. (1993). Large-scale parallel computing, optimization, and operations research: A survey. *ORSA/CSTS Newsletter, 14*(2), 11–12, 25–28.
Fourer, R. (1983). Modeling languages versus matrix generators for linear programming. *ACM Transactions on Mathematical Software, 9*, 143–183.
Fujimoto, R. M. (1993). Parallel discrete event simulation: Will the field survive? *ORSA Journal on Computing, 5*, 213–230.
Geoffrion, A. M. (1996). Structured modeling: Survey and future research directions. *Interactive Transactions of ORMS, 1*(3).
Glover, F., Laguna, M., Marti, R., & Womer, K. (2000). Fundamentals of scatter search and path relinking. *Control and Cybernetics, 39*, 653–684.
Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Reading: Addison-Wesley Publishing.
Greenberg, H. J. (1988). Interfaces between operations research and computer science. *OR/MS Today, 15*, 5.
Greenberg, H. J. (1992). Intelligent analysis support for linear programs. *Computers and Chemical Engineering, 16*, 659–674.
Hooker, J. N. (2007). *Integrated methods for optimization*. New York: Springer.
Jones, C. V. (1998). Visualization and modeling. *Interactive Transactions of ORMS, 2*(1).
Kraft, D. H. (1985). Advances in information retrieval: Where is that/* & % record? *Advances in Computers, 24*, 277–318.
Krishnan, R. (1993). Model management: Survey, future research directions, and a bibliography. *ORSA/CSTS Newsletter, 14*(1), 1–16.
Lijima, J. (1996). Automatic model building and solving for optimization problems. *Decision Support Systems, 18* (3&4), 293–300.
Nemhauser, G. L. (1994). The age of optimization: Solving large scale real-world problems. *Operations Research, 42*, 5–13.
Potvin, J.-Y. (2009). State-of-the art review: Evolutionary algorithms for vehicle routing. *INFORMS Journal on Computing, 21*, 518–548.
Powell, W. B. (2010). Merging AI and OR to solve high-dimensional stochastic optimization problems using approximate dynamic programming. *INFORMS Journal on Computing, 22*, 2–17.
Sharda, R. (1993). *Linear and discrete optimization modeling and optimization software: An industry resource guide*. Atlanta, GA: Lionheart Publishing.
Srikanth, R. (1995). A graph theory-based approach for partitioning knowledge bases. *INFORMS Journal on Computing, 7*, 286–297.
Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: The condor experience. *Concurrency and Compution: Practice and Experience, 17*, 323–356.
Zenios, S. (1989). Parallel numerical optimization: Current status and an annotated bibliography. *ORSA Journal on Computing, 1*, 20–43.

Concave Function

A function that is never below its linear interpolation. Mathematically, a function $f(x)$ is concave over a convex set S, if for any two points, x_1 and x_2 in S and for any $0 \leq \alpha \leq 1$, $f[\alpha x_1 + (1 - \alpha)x_2] \geq \alpha f(x_1) + (1 - \alpha) f(x_2)$.

Conclusion

A portion of a rule composed of series of one or more actions that the inference engine can carry out if a rule's premise can be established to be true.

See

- ▶ Artificial Intelligence
- ▶ Expert Systems

Condition Number

- ▶ Numerical Analysis

Conditional Value-at-Risk (CVaR)

Gaia Serraino[1] and Stanislav Uryasev[2]
[1]American Optimal Decisions, Gainesville, FL, USA
[2]University of Florida, Gainesville, FL, USA

Introduction

Conditional Value-at-Risk (CVaR), introduced by Rockafellar and Uryasev (2000), is a popular tool for managing risk. CVaR approximately (or exactly, under certain conditions) equals the average of some percentage of the worst case loss scenarios. CVaR risk measure is similar to the Value-at-Risk (VaR) risk measure which is a percentile of a loss distribution. VaR is heavily used in various engineering applications, including financial ones. VaR risk constraints are equivalent to the so called chance constraints on probabilities of losses. Some risk communities prefer VaR, others prefer chance (or probabilistic) functions. There is a close correspondence between CVaR and VaR: with the same confidence level, VaR is a lower bound for CVaR. Rockafellar and Uryasev (2000, 2002) showed that CVaR is superior to VaR in optimization applications. The problem of choice between VaR and CVaR, especially in financial risk management, has been quite popular in academic literature. Reasons affecting the choice between VaR and CVaR are based on the differences in mathematical properties, stability of statistical estimation, simplicity of optimization procedures, acceptance by regulators, etc.

Definition of VaR and CVaR

Let X be a random variable with the cumulative distribution function $F_X(z) = P\{X \le z\}$. X may have meaning of loss or gain. In what follows, X has meaning of loss and this impacts the sign of functions in the definition of VaR and CVaR. Figure 1 presents the graphical representation of VaR and CVaR.

Definition 1: Value-at-Risk. Value-at-Risk (VaR) of X with confidence level $\alpha \in]0,1[$ is

$$\mathrm{VaR}_\alpha(X) = \min\{z | \boldsymbol{F_X}(z) \ge \alpha\}. \tag{1}$$

By definition, $\mathrm{VaR}_\alpha(X)$ is a lower α-percentile of the random variable X. Value-at-Risk is commonly used in many engineering areas involving uncertainties, such as military, nuclear, material, air and space, finance, etc. For instance, finance regulations like Basel I and Basel II, use VaR-deviation measuring the width of daily loss distribution of a portfolio.

For normally distributed random variables, VaR is proportional to the standard deviation. If $X \sim N(\mu, \sigma^2)$ and $F_X(z)$ is the cumulative distribution function of X, then (see Rockafellar and Uryasev 2000),

$$\mathrm{VaR}_\alpha(X) = F_X^{-1}(\alpha) = \mu + k(\alpha)\sigma, \tag{2}$$

where $k(\alpha) = \sqrt{2}\mathrm{erf}^{-1}(2\alpha - 1)$ and $\mathrm{erf}(\boldsymbol{z}) = \frac{2}{\sqrt{\pi}}\int_0^z \boldsymbol{e}^{-t^2}\boldsymbol{dt}$.

Ease and intuitiveness of VaR is counterbalanced by its mathematical properties. As a function of the confidence level, for discrete distributions $\mathrm{VaR}_\alpha(X)$ is a non-convex, discontinuous function. For discussion of numerical difficulties of VaR optimization see, for example, Rockafellar (2007), and Rockafellar and Uryasev (2000).

Definition 2: Conditional Value-at-Risk. An alternative percentile measure of risk is the Conditional Value-at-Risk (CVaR). For random variables with continuous distribution functions, $\mathrm{CVaR}_\alpha(X)$ equals the conditional expectation of X subject to $X \ge \mathrm{VaR}_\alpha(X)$. This definition is the basis for the name of Conditional Value-at-Risk. The term Conditional Value-at-Risk has been introduced by Rockafellar and Uryasev (2000). The general definition of CVaR for random variables with possibly discontinuous distribution function is as follows (see Rockafellar and Uryasev 2002).

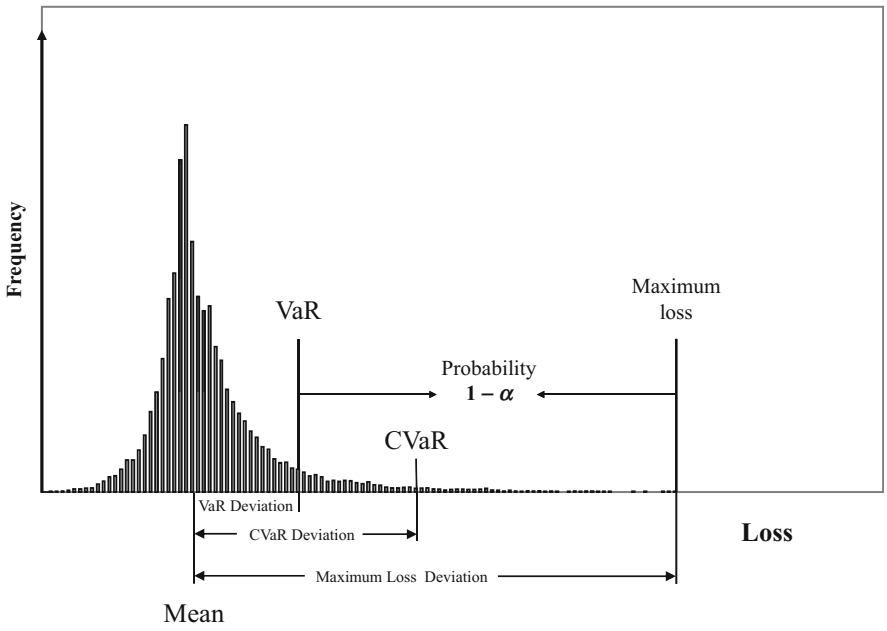
Conditional Value-at-Risk (CVaR) of X with confidence level $\alpha \in]0,1[$ is the mean of the generalized α-tail distribution:

$$\mathrm{CVaR}_\alpha(X) = \int_{-\infty}^{\infty} z dF_X^\alpha(z), \tag{3}$$

where

$$F_X^\alpha(z) = \begin{cases} 0, & \text{when } z < \mathrm{VaR}_\alpha(X), \\ \frac{F_X(z)-\alpha}{1-\alpha}, & \text{when } z \ge \mathrm{VaR}_\alpha(X). \end{cases}$$

Contrary to popular belief, in the general case, $\mathrm{CVaR}_\alpha(X)$ is not equal to an average of outcomes



Conditional Value-at-Risk (CVaR), Fig. 1 Risk Functions. Graphical Representation of VaR, VaR Deviation, CVaR, CVaR Deviation, Max Loss, Max Loss Deviation

greater than $\mathrm{VaR}_\alpha(X)$. For general distributions, one may need to split a probability atom. For example, when the distribution is modeled by scenarios, CVaR may be obtained by averaging a fractional number of scenarios. To explain this idea in more detail, alternative definitions of CVaR are presented in the following. Let $\mathrm{CVaR}^+_\alpha(X)$, called upper CVaR, be the conditional expectation of X subject to $X > \mathrm{VaR}_\alpha(X)$.

$$\mathrm{CVaR}^+_\alpha(X) = E[X \,|\, X > \mathrm{VaR}_\alpha(X)].$$

$\mathrm{CVaR}_\alpha(X)$ can be alternatively defined as the weighted average of $\mathrm{VaR}_\alpha(X)$ and $\mathrm{CVaR}^+_\alpha(X)$, as follows. If $F_X(\mathrm{VaR}_\alpha(X))<1$, so there is a chance of a loss greater than $\mathrm{VaR}_\alpha(X)$, then

$$\mathrm{CVaR}_\alpha(X) = \lambda_\alpha(X)\,\mathrm{VaR}_\alpha(X) + (1-\lambda_\alpha(X))\,\mathrm{CVaR}^+_\alpha(X), \tag{4}$$

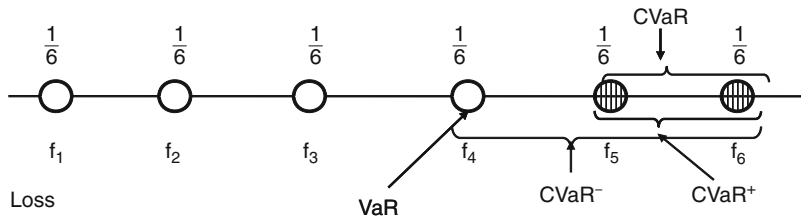
$$\text{where } \lambda_\alpha(X) = \frac{F_X(\mathrm{VaR}_\alpha(X)) - \alpha}{1-\alpha}, \tag{5}$$

whereas if $F_X(\mathrm{VaR}_\alpha(X)) = 1$, so that $\mathrm{VaR}_\alpha(X)$ is the highest loss that can occur, then

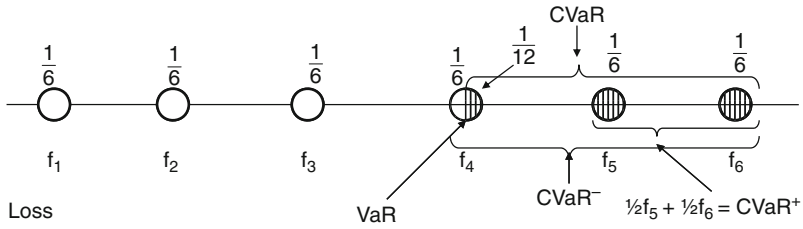
$$\mathrm{CVaR}_\alpha(x) = \mathrm{VaR}_\alpha(x)\ . \tag{6}$$

Definition (4) demonstrates that CVaR is not defined as a conditional expectation. The function $\mathrm{CVaR}^-_\alpha(X) = E[X \,|\, X \geq \mathrm{VaR}_\alpha(X)]$, called "lower CVaR", coincides with $\mathrm{CVaR}_\alpha(X)$ for continuous distributions; however, for general distributions it is discontinuous with respect to α and not convex. The construction of CVaR_α as a weighted average of VaR_α and $\mathrm{CVaR}^+_\alpha(X)$ is a major innovation. Neither VaR nor $\mathrm{CVaR}^+_\alpha(X)$ behaves well as a measure of risk for general loss distributions (both are discontinuous functions), but CVaR is a very attractive function. It is continuous with respect to α and jointly convex in (X, α). The unusual feature in the definition of CVaR is that VaR atom can be split. If $F_X(x)$ has a vertical discontinuity gap, then there is an interval of confidence level α having the same VaR. The lower and upper endpoints of that interval are $\alpha^- = F_X(\mathrm{VaR}^-_\alpha(X))$ and $\alpha^+ = F_X(\mathrm{VaR}_\alpha(X))$ where $F_X(\mathrm{VaR}^-_\alpha(X)) = P\{X < \mathrm{VaR}_\alpha(X)\}$. When $F_X(\mathrm{VaR}^-_\alpha(X)) < \alpha < F_X(\mathrm{VaR}_\alpha(X)) < 1$ the atom $\mathrm{VaR}_\alpha(X)$ having total probability $\alpha^+ - \alpha^-$ is split by the confidence level α in two pieces with probabilities $\alpha^+ - \alpha$ and $\alpha - \alpha^-$. Equation 4 highlights this splitting.

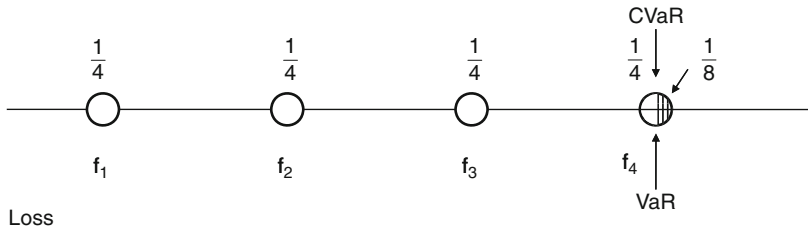
CVaR definition is illustrated further with the following examples, in which 6 equally likely scenarios have losses $f_1 \ldots f_6$. Let $\alpha = \frac{2}{3}$, see Fig. 2.



Conditional Value-at-Risk (CVaR), Fig. 2 CVaR Example 1. Computation of CVaR when α does not spilt the atom



Conditional Value-at-Risk (CVaR), Fig. 3 CVaR Example 2. Computation of CVaR when α splits the atom



Conditional Value-at-Risk (CVaR), Fig. 4 CVaR Example 3. Computation of CVaR when α splits the last atom

In this case α does not split any probability atom. Then $\mathrm{VaR}_\alpha(X) < \mathrm{CVaR}_\alpha^-(X) < \mathrm{CVaR}_\alpha(X) = \mathrm{CVaR}_\alpha^+(X)$, $\lambda_\alpha(X) = \frac{F_X(\mathrm{VaR}_\alpha(X)) - \alpha}{1-\alpha} = 0$ and $\mathrm{CVaR}_\alpha(X) = \mathrm{CVaR}_\alpha^+(X) = \frac{1}{2}f_5 + \frac{1}{2}f_6$, where f_5, f_6 are losses number five and six.

Let now $\alpha = \frac{7}{12}$, see Fig. 3. In this case α does split the $\mathrm{VaR}_\alpha(X)$ atom, $\lambda_\alpha(X) = \frac{F_X(\mathrm{VaR}_\alpha(X)) - \alpha}{1-\alpha} > 0$ and $\mathrm{CVaR}_\alpha(X)$ is given by:

$$\begin{aligned}\mathrm{CVaR}_\alpha(X) &= \frac{1}{5}\mathrm{VaR}_\alpha(X) + \frac{4}{5}\mathrm{CVaR}_\alpha^+(X) \\ &= \frac{1}{5}f_4 + \frac{2}{5}f_5 + \frac{2}{5}f_6 \ .\end{aligned}$$

In the last case, there are four equally likely scenarios and $\alpha = \frac{7}{8}$ splits the last atom; see Fig. 4. Now $\mathrm{VaR}_\alpha(X) = \mathrm{CVaR}_\alpha^-(X) = \mathrm{CVaR}_\alpha(X)$, upper CVaR, $\mathrm{CVaR}_\alpha^+(X)$ is not defined, $\lambda_\alpha(X) = \frac{F_X(\mathrm{VaR}_\alpha(X)) - \alpha}{1-\alpha} > 0$ and $\mathrm{CVaR}_\alpha(X) = \mathrm{VaR}(X) = f_4$. Portfolio Safeguard package (see American Optimal Decisions 2009), defines CVaR function for discrete distributions equivalently to (4) through the lower CVaR and upper CVaR. Suppose that $\mathrm{VaR}_\alpha(X)$ atom having total probability $\alpha^+ - \alpha^-$ is split by the confidence level α in two pieces with probabilities $\alpha^+ - \alpha$ and $\alpha - \alpha^-$. Then,

$$\begin{aligned}\mathrm{CVaR}_\alpha(X) = &\frac{\alpha^+ - \alpha}{\alpha^+ - \alpha^-}\,\frac{1-\alpha^-}{1-\alpha}\,\mathrm{CVaR}_\alpha^-(X) \\ &+ \frac{\alpha - \alpha^-}{\alpha^+ - \alpha^-}\,\frac{1-\alpha^+}{1-\alpha}\,\mathrm{CVaR}_\alpha^+(X),\end{aligned} \tag{7}$$

$$\begin{aligned}\text{where } \mathrm{CVaR}_\alpha^-(X) &= E[X \,|\, X \geq \mathrm{VaR}_\alpha(X)], \\ \mathrm{CVaR}_\alpha^+(X) &= E[X \,|\, X > \mathrm{VaR}_\alpha(X)] \,.\end{aligned} \tag{8}$$

Pflug (2000) followed a different approach and suggested to define CVaR via an optimization problem which he borrowed from Rockafellar and Uryasev (2000)

$$\begin{aligned}\mathrm{CVaR}_\alpha(X) = &\\ &\min_C \left\{ C + \frac{1}{1-\alpha} E[X - C]^+ \right\}, \\ &\text{where } [t]^+ = \max\{0, t\}.\end{aligned} \tag{9}$$

One more equivalent representation of CVaR was given by Acerbi (2002), who showed that CVaR is equal to "expected shortfall" defined by

$$\mathrm{CVaR}_{\alpha}(X) = \frac{1}{\alpha}\int_0^{\alpha} \mathrm{VaR}_{\beta}(X)d\beta\,.$$

For normally distributed random variables, CVaR deviation is proportional to the standard deviation. If $X \sim N(\mu, \sigma^2)$, then (see Rockafellar and Uryasev 2000),

$$\begin{aligned}\mathrm{CVaR}_{\alpha}(X) &= E[X \mid X \geq \mathrm{VaR}_{\alpha}(X)] \\ &= \mu + k_1(\alpha)\sigma,\end{aligned} \tag{10}$$

where

$$k_1(\alpha) = \left(\sqrt{2\pi} exp\left(\mathrm{erf}^{-1}(2\alpha - 1)\right)^2 (1-\alpha)\right)^{-1}$$

and $\mathrm{erf}(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$.

CVaR Optimization

CVaR optimization has been researched in Rockafellar and Uryasev (2000) and Uryasev (2000). Nowadays VaR has achieved the high status of being written into industry regulations (for instance, in regulations for financial companies). It is difficult to optimize VaR numerically when losses are not normally distributed. Only recently VaR optimization was included in commercial packages such as Portfolio Safeguard (see American Optimal Decisions 2009). As a tool in optimization modeling, CVaR has superior properties in many respects. CVaR optimization is consistent with VaR optimization and yield the same results for normal or elliptical distributions (see definition of elliptical distribution in (see definition of elliptical distribution in Embrechts et al. (2001)); for models with such distributions, working with VaR, CVaR or minimum variance (Markowitz 1952) is equivalent (see Rockafellar and Uryasev 2000). Most importantly, CVaR can be expressed by a minimization formula suggested by Rockafellar and Uryasev (2000). This formula can be incorporated into the optimization problem with respect to decision variables $x \in X \in \Re^n$ that are designed to minimize risk or shape it within bounds. Significant shortcuts are thereby achieved while preserving the crucial problem features like convexity. Let the random loss function $f(x, y)$ depends upon the decision vector x and a random vector y of risk factors. For instance, $f(x, y) = -(y_1x_1 + y_2x_2)$ is the negative return of a portfolio involving two instruments. Here x_1, x_2 are positions and y_1, y_2 are rates of returns of two instruments in the portfolio. The main idea in Rockafellar and Uryasev (2000) is to define a function that can be used instead of CVaR:

$$F_{\alpha}(x, \zeta) = \zeta + \frac{1}{1-\alpha}E\{[f(x, y) - \zeta]^+\}\,. \tag{11}$$

The authors proved that:

1. $F_{\alpha}(x, \zeta)$ is convex w.r.t. α,
2. $\mathrm{VaR}_{\alpha}(x)$ is a minimum point of function $F_{\alpha}(x, \zeta)$ w.r.t. ζ,
3. Minimizing $F_{\alpha}(x, \zeta)$ w.r.t. ζ gives $\mathrm{CVaR}_{\alpha}(x)$:

$$\mathrm{CVaR}_{\alpha}(x) = \min_{\alpha} F_{\alpha}(x, \zeta)\,. \tag{12}$$

In optimization problems, CVaR can enter into the objective or constraints or both. A big advantage of CVaR over VaR in that context is the preservation of convexity, i.e., if $f(x, y)$ is convex in x than $\mathrm{CVaR}_{\alpha}(x)$ is convex in x. Moreover, if $f(x, y)$ is convex in x then the function $F_{\alpha}(x, \zeta)$ is convex in both x and ζ. This convexity is very valuable because minimizing $F_{\alpha}(x, \zeta)$ over $(x, \zeta) \in X \times \Re$, results in minimizing $\mathrm{CVaR}_{\alpha}(x)$

$$\min_{x \in X} \mathrm{CVaR}_{\alpha}(x) = \min_{(x,\zeta) \in X \times \Re} F_{\alpha}(x, \zeta)\,. \tag{13}$$

In addition, if (x^*, ζ^*) minimizes F_{α} over $X \times \Re$, then not only does x^* minimize $\mathrm{CVaR}_{\alpha}(x)$ over X but also

$$\mathrm{CVaR}_{\alpha}(x^*) = F_{\alpha}(x^*, \zeta^*)\,.$$

In risk management CVaR can be utilized to "shape" the risk in an optimization model. For that purpose several confidence levels can be specified. Rockafellar and Uryasev (2000) showed that for any selection of confidence levels α_i and loss tolerances ω_i, $i = 1, \ldots, l$, the problem:

$$\begin{aligned} &\min_{x \in X} \; g(x) \\ s.\,t. \quad &\mathrm{CVaR}_{\alpha_i}(x) \leq \omega_i, \quad i = 1, \ldots, l \end{aligned} \tag{14}$$

is equivalent to the problem:

$$\min_{x,\zeta_1,..,\zeta_l,\in X\times\Re\times...\Re} g(x) \quad s.\,t. \quad F_{\alpha_i}(x,\zeta_i) \le \omega_i, \quad i=1,\ldots,l. \tag{15}$$

When X and g are convex and $f(x,y)$ is convex in x, the optimization problems (13) and (14) are ones of convex programming and thus especially favorable for computation. When Y is a discrete probability space with elements y_k, $k=1,..,N$ having probabilities p_k, $k=1,..,N$:

$$F_{\alpha_i}(x,\zeta_i) = \zeta_i + \frac{1}{1-\alpha_i} \times \sum_{k=1}^{N} p_k[\,f(x,y_k)-\zeta_i]^+ \,. \tag{16}$$

The constraint $F_\alpha(x,\zeta)\le\omega$ can be replaced by a system of inequalities by introducing additional variables η_k :

$$\eta_k \ge 0, \quad f(x,y_k)-\zeta-\eta_k \le 0, \quad k=1,..,N, \tag{17}$$

$$\zeta + \frac{1}{1-\alpha}\sum_{k=1}^{N} p_k\eta_k \le \omega \,.$$

The minimization problem in (14) can be converted into the minimization of $g(x)$ with the constraints $F_{\alpha_i}(x,\zeta_i)\le\omega_i$ being replaced as presented in (17). When f is linear in x, constraints (17) are linear.

Risk Measures

Axiomatic investigation of risk measures was suggested by Artzner et al. (1999). Rockafellar (2007) defined a functional $\mathcal{R}:\mathcal{L}^2\to]-\infty,\infty]$ as a coherent risk measure in the extended sense if:

R1: $\mathcal{R}(\mathrm{C})=\mathrm{C}$ for all constant C,

R2: $\mathcal{R}((1-\lambda)X+\lambda X')\le(1-\lambda)\mathcal{R}(X)+\lambda R(X')$ for $\lambda\in]0,1[$ (convexity),

R3: $\mathcal{R}(X)\le\mathcal{R}(X')$ when $X\le X'$ (monotonicity),

R4: $\mathcal{R}(X)\le 0$ when $||X^k-X||_2\to 0$ with $\mathcal{R}(X^k)\le 0$ (closedness).

A functional $\mathcal{R}:\mathcal{L}^2\to]-\infty,\infty]$ is called a *coherent risk measure in the basic sense* if it satisfies axioms R1, R2, R3, R4 and additionally the axiom 4

R5: $\mathcal{R}(\lambda X)=\lambda\mathcal{R}(X)$ for $\lambda>0$ (positive homogeneity).

A functional $\mathcal{R}:\mathcal{L}^2\to]-\infty,\infty]$ is called an averse risk measure in the extended sense if it satisfies axioms R1, R2, R4 and

R6: $\mathcal{R}(X)>EX$ for all nonconstant X (aversity).

Aversity has the interpretation that the risk of loss in a nonconstant random variable X cannot be acceptable, i.e. $\mathcal{R}(X)<0$, unless $EX<0$.

A functional $\mathcal{R}:\mathcal{L}^2\to]-\infty,\infty]$ is called an averse risk measure in the basic sense if it satisfies R1, R2, R4, R6 and also R5.

Examples of coherent measures of risk are $\mathcal{R}(X)=\mu X=E[X]$ or $\mathcal{R}(X)=\sup X$. However, $R(X)=\mu(X)+\lambda\sigma(X)$ for some $\lambda>0$ is not a coherent measure of risk since it does not satisfies the monotonicity axiom R3.

$R(X)=\mathrm{VaR}_\alpha(X)$ is not a coherent nor an averse risk measure. The problem lies in the convexity axiom $R2$, which is equivalent to the combination of positive homogeneity and subadditivity, this last defined as $\mathcal{R}(X+X')\le\mathcal{R}(X)+\mathcal{R}(X')$. Although positive homogeneity is obeyed, the subadditivity is violated. The lack of coherency can destroy convexity; this can still be present if the distribution of the random variable X belongs to the log-concave class, but even then there are technical hurdles because the convexity of R is missing relative to the entire space $\mathcal{L}^2$. It has been proved, for example in Acerbi and Tasche (2002), Pug (2000), Rockafellar and Uryasev (2002), that for any probability level $\alpha\in]0,1[$, $\mathcal{R}(X)=\mathrm{CVaR}_\alpha(X)$ is a coherent measure of risk in the basic sense. $\mathrm{CVaR}_\alpha(X)$ is also an averse measure of risk for $\alpha\in]0,1]$ An averse measure of risk might not be coherent; a coherent measure might not be averse.

Deviation Measures

This section refers to Rockafellar (2007) and Rockafellar et al. (2006). A functional $\mathcal{D}:\mathcal{L}^2\to[0,\infty]$ is called a deviation measure in the extended sense if it satisfies

D1: $\mathcal{D}(C)=0$ for constant C, but $D(X)>0$ for nonconstant X,

D2: $\mathcal{D}((1-\lambda)X+\lambda X')\le(1-\lambda)\mathcal{D}(X)+\lambda D(X')$ for $\lambda\in]0,1[$ (convexity),

D3: $\mathcal{D}(X) \leq d$ when $||X^k - X||_2 \to 0$ with $\mathcal{D}(X^k) \leq$ d (closedness).

A functional is called a deviation measure in the basic sense when it satisfies axioms D1, D2, D3, and furthermore

D4: $\mathcal{D}(\lambda X) = \lambda \mathcal{D}(X)$ for $\lambda > 0$ (positive homogeneity).

A deviation measure in extended or basic sense is called a coherent measure in extended or basic sense if it additionally satisfies

D5: $\mathcal{D}(X) \leq \sup X - E[X]$ for all X (upper range boundedness).

An immediate example of a deviation measure in the basic sense is the standard deviation:

$$\sigma(X) = \left(E[X - EX]^2\right)^{1/2},$$

which satisfies axioms D1, D2, D3, D4, but not D5. I.e., standard deviation is not a coherent deviation measure. Here are more examples of deviation measures in the basic sense:

Standard semideviations

$$\sigma_+(X) = \left(E[\max\{X - EX, 0\}]^2\right)^{1/2},$$

$$\sigma_-(X) = \left(E[\max\{EX - X, 0\}]^2\right)^{1/2},$$

Mean Absolute Deviation

$$MAD(X) = E[|X - EX|].$$

Moreover it is possible to define the α-Value-at-Risk deviation measure and the α-Conditional Value-at-Risk deviation measure as:

$$\mathrm{VaR}_\alpha^\Delta(X) = \mathrm{VaR}_\alpha(X - EX) \tag{18}$$

and

$$\mathrm{CVaR}_\alpha^\Delta(X) = \mathrm{CVaR}_\alpha(X - EX)\ . \tag{19}$$

VaR deviation measure $\mathrm{VaR}_\alpha^\Delta(X)$ is not a deviation measure in the general or basic sense because the convexity axiom D2 is not satisfied. CVaR deviation measure $\mathrm{CVaR}_\alpha^\Delta(X)$ is a coherent deviation measure in the basic sense.

Risk Measures Versus Deviation Measures

Rockafellar et al. originally in Rockafellar et al. (2006), and then in Rockafellar (2007) obtained the following result:

Theorem 1. *A one-to-one correspondence between deviation measures D in the extended sense and averse risk measures $\mathcal{R}$ in the extended sense is expressed by the relations*

$$\mathcal{R}(X) = \mathcal{D}(X) + EX,$$

$$\mathcal{D}(X) = \mathcal{R}(X - EX),$$

additionally,

$$\mathcal{R}\ \textit{is coherent}\ \leftrightarrow\ \mathcal{D}\ \textit{is coherent}\ .$$

Moreover the positive homogeneity is preserved:

$$\mathcal{R}\ \textit{is positively homogeneous} \leftrightarrow\ \mathcal{D}\ \textit{is positively homogeneous}\ .$$

i.e., for an averse risk measures $\mathcal{R}$ in the basic sense and a deviation measures $\mathcal{D}$ in the basic sense the one-to-one correspondence is valid, and additionally, coherent $\mathcal{R}$ $\leftrightarrow$ coherent $\mathcal{D}$.

With this theorem it is obtained that for the standard deviation, $\sigma(X)$, which is a deviation measure in the basic sense, the counterpart is the standard risk $EX + \sigma(X)$, which is a risk averse measure in the basic sense. For CVaR deviation, $\mathrm{CVaR}_\alpha^\Delta(X)$, which is a coherent deviation measure in the basic sense, the counterpart is CVaR risk, $\mathrm{CVaR}_\alpha(X)$, which is a risk averse coherent measure in the basic sense.

Another coherent deviation measure in the basic sense is the so-called Mixed Deviation CVaR, quite promising for risk management purposes. Mixed Deviation CVaR is defined as:

$$Mixed - \mathrm{CVaR}_\alpha^\Delta(X) = \sum_{k=1}^{K} \lambda_k \mathrm{CVaR}_{\alpha_k}^\Delta(X)$$

for $\lambda_k \geq 0$, $\sum_{k=1}^K \lambda_k = 1$ and α_k in $]0, 1[$. The counterpart to the Mixed Deviation CVaR is the

Mixed CVaR, which is the coherent averse risk measure in the basic sense, defined by

$$Mixed - \mathrm{CVaR}_{\alpha}(X) = \sum_{k=1}^{K} \lambda_k \mathrm{CVaR}_{\alpha_k}(X) \ .$$

Generalized Regression Problem

In linear regression a random variable Y is approximated in terms of random variables $X_1, X_2, \ldots, X_n$ by an expression $c_0 + c_1X_1 + \ldots + c_nX_n$. The coefficients are chosen by minimizing mean square error:

$$\min_{c_0, c_1, .., c_n} E(Y - [c_0 + c_1X_1 + \ldots + c_nX_n])^2 \ . \qquad (20)$$

Mean square error minimization is equivalent to minimizing standard deviation with the unbiasedness constraint (see Rockafellar et al. 2002, 2008):

$$\begin{array}{ll} \min & \sigma(Y - [c_0 + c_1X_1 + \ldots + c_nX_n]) \\ s.\,t. & E[c_0 + c_1X_1 + \ldots + c_nX_n] = EY \ . \end{array} \qquad (21)$$

Rockafellar et al. (2002, 2008) considered a general axiomatic setting for error measures and corresponding deviation measures. They defined an error measure as a functional $\mathcal{E} : \mathcal{L}^2(\Omega) \to [0, \infty]$ satisfying the axioms

E1: $\mathcal{E}(0) = 0$, $\mathcal{E}(X) > 0$ for $X \neq 0$, $\mathcal{E}(C) < \infty$ for constant C

E2: $\mathcal{E}(\lambda X) = \lambda \mathcal{E}(X)$ for $\lambda > 0$ (positive homogeneity)

E3: $\mathcal{E}(X + X') \leq \mathcal{E}(X) + \mathcal{E}(X')$ for all X and X' (subadditivity)

E4: $\{X \in \mathcal{L}^2(\Omega) | \mathcal{E}, (X) \leq c\}$ is closed for all $c < \infty$ (lower semicontinuity)

For an error measure $\mathcal{E}$ the projected deviation measure $\mathcal{D}$ is defined by the equation, $\mathcal{D}(X) = \min_C \mathcal{E}(X - C)$, and the statistic, $S(X)$, is defined by $\mathcal{S}(X) = \arg\min_C \mathcal{E}(X - C)$. Their main finding is that the general regression problem:

$$\min_{c_0, c_1, .., c_n} \mathcal{E}(Y - [c_0 + c_1X_1 + \ldots + c_nX_n]) \qquad (22)$$

is equivalent to:

$$\min_{c_1, .., c_n} \mathcal{D}(Y - [c_1X_1 + \ldots + c_nX_n])$$
$$c_0 \in \mathcal{S}(Y - [c_1X_1 + \ldots + c_nX_n]) \ .$$

The equivalence of optimization problems (20) and (21) is a special case of this theorem This leads to the identification of a link between statistical work on percentile regression (see Koenker and Bassett 1978) and CVaR deviation measure: minimization of the Koenker and Bassett error measure is equivalent to minimization of CVaR deviation. Rockafellar et al. (2008) show that when the error measure is the Koenker and Basset function: $\mathcal{E}^{\alpha}_{KB}(X) = E[\max\{0, X\} + (\alpha^{-1} - 1)\max\{0, -X\}]$ the projected measure of deviation is: $\mathcal{D}(X) = \mathrm{CVaR}^{\Delta}_{\alpha}(X) = \mathrm{CVaR}_{\alpha}(X - EX)$ with the corresponding averse measure of risk and associated statistic given by

$$\mathcal{R}(X) = \mathrm{CVaR}_{\alpha}(X),$$

$$\mathcal{S}(X) = \mathrm{VaR}_{\alpha}(X) \ .$$

Then:

$$\min_{C \in \Re} (E[X - C]_+ + (\alpha^{-1} - 1)E[X - C]_-)$$
$$= \mathrm{CVaR}^{\Delta}_{\alpha}(X),$$

$$\arg\min_{C \in \Re} (E[X - C]_+ + (\alpha^{-1} - 1)E[X - C]_-)$$
$$= \mathrm{VaR}_{\alpha}(X) \ .$$

Similar result is available for the "mixed Koenker and Bassett error measure" and the corresponding mixed deviation CVaR (see Rockafellar et al. 2008).

Comparative Analysis of VaR and CVaR

VaR is a relatively simple risk management notion. Intuition behind α-percentile of a distributions is easily understood and VaR has a clear interpretation: how much it is possible to lose with certain confidence level. VaR is a single number measuring risk, defined by some specified confidence level, e.g., $\alpha = 0.95$.

Two distributions can be ranked by comparing their VaR's for the same confidence level. Specifying VaR for all confidence levels completely defines the distribution. In this sense, VaR is superior to the standard deviation. Unlike the standard deviation, VaR focuses on a specific part of the distribution specified by the confidence level. This is what is often needed, which made VaR popular in risk management, including finance, nuclear, air and space, material science, and various military applications. One of important properties of VaR is stability of estimation procedures. Since VaR disregards the tail, it is not affected by very high tail losses, which are usually difficult to measure. VaR is estimated with parametric models, for instance Covariance-VaR based on the normal distribution assumption is very well known in finance, with simulation models such as historical or Monte Carlo or by using approximations based on second order Taylor expansion.

VaR does not account for properties of the distribution beyond the confidence level. This implies that $\mathrm{VaR}_\alpha(X)$ may increase dramatically with a small increase in α. In order to adequately estimate risk in the tail, one may need to calculate several VaRs with different confidence levels. The fact that VaR disregards the tail of the distribution may lead to unintentional bearing of high risks. In financial setting, for instance, the strategy of "naked" shorting deep out-of-the-money options will result most of the time in receiving an option premium without any loss at expiration. However, there is a chance of a big adverse market movement leading to an extremely high loss. VaR cannot capture this risk. Risk control using VaR may lead to undesirable results for skewed distributions. VaR is a non-convex and discontinuous function for discrete distributions. For instance, in financial setting, VaR is a non-convex and discontinuous function w.r.t. portfolio positions when returns have discrete distributions. This makes VaR optimization a challenging computational problem. There are codes, such as Portfolio Safeguard (PSG), that can work with VaR functions very efficiently. Portfolio Safeguard can optimize VaR performance function and also shape distributions with multiple VaR constraints. For instance, in portfolio optimization it is possible to maximize expected return with several VaR constraints at different confidence levels.

CVaR has a clear engineering interpretation. It measures outcomes which hurt the most. For example, if L is a loss then the constraint $\mathrm{CVaR}_\alpha(L) \le \bar{L}$ ensures that the average of $(1-\alpha)\%$ highest losses does not exceed $\bar{L}$. Defining $\mathrm{CVaR}_\alpha(X)$ for all confidence levels α in $(0,1)$ completely specifies the distribution of X. In this sense it is superior to standard deviation. Conditional Value-at-Risk has several attractive mathematical properties. CVaR is a coherent risk measure. $\mathrm{CVaR}_\alpha(X)$ is continuous with respect to α. CVaR of a convex combination of random variables $\mathrm{CVaR}_\alpha(w_1X_1 + \ldots + w_nX_n)$ is a convex function with respect to $(w_1, \ldots, w_n)$. In financial settings, CVaR of a portfolio is a convex function of portfolio positions. CVaR optimization can be reduced to convex programming, in some cases to linear programming (i.e. for discrete distributions).

CVaR is more sensitive than VaR to estimation errors. If there is no good model for the tail of the distribution, CVaR value may be quite misleading; accuracy of CVaR estimation is heavily affected by accuracy of tail modelling. For instance, historical scenarios often do not provide enough information about tails, hence it is necessary to assume a certain model for the tail to be calibrated on historical data. In the absence of a good tail model, one should not count on CVaR. In financial settings, equally weighted portfolios may outperform CVaR-optimal portfolios out of sample when historical data have mean reverting characteristics. VaR and CVaR measure different parts of the distribution. Depending on what is needed, one may be preferred over the other. This topic can be illustrated with financial applications of VaR and CVaR, to examine which one of these measures is better for portfolio optimization. A trader may prefer VaR to CVaR, as he may like high uncontrolled risks; VaR is not as restrictive as CVaR with the same confidence level. Nothing dramatic happens to a trader in case of high losses. He will not pay losses from his pocket; if fired, he may move to some other company. A company owner will probably prefer CVaR; he has to cover large losses if they occur, hence he "really" needs to control tail events. A board of directors of a company may prefer to provide VaR based reports to shareholders and regulators since it is less than CVaR with the same confidence level. However,

CVaR may be used internally, thus creating asymmetry of information between different parties.

In financial optimization, VaR may be better for optimizing portfolios when good models for tails are not available. VaR disregards the hardest to measure events. CVaR may not perform well out of sample when portfolio optimization is run with poorly constructed set of scenarios. Historical data may not give right predictions of future tail events because of mean-reverting characteristics of assets. High returns typically are followed by low returns, hence CVaR based on history may be quite misleading in risk estimation. If a good model of tail is available, then CVaR can be accurately estimated and CVaR should be used. CVaR has superior mathematical properties and can be easily handled in optimization and statistics. When comparing stability of estimation of VaR and CVaR, appropriate confidence levels for VaR and CVaR must be chosen, avoiding comparison of VaR and CVaR for the same level of α, as they refer to different parts of the distribution (Sarykalin et al. 2008).

References

Acerbi, C. (2002). Spectral measures of risk: A coherent representation of subjective risk aversion. *Journal of Banking and Finance, 26*, 1505–1518.

Acerbi, C., & Tasche, D. (2002). On the coherence of expected shortfall. *Journal of Banking and Finance, 26*, 1487–1503.

American Optimal Decisions. (2009). Portfolio Safeguard (PSG).

Artzner, P., Delbaen, F., Eber, J. M., & Heath, D. (1999). Coherent measures of risk. *Mathematical Finance, 9*, 203–227.

Embrechts, P., Mc Neil, A. J., & Straumann, D. (2001). Correlation and dependency in risk management: Properties and pitfalls. In M. Dempster (Ed.), *Risk management: Value at risk and beyond*. Cambridge: Cambridge University Press.

Koenker, R., & Bassett, G. W. (1978). Regression quantiles. *Econometrica, 46*, 33–50.

Markowitz, H. M. (1952). Portfolio selection. *Journal of Finance, 7*(1), 77–91.

Pflug, G. C. (2000). Some remarks on the value-at-risk and the conditional value-at-risk. In S. P. Uryasev (Ed.) *Probabilistic constrainted optimization: Methodology and applications*. (pp. 278–287). Kluwer Academic Publishers.

Rockafellar, R. T. (2007). Coherent approaches to risk in optimization under uncertainty. In INFORMS (Ed.). *Tutorials in operations research*, (pp. 38–61).

Rockafellar, R. T., & Uryasev, S. P. (2000). Optimization of conditional value-at-risk. *Journal of Risk, 2*, 21–42.

Rockafellar, R. T., & Uryasev, S. P. (2002). Conditional value-at-risk for general loss distributions. *Journal of Banking and Finance, 26*, 1443–1471.

Rockafellar, R. T., Uryasev, S., & Zabarankin, M. (2002). Deviation measures in generalized linear regression, Research Report 2002–9, ISE Dept., University of Florida.

Rockafellar, R. T., Uryasev, S., & Zabarankin, M. (2008). Risk tuning with generalized linear regression. *Mathematics of Operations Research, 33*(3), 712–729.

Rockafellar, R. T., Uryasev, S., & Zabarankin, M. (2006). Generalized deviations in risk analysis. *Finance and Stochastics, 10*, 51–74.

Sarykalin, S., Serraino, G., & Uryasev, S. (2008). Value-at-risk vs conditional value-at-risk in risk management and optimization

Uryasev, S. (2000). Conditional value-at-risk: Optimization algorithms and applications, *Financial Engineering News,* 14, February,1-5.

Cone

A set which contains the ray generated by any of its points. Mathematically, a set S is a cone if the point x in S implies that αx is in S for all $\alpha \geq 0$.

Congestion System

Often used to be synonymous with queueing system because congestion refers to the inability of arriving customers to get immediate service, which is the reason behind doing queueing analyses.

See

► Queueing Theory

Conjoint Analysis

Situations are presented to subjects, with the features of the situations varied by experimental design. The subjects are asked to state their preferences among the situations, and the importance of each feature is assessed by statistical analysis.

See

► Forecasting

Conjugate Gradient Method

► Quadratic Programming

Connected Graph

A graph (or network) in which any two distinct nodes are connected by a path.

Conservation of Flow

(1) A set of flow-balance equations governing the flow of a commodity in a network that state that the difference between the amount of flow entering and leaving a node equals the supply or demand of the commodity at the node. (2) A set of equations that state that the limiting rates that units enter and leave a state or entity of a queueing system or related random process must be equal. The entities may be service facilities (stages), where the limiting number of units coming in must equal the limiting departing; balance at a state might mean, for example, that the rate at which a queueing system goes up to n customers equals the rate at which it goes down to n from above.

See

► Balance Equations
► Markov Chains
► Network Optimization
► Queueing Theory

Constrained Optimization Problem

A problem in which a function $f(X)$ is to be optimized (minimized or maximized), where the possible solutions X lie in a defined solution subspace S, which is usually determined by a set of linear and/or nonlinear constraints.

Constraint

An equation or inequality relating the variables in an optimization problem; a restriction on the permissible values of the decision variables of a given problem.

Constraint Programming

Irvin Lustig[1] and Jean-Francois Puget[2]
[1]IBM, Somers, NY, USA
[2]IBM, Valbonne, France

Introduction

Arising from research in the computer science community, constraint programming is a technique for solving optimization problems. It often is applied to difficult combinatorial optimization problems arising in configuration, sequencing, and scheduling. To apply constraint programming, users must write software that includes both a model of an optimization problem plus an algorithmic search procedure that indicates how to search for a solution.

Background

Constraint programming is often called constraint logic programming and originates in the artificial intelligence literature in the computer science community. Here, the word programming refers to computer programming. Knuth (1968) defines a computer program as "an expression of a computational method in a computer language." A computer program can be viewed as a plan of action of operations of the computer, and, hence, the common concept of a plan is shared with the origins of linear programming. With respect to constraint programming, it is a computer programming technique, with a name that is in the spirit of other programming techniques such as object-oriented programming, functional programming, and structured programming. Van Hentenryck (1999) wrote:

> The essence of constraint programming is a two-level architecture integrating a constraint and a programming component. The constraint component provides the basic

operations of the architecture and consists of a system reasoning about fundamental properties of constraint systems such as satisfiability and entailment. The constraint component is often called the constraint store, by analogy to the memory store of traditional programming languages. Operating around the constraint store is a programming-language component that specifies how to combine the basic operations, often in non-deterministic ways.

Hence, a constraint program is not a statement of a problem as in mathematical programming, but is rather a computer program that indicates a method for solving a particular problem. It is important to emphasize the two-level architecture of a constraint programming system. Because it is first and foremost a computer programming system, the system contains representations of programming variables, which are representations of memory cells in a computer that can be manipulated within the system. The first level of the constraint programming architecture allows users to state constraints over these programming variables. The second level of this architecture allows users to write a computer program that indicates how the variables should be modified so as to find values of the variables that satisfy the constraints.

The roots of constraint programming can be traced back to the work on constraint satisfaction problems in the 1970s, with the advent of arc consistency techniques (Mackworth 1977) on the one hand, and the language ALICE (Lauriere 1978) that was designed for stating and solving combinatorial problem on the other hand. In the 1980s, work in the logic programming community showed that the PROLOG language could be extended by replacing the fundamental logic programming algorithms with more powerful constraint solving algorithms. For instance, in 1980, PROLOG II used a constraint solver to solve equations and disequations on terms. This idea was further generalized in the constraint logic programming scheme and implemented in several languages (Colmerauer 1990; Jaffar and Lassez 1987; Van Hentenryck 1989). Van Hentenryck (1989) used the arc-consistency techniques developed in the constraint satisfaction problem (CSP) framework as the algorithm for the basic constraint solving. This was termed finite domain constraints.

In the 1990s, a rich area of research in constraint programming was the development of special purpose programming languages to allow people to apply the techniques of constraint programming to different classes of problems. Constraint logic programming was first proposed in the context of the programming language PROLOG, and there are many other specialized languages that have been developed that offer extended functionalities compared to traditional constraint logic programming systems. Some of these are implemented as libraries in mainstream languages, such as ILOG Solver C++ (Puget 1994) or Lisp (Puget 1992). Some others are special purpose languages, such as Oz (Smolka 1995) and Claire (Caseau and Laburthe 1995).

In the design of such languages, an axiom of their development is that they provide completeness with respect to being languages for doing computer programming. A recent innovative approach with respect to languages for constraint programming is in the design of the Optimization Programming Language (OPL) (Van Hentenryck 1999), where the language was designed with the purpose of making it easy to solve optimization problems by supporting constraint programming and mathematical programming techniques. Here, the completeness of the language for computer programming is not important. Instead, the language is designed to support the representation of optimization problems and includes the facilities to use an underlying constraint programming engine, with the ability to program a search strategy to find solutions to problems. The OPL language is not a complete programming language, but rather a language that is designed to solve optimization problems using either constraint programming or mathematical programming techniques. An advantage of OPL is that the same language is used to unify the representations of decision variables from traditional mathematical programming with programming variables from traditional constraint programming.

Constraint Satisfaction Problems

To understand the constraint programming framework, a formal definition of a constraint satisfaction problem is given next using the terminology of mathematical programming. Given a set of n decision variables $x_1, x_2, \ldots, x_n$, the set D_j of allowable values for each decision variable x_j, $j = 1, \ldots, n$, is called the

domain of the variable x_j. The domain of a decision variable can be any possible set, operating over any possible set of symbols. For example, the domain of a variable could be the even integers between 0 and 100, or the set of real numbers in the interval [1,100], or a set of people {Tom, John, Jim, Jack}. There is no restriction on the type of each decision variable, and, thus, decision variables can take on integer values, real values, set elements, or even subsets of sets.

Formally, a constraint $c(x_1, x_2,\ldots, x_n)$ is a mathematical relation, that is, a subset S of the set $D_1 \times D_2 \times \ldots \times D_n$, such that if $(x_1, x_2,\ldots,x_n) \varepsilon S$, then the constraint is said to be satisfied. Alternatively, a constraint can be defined as a mathematical function f: $D_1 \times D_2 \times \ldots D_n \rightarrow \{0,1\}$ such that $f(x_1, x_2,\ldots, x_n) = 1$ if and only if $c(x_1, x_2,\ldots, x_n)$ is satisfied. Using this functional notation, a constraint satisfaction problem (CSP) can be defined as follows:

Given n domains $D_1, D_2,\ldots, D_n$ and m constraints $f_1, f_2,\ldots, f_m$, find $x_1, x_2,\ldots, x_n$ such that

$$f_k(x_1, x_2, \ldots, x_n) = 1, \quad 1 \le k \le m$$
$$x_j \varepsilon D_j, \quad 1 \le j \le n$$

Note that this problem is only a feasibility problem, and that no objective function is defined. It is important to note here that the functions f_k do not necessarily have closed mathematical forms and can simply be defined by providing the set S described above. A solution to a CSP is simply a set of values of the variables such that the values are in the domains of the variables, and all of the constraints are satisfied.

Algorithms for Constraint Satisfaction

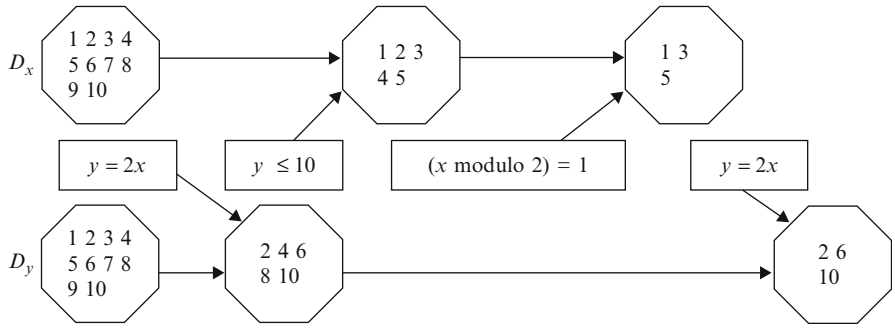
Up to now, there has been no discussion about the algorithm that a constraint programming system uses to determine solutions to constraint satisfaction problems. As mentioned earlier, a constraint programming system requires that the user programs a search strategy that indicates how the values of the variables should change so as to find values that satisfy the constraints. In OPL, there is a default search strategy that is used if the user does not program a search strategy. Most constraint programming systems require the user to program a search strategy. The first fundamental algorithm underlying a constraint programming system is given next, followed by a discussion of the methodologies used to program search.

Constraint Propagation and Domain Reduction

A constraint is defined as a mathematical function $f(x_1, x_2,\ldots, x_n)$ of the variables. Because constraint programming has its roots in computer programming, the variables can be viewed as programming language variables within a computer programming environment. Within this environment, assume there is an underlying mechanism that allows the domains of the variables to be maintained and updated. When a variable's domain is modified, the effects of this modification are then propagated to any constraint that interacts with that variable. For each constraint, a domain reduction algorithm is then programmed that modifies the domains of all the variables in that constraint, given the modification of one of the variables in that constraint. The domain reduction algorithm for a particular kind of constraint discovers inconsistencies among the domains of the variables in that constraint by removing values from the domains of the variables. If a particular variable's domain becomes empty, then it can be determined that the constraint cannot be satisfied, and an earlier choice can be undone.

This is best illustrated by the example in Fig. 1. Consider two variables x and y, where the domains of each variable are given as $D_x = \{1, 2, 3, 4, \ldots, 10\}$ and $D_y = \{1, 2, 3, 4, \ldots, 10\}$, and the single constraint $y = 2x$. For the variable y and this constraint, it is clear that y must be even and the domain of y can be changed to $D_y = \{2, 4, 6, 8, 10\}$. Now, considering the variable x, since $y \le 10$, it then follows that $x \le 5$, and the domain of x can be changed to $D_x = \{1, 2, 3, 4, 5\}$. Suppose that now a constraint is added of the form x (modulo 2) = 1. This is equivalent to the statement that x is odd. This reduces the domain of x to be $D_x = \{1, 3, 5\}$. Now, reconsidering the original constraint $y = 2x$, the values of 4 and 8 can be removed from the domain of y and obtain $D_y = \{2, 6, 10\}$.

A typical constraint programming system allows the programmer to take advantage of the existing



Constraint Programming, Fig. 1 Illustration of constraint propagation and domain reduction

propagators for built-in constraints that cause domain reductions, and to build one's own propagation and domain reduction schemes for user-defined constraints. Some systems, however, for example OPL built on top of ILOG Solver (ILOG 1999), are robust enough that large libraries of predefined constraints are provided as part of the constraint programming system, along with associated propagation and domain reduction algorithms, and it is often not necessary to create new constraints with specialized propagation and domain reduction algorithms.

Given a set of variables with their domains and a set of constraints on those variables, a constraint programming system will apply the constraint propagation and domain reduction algorithm in an iterative fashion to make the domains of each variable as small as possible, while making the entire system arc consistent. Given a constraint f_k as stated above and a variable x_j, a value $d \varepsilon D_j$ is consistent with f_k if there is at least one assignment of the variables such that $x_j = d$ and $f_k = 1$ with respect to that assignment. A constraint is then arc consistent if all of the values of all the variables involved in the constraint are consistent. A constraint system is arc consistent if all of the corresponding constraints are arc consistent. The term arc is used because the first CSPs were problems with constraints stated on pairs of variables, and this system could be viewed as a graph, with nodes corresponding to the variables and arcs corresponding to the constraints.

A number of algorithms have been developed to efficiently propagate constraints and reduce domains so as to create systems that are arc consistent. The predominant algorithm is called AC-5, developed by Van Hentenryck et al. (1992). This latter article unified the directions of the constraint satisfaction community and the logic programming community by introducing the concept of developing different algorithms for different constraints as implementations of the basic constraint propagation and domain reduction principle.

Programming Search

Given a CSP, the constraint propagation/domain reduction algorithm can be applied to reduce the domains of the variables so as to arrive at an arc consistent system. However, while this may determine if the CSP is infeasible, it does not necessarily find solutions of a CSP. To do this, one must program a search strategy. Traditionally, the search facilities provided by a constraint programming system have been based on depth-first search. The root node of the search tree contains the initial values of the variables. At each node, the user programs a *goal*, which is a strategy that breaks the problem into two (or more) parts, and decides which part should be evaluated first. A simple strategy might be to pick a variable and to try to set that variable to the different values in the variable's domain. This strategy creates a set of leaves in the search tree and creates what is called a *choice point*, with each leaf corresponding to a specific choice. The goal also orders the leaves among themselves within the choice point. In the next level of the tree, the results of the choice made at the leaf are propagated, and the domains are reduced locally in that part of the tree. This will either produce a smaller arc consistent system, or a proof that the choice made for this leaf is not possible. In this case, the system automatically backtracks to the parent and tries other leaves of that parent. The search, thus, proceeds in a depth-first

manner, until at a node low in the tree a solution is found, or until the entire tree is explored, in which case the CSP is found to be infeasible. The search strategy is enumerative, and, at each node, constraint propagation and domain reduction are used to help prune the search space.

A recent innovation in constraint programming systems is found in ILOG Solver 4.4 (ILOG 1999), where the idea of allowing the programmer to use other strategies beyond depth-first search is provided. Depth-first search has traditionally been used because in the context of computer programming, the issues regarding memory management are dramatically simplified. ILOG Solver 4.4 allows the programmer to use best first search (Nilsson 1971), limited discrepancy search (Harvey and Ginsberg 1995), depth-bounded discrepancy search (Walsh 1997), and interleaved depth-first search (Meseguer 1997). In ILOG Solver, the basic idea is that the user programs *node evaluators*, *search selectors*, and *search limits*. Node evaluators contain code that looks at each open node in the search tree and chooses one to explore next. Search selectors order the different choices within a node, and search limits allow the user to terminate the search after some global limit is reached (e.g., time, node count, etc.). With these basic constructs in place, it is then possible to easily program any search strategy that systematically searches the entire search space by choosing nodes to explore (i.e., programming node evaluators), dividing the search space at nodes (i.e., programming goals and creating choice points), and picking the choice to evaluate next within a specific node (i.e., programming search selectors). Constraint programming systems provide a framework for describing enumeration strategies for solving search problems in combinatorial optimization.

Constraint Programming and Branch and Bound

For those familiar with integer programming, the concept of search strategies should seem familiar. In fact, branch and bound, which is an enumerative search strategy, has been used to solve integer programs since the middle 1960s. Lawler and Wood (1966) present a survey, while the text by Garfinkel and Nemhauser (1972) describes branch and bound in the context of an enumerative procedure. In systems that have been developed for integer programming, users are often given the option of selecting a variable selection strategy and a node selection strategy. These are clearly equivalent to the descriptions of search electors and node evaluators described above.

There are two fundamental ways in which a constraint programming framework extends the basic branch and bound procedures. First, in a branch and bound procedure, two branches are created at each node after a variable x with a fractional value v has been chosen to branch on. The search space is then divided into two parts, by creating a choice point based on the two choices of $(x = \lfloor v \rfloor)$ and $(x \geq \lceil v \rceil)$. In the constraint programming framework, the choices that are created can be any set of constraints that divides the search space. For example, given two integer variables x_1 and x_2, a choice point could be created consisting of the three choices $(x_1 < x_2)$, $(x_1 > x_2)$, and $(x_1 = x_2)$.

The second way in which a constraint programming framework extends the basic branch and bound procedures is with respect to the variable selection strategy. In most branch and bound implementations, the variable selection strategy uses no knowledge about the problem to make the choice of variable to branch on. The integer program is treated in its matrix form, and different heuristics are used to choose the variable to branch on based on the solution of the linear programming relaxation that is solved at each node. In a constraint programming approach, the user specifies the branching strategy in terms of the formulation of the problem. Because a constraint program is a computer program, the decision variables of the problem can be treated as computer programming variables, and a strategy is programmed in the context of the problem formulation. Hence, to effectively apply constraint programming techniques, one uses problem-specific knowledge to help guide the search strategy so as to efficiently find a solution. In this way, a constraint programming system, when combined with a linear programming optimizer, can be viewed as a framework that allows users to program problem-specific branch and bound search strategies for solving mixed-integer programming problems. This capability has been available since 1996 by combining the products ILOG Solver for constraint programming and ILOG Planner for linear programming. Similar concepts also appeared in PROLOG III (Colmerauer 1990), CLP(R) (Jaffar and Lassez 1987), and CHIP (Dincbas et al. 1988).

Optimization in Constraint Programming

A constraint satisfaction problem was defined as a feasibility problem. With regard to optimization, constraint programming systems allow an objective function to be specified. Notationally, the objective function will be denoted as g: $D_1 \times D_2 \times \ldots \times D_n \rightarrow \Re$, so that at any feasible point to the CSP, the function $g(x_1, x_2, \ldots, x_n)$ can be evaluated, with the objective function to be minimized. A weakness of a constraint programming approach is that there is not necessarily a lower bound present when minimizing an objective function. This is unlike integer programming, where a lower bound exists due to the linear programming relaxation of the problem. Constraint programming systems offer two methods for optimizing problems, called standard and dichotomic search.

Standard and Dichotomic Search

The standard search procedure used is to first find a feasible solution to the CSP, while ignoring the objective function $g(x_1, x_2, v, x_n)$. Let $y_1, y_2, \ldots, y_n$ represent such a feasible point. The search space can then be pruned by adding the constraint $g(y_1, y_2, \ldots, y_n) > g(x_1, x_2, \ldots, x_n)$ to the system, and continuing the search. The constraint that is added specifies that any new feasible point must have a better objective value than the current point. As the search progresses, new points will have progressively better objective values. The procedure concludes until no feasible point is found. When this happens, the last feasible point can be taken as the optimal solution.

Dichotomic search depends on having a good lower bound L on the objective function $g(x_1, x_2, \ldots, x_n)$. Before optimizing the objective function, an initial feasible point is found, that determines an upper bound U on the objective function. A dichotomic search procedure is essentially a binary search on the objective function. The midpoint $M = (U + L)/2$ of the two bounds is computed, and a CSP is solved by taking the original constraints and adding the constraint $g(x_1, x_2, \ldots, x_n) < M$. If a new feasible point is found, then the upper bound is updated, and the search continues in the same way with a new midpoint M. If the system is found to be infeasible, then the lower bound is updated, and the search again continues with a new midpoint M. Dichotomic search is effective when the lower bound is strong, because the computation time to prove that a CSP is infeasible can often be large. The use of dichotomic search in cooperation with a linear programming solver might be effective if the linear programming representation can provide a good lower bound.

See

- Artificial Intelligence
- Branch and Bound
- Integer and Combinatorial Optimization
- Linear Programming

References

Abdennadher, S., & Frühwirth, T. (2003). *Essentials of constraint programming*. Heidelberg: Springer.

Apt, K. (2003). *Principles of constraint programming*. Cambridge, UK: University of Cambridge Press.

Caseau, Y. & Laburthe, F. (1995). *The Claire documentation* (LIENS report 96–15), *Ecole Normale Superieure*, Paris.

Colmerauer, A. (1990). An introduction to PROLOG III. *Communications of the ACM, 33*(7), 70–90.

Dincbas, M., Van Hentenryck, P., Simonis, H., Aggoun, A., Graf, T., & Berthier, F. (1988). The constraint logic programming language CHIP. *Proceedings of the International Conference on fifth generation computer systems*, Tokyo.

Garfinkel, R. S., & Nemhauser, G. L. (1972). *Integer programming*. New York: Wiley.

Harvey, W. D. & Ginsberg, M. L. (1995). Limited discrepancy search. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Vol. 1. pp. 607–613.

ILOG (1999). *ILOG solver 4.4 users manual*. Gentilly, France: ILOG.

Jaffar, J., & Lassez, J.-L. (1987). Constraint logic programming. In *Conference Record of the Fourteenth Annual ACM Symposium on principles of programming languages*, Munich, pp. 111–119.

Knuth, D. E. (1968). *Fundamental algorithms, the art of computer programming* (2nd ed., Vol. 1). Reading, MA: Addison-Wesley.

Lauriere, J.-L. (1978). A language and a program for stating and solving combinatorial problems. *Artificial Intelligence, 10*, 29–127.

Lawler, E. L., & Wood, D. E. (1966). Branch-and-bound methods: A survey. *Operations Research, 14*, 699–719.

Mackworth, A. K. (1977). Consistency in networks of relations. *Artificial Intelligence, 8*, 99–118.

Meseguer, P. (1997). Interleaved depth-first search. In *Proceedings of the International Joint Conference on artificial intelligence (IJCAI)*, Vol. 2. pp. 1382–1387.

Nilsson, N. J. (1971). *Problem solving methods in artificial intelligence*. New York: McGraw-Hill.
Puget, J.-F. (1992). Pecos: A high level constraint programming language. *Proceedings of the 1st Singapore International Conference on intelligent systems*.
Puget, J.-F. (1994). A C++ implementation of CLP. *Proceedings of the 2nd Singapore International Conference on intelligent systems*. See also the current web site http://www.ilog.com/products/optimization/research/spicis94.pdf
Rossi, F., van Beek, P., & Walsh, T. (Eds.). (2006). *Handbook of constraint programming*. New York: Elsevier.
Smolka, G. (1995). The Oz programming model. In J. van Leeuwen (Ed.), *Computer science today*. Lecture notes in computer science (Vol. 1000, pp. 324–343). Springer-Verlag.
Van Hentenryck, P. (1989). *Constraint satisfaction in logic programming*. Cambridge, MA: MIT Press.
Van Hentenryck, P. (1999). *The OPL optimization programming language*. Cambridge, MA: MIT Press.
Van Hentenryck, P., Deville, Y., & Teng, C. M. (1992). A generic arc-consistency algorithm and its specializations. *Artificial Intelligence, 57*, 291.
Walsh, T. (1997). Depth-bounded discrepancy search. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, Vol. 2. pp. 1388–1395.

Constraint Qualification

A condition imposed on the constraints of an optimization problem so that local minimum points will satisfy the Karush-Kuhn-Tucker conditions.

See

- ▶ Karush-Kuhn-Tucker (KKT) Conditions
- ▶ Nonlinear Programming

Construction Applications

Chad Perry
Queensland University of Technology,
Brisbane, Australia

Introduction

Due to their size and complexity, most construction projects would appear to offer a wide potential for MS/OR applications. For example, the standard critical path models of PERT, CPM and precedence diagrams are particularly successful in construction. However, apart from these models, MS/OR methods and models are not often used in construction. Schelle (1990, p. 111) summarizes, "In project management the large number of publications about operations research topics contrast to the small number of real applications."

This entry reviews three major areas of construction where MS/OR applications could occur — job estimation and tendering, project planning, and project management and control. Factors inhibiting the application of MS/OR in construction projects are discussed and possible future developments are canvassed.

Job Estimation and Tendering

Some MS/OR models have been applied to job estimation. Job estimation requires trade-offs between time and cost. Early MS/OR work assumed direct costs for each activity increased linearly with time, and therefore, used linear programming. But construction usually does not fit this assumption. Dynamic programming and integer linear programming have also been used, but the large number of variables and constraints of construction projects made them unworkable. Models based on heuristic and nonlinear curves have been found to be almost as accurate and more friendly for construction managers, and have been tried (Cusack 1985). In addition, the Line of Balance (LOB) model, originally developed for the U.S. Navy, is used to make trade-offs between alternative schedules, and a modified LOB model called Time Chainage is used in the U.K. for estimating schedules for construction of roads, bridges and other civil engineering projects (Wager and Pittard 1991).

Allied to job estimation is tendering, which must consider competitors' likely actions along with the bidder's decisions. It is a relatively more open and therefore more difficult system to model. Hence, although ARIMA and regression, plus other statistical and simulation models, have been developed to assist tendering, they have rarely been applied.

If tendering is considered from the selector's point of view, rather than a bidder's point of view, variables are not so uncertain because the selector will have certain information about all the bids.

Nevertheless, the complexity of construction projects again makes application of conventional MS/OR models difficult, especially as prior knowledge about bidders is an important choice factor. A hybrid model using linear programming, multiattribute utility, regression and expert systems seems appropriate here (Russell 1992).

Project Planning

While preparing a tender, construction managers must start to plan the project in more detail. The critical path models, integrated with cost control and reporting models, are widely used in construction for this purpose (Wager and Pittard 1991). Their application in complex construction projects has suggested theoretical extensions, for example, incorporating the stochastic relationship of cost with time. One such extension for the complex construction industry is a suite of PC programs, Construction Project Simulator (CPS), which incorporates productivity variability and external interferences to the construction process on site. It then produces bar charts, cost and resource schedules like the critical path models (Bennett and Ormerod 1984). However, most of these extensions have unrealistic data requirements and are rarely applied even if they are tried.

Modeling could be especially useful in planning tunnel construction projects. For example, Touran and Toshiyuki (1987) demonstrated a simulation model for tunnel construction and design. But model use is limited to very large projects.

Project planning usually involves more than cost minimization with constraints, for example, environmental considerations. Some MS/OR multi-objective models have provided assistance here. For example, Scott (1987) applied multi-objective valuation to roads construction, using a step-by-step procedure to evaluate all objectives, without having to assume all quantified data as being equally accurate and reliable.

Management and Control

After a project is planned, it must be managed and controlled. Linked with the project plan are straightforward accounting models. With increasing use of real time reporting, they allow closer management of costs. It is in this relatively stable field of managing and controlling the project after it has begun that conventional MS/OR models offer most promise, that is, at a tactical and relatively deterministic and repetitive level. For example, standard cost-minimization models could be applied to the management of construction equipment, to location and stocking of spare parts ware-houses, and to selecting material handling methods. In one of few actual MS/OR applications, Perry and Iliffe (1983) used a transhipment model to manage movement of sand during an airport construction project. Two other possible areas in where MS/OR models might be applied are multiple projects (where several projects are designed and built somewhat concurrently to minimize costs), and marketing.

In summary, although potential applications of MS/OR in construction appear at first glance to be plentiful, progress with actual MS/OR applications is slow. One reason for this is that risks in using unproven MS/OR models are high in commercial operations where claims resulting from mistakes can be taken to court. Moreover, each construction appears to be one-off, that is, the building is more or less different than previous ones of the constructor: at a different site with different subsurface conditions; involving different organizations and individuals with different goals; different weather; different material, labor requirements and shortages; different errors in estimates of time and cost; and different levels of interference from outside. Given this lack of standardization, MS/OR modeling has tended to move towards more general simulation models (which have large data requirements) or heuristic models. Still, MS/OR applications are few and although "computers are installed extensively throughout ... consultants and construction site offices ... their role appears to make the former manual processes more efficient rather than exploit the increased potential brought by the machine" (Brandon 1990, p. 285).

What does the future hold for MS/OR applications in the construction industry? A probable development is their increasing use in conjunction with user-friendly software on PCs. Research in the construction industry suggests that the key to successful implementation of research is a powerful intermediary like

construction managers. Developments in PC-based software such as simulations and expert systems, which assist rather than replace the experience-based knowledge of people like site managers, offer promise of more MS/OR applications, especially in the complex and expensive field of contractual disputes. These possibilities will be enhanced by interactive, three-dimensional graphical interfaces. In particular, expert systems should be used more frequently because they incorporate the existing knowledge of construction managers.

See

- ▶ Bidding Models
- ▶ CPM
- ▶ Engineering Applications
- ▶ Expert Systems
- ▶ Gantt Charts
- ▶ Linear Programming
- ▶ Multiobjective Programming
- ▶ PERT
- ▶ Project Management

References

Bennett, J., & Ormerod, R. N. (1984). Simulation applied to construction projects. *Construction Management and Economics, 2*, 225–263.

Brandon, P. S. (1990). The development of an expert system for the strategic planning of construction projects. *Construction Management and Economics, 8*, 285–300.

Cusack, M. M. (1985). A simplified approach to the planning and control of cost and project duration. *Construction Management and Economics, 3*, 183–198.

Gupta, V., Fisher, D., & Murtaza, M. (1996). A consortium sponsored knowledge-based system for managerial decision making in industrial construction. *Interfaces, 26*, 9–23, November/December.

Lewis, J. (2005). *Project planning, scheduling, & control: A hands-on guide to bringing projects in on time and on budget* (4th ed.). New York: McGraw-Hill Osborne Media.

Perry, C., & Iliffe, M. (1983). Earthmoving on construction sites. *Interfaces, 13*(1), 79–84.

Russell, J. S. (1992). Decision models for analysis and evaluation of construction contractors. *Construction Management and Economics, 10*, 185–202.

Schelle, H. (1990). Operations research and project management past, present and future. In H. Schelle & H. Reschke (Eds.), *Dimensions of project management*. Berlin: Springer-Verlag.

Scott, D. (1987). Multi-objective economic evaluation of minor roading projects. *Construction Management and Economics, 5*, 169–181.

Slowinski, R., & Weglarz, R. (Eds.). (1989). *Advances in project scheduling* (Studies in production and engineering economics, 9). Amsterdam: Elsevier.

Touran, A., & Toshiyuki, A. (1987). Simulation of tunnelling operations. *Construction Engineering and Management, 113*, 554–568.

Wager, D. M., & Pittard, S. J. (1991). *Using computers in project management*. Cambridge, UK: Construction Industry Computing Association.

Continuous-Time Markov Chain (CTMC)

A Markov process with a continuous parameter but countable state space. The stochastic process $\{X(t)\}$ has the property that, for all $s, t \geq 0$ and nonnegative integers i, j, and $x(u)$, $0 \leq u < \infty$,

$$\Pr\{X(t+s) = j | X(s) = i,\ X(u) = x(u),\ 0 \leq u < s\}$$

$$= \Pr\{X(t+s) = j | X(s) = i\}.$$

See

- ▶ Markov Chains
- ▶ Markov Processes

Control Charts

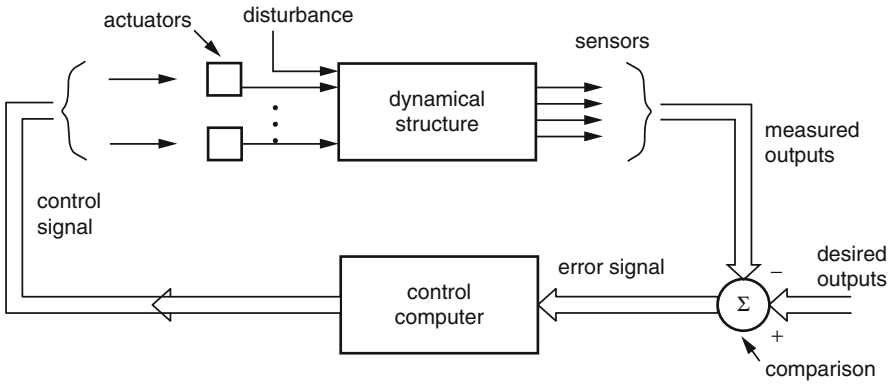
- ▶ Quality Control

Control Theory

Andre Z. Manitius
George Mason University, Fairfax, VA, USA

Introduction

Although the use of control theory is normally associated with applications in electrical and mechanical engineering, it shares much of its mathematical foundations with operations research and management science. These foundations include



Control Theory, Fig. 1 Closed loop multivariate system

differential and difference equations, stochastic processes, optimization, calculus of variations, and others.

In application, control theory is concerned with steering dynamical systems to achieve desired results. Both types of systems to be controlled and the goals of control include a wide variety of cases. Control theory is strongly related to control systems engineering, which is fundamental to many advanced technologies. In a broader sense, control theoretic concepts are applicable not just to technological systems, but also to dynamical systems encountered in biomedical, economic and social sciences. Control theory has also had a fundamental impact on many areas of applied mathematics and continues to be a rich source of research problems.

Systems to be controlled may be of various forms: they could be mechanical, electrical, chemical, thermal or other systems that exhibit dynamical behavior. Control of such systems requires that the system dynamics be well understood. This is usually accomplished by formulating and analyzing a mathematical model of the system. Physical properties of the system play an important role in establishing the mathematical model. However, once the model is established, the control theoretic considerations are independent of the exact physical nature of the system. Since different physical systems often have similar mathematical models, similar control principles are applicable to them. For example, a mechanical system of interconnected masses and springs is described by the same mathematical model as an electrical circuit of interconnected capacitors and inductors. From the control theoretic point of view, the two systems can be treated in the same way.

The control of a system is usually accomplished by providing an input signal which affects the system behavior. Physically, the input signal often changes the energy flow in the system, much like the pilot's commands change the thrust of the engines in the aircraft. The conversion of input signals into physical variables, such as the energy of the mass flow, is done by devices called actuators. System response is measured by various instruments, called sensors. The measurements, called output signals, are fed to a controller, which usually means a control computer. The controller determines the successive values of the input signals that are then passed on to the actuators. While the control computer hardware is the physical location where the control decisions are being made, the essence of the control is a control algorithm imbedded in the computer software. The development of control algorithms is often based on sophisticated mathematical theory of control and on specific models of systems under control, Fig. 1.

One of the key difficulties of control is the uncertainty about the system model and system outputs. The uncertainty has several origins. Mathematical models of systems under control are based on many simplifying assumptions and thus contain errors due to approximations. Properties or parameters of the system may change in unpredictable ways. Systems may be subject to unknown external inputs, such as, gusts of winds acting on the aircraft. Output signals provided by sensors contain sensor noise or communication channel noise. By its very nature, the control problem formulation usually includes uncertain parameters and signals. The task of control theory is to provide solutions which guarantee, whenever possible, good system performance in spite of the uncertainties.

Historical Development

The first systematic study of feedback control of steam engines by J. C. Maxwell appeared in 1868. In 1893, A.M. Lyapunov published a first paper on the stability of motion, but his work made an impact on the control theory literature only 55 years later. When the first electronic amplifiers appeared in the long-distance telephone lines after World War I, high-gain feedback coupled with high-order dynamics of amplifiers led to stability problems. In 1932, H. Nyquist provided a method of feedback stability analysis based on the frequency response. In the late 1930s, devices for controlling aircraft were introduced. World War II gave a big boost to the field of feedback control. Norbert Wiener's theory of filtering of stochastic processes, combined with the servomechanism theory, provided a unified framework for the design of control mechanisms in aircraft and ships and became what is known as classical control theory.

In the late 1950s and in the 1960s, an extensive development of control theory took place, coinciding with manned space flight and other aerospace applications, and with the advent of computers. Bellman's principle of optimality embedded in Dynamic Programming, Pontryagin's Maximum Principle of Optimal Control, and the Kalman Filter, were invented between 1956 and 1960. State-space methods of analysis, based on differential equations and matrix computations, have become the main tools of what was then named modern control theory. Control theory played a crucial role in the success of the Apollo moon-landing project in 1969. In the 1970s, substantial progress was made in the control of systems governed by partial differential equations, adaptive control and nonlinear control. The applications of control theory became very diverse, including complex material processing, bio-medical problems, and economic studies. In the 1980s, robust control theory was formulated and reached a significant level of maturity. Robust control theory has by now provided a synthesis of the classical and the modern (state-space based) control theory.

In general, research in control aims at studying the limits of performance of feedback control systems in some advanced applications. Computational tools of control have been coded in MATLAB software system and in similar software. Control hardware has been revolutionized by microprocessors and new sensor and actuator technologies, such as smart materials. Some tools of the intelligent control approach have been applied to on-board guidance and navigation systems. Anti-lock brake systems, computerized car engine control, and geographic positioning systems are a few examples of systems where the principles and tools of control theory are at work.

Mathematical Models for Control and the Identification Issue

The most commonly used mathematical control is the linear state-space model. This is a system of first-order, time-invariant, linear differential equations with inputs and outputs. Such a linear system can be written as:

$$\begin{cases} \frac{d}{dt}x(t) = \boldsymbol{A}x(t) + \boldsymbol{B}u(t) \\ y(t) = \boldsymbol{C}x(t) + \boldsymbol{D}u(t) \end{cases}$$

where $\boldsymbol{x}(t) =$ state vector, $\boldsymbol{u}(t) =$ control, $\boldsymbol{y}(t) =$ output, and $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{D}$ are matrices of appropriate dimensions.

In practical applications, engineers often use scalar or matrix transfer functions. These are rational functions of the complex variable s arising in the Laplace transform or the variable z from the Z-transform, the latter being used for discrete-time systems. There are close relationships between state-space and transfer-function models.

In the past, many other systems have been analyzed in the control theory literature, such as nonlinear ordinary differential systems, differential equations with delay, integro-differential equations, linear and nonlinear partial differential equations, stochastic differential equations, both ordinary and partial, semigroup theory, discrete-event systems, queueing systems, Markov chains, Petri nets, neural network models, and others. In many cases, research on those systems has resulted in precise mathematical conditions under which the main paradigms of linear system theory extend to those systems.

Given an existing physical system, one of the most challenging tasks is the determination of the mathematical model for control. This is usually done in one of two ways: either the model equations are derived from physical laws and the few unknown parameters are estimated from input and output data,

or a general model form is assumed with all the parameters unknown, which then requires a more extensive parameter estimation and model validation procedure. In either case, the overall step of model determination from experimental data is called system identification. Well developed methods and computer algorithms exist to assist the control designer in this task.

The Main Ideas

Feedback is a scheme in which the control of the system is based on a concurrent measurement of the system's output. Usually, the system output is being compared to a given desired value of the output and the control is adjusted so as to steer the system output closer to the desired value. Feedback creates a directed loop linking the output to the input.

Complicated systems may have many feedback loops, either nested or intersecting one another. Feedback results in a change of a system's internal dynamics and system's input–output characteristics. A system with a properly designed feedback is capable of responding correctly to input commands even in the presence of uncertainties about the model and the external perturbations. An effective feedback reduces the effects of uncertainties, regardless of their origin. Feedback is also being used to improve stability margins, eliminate or attenuate some undesirable nonlinearities, or to shape system's bandwidth. Some systems cannot even function in a stable way without feedback. An example is the fly-by-wire fighter jet in which the feedback control loop keeps the aircraft in a stable flight envelope. The mechanism of feedback is well understood in case of linear systems. However, feedback mechanisms in nonlinear systems, especially those with many degrees of freedom, remain the subject of continued investigations. In a broader sense, the concept of feedback may be used to interpret various closed-loop interactions taking place in dynamical systems in physics, biology, economics, etc. (e.g., see Franklin et al. 2006; SIAM 1988).

Optimal Control

In many cases, the goal of control may be mathematically formulated as the optimization of a certain performance measure. The tools of optimization theory and calculus of variations have been applied to derive certain optimal control principles. For example, one of the fundamental results valid for a broad class of linear systems with a quadratic performance measure says that the optimal control is accomplished by a linear feedback based on the measurement of the internal state vector of the system. Parameters of that linear feedback are obtained by solving a quadratic equation called the Riccati equation. Another fundamental result says that the control of linear systems with bounded control function and the transition time as a performance measure is accomplished by using only the extremum values of control (a bang-bang control). Solution of optimal control problems often requires iterative numerical computations to find a control that yields the best performance.

Robust Control

Control methods have been developed to design feedback that minimizes the effect of uncertainty. Systems of this type are called robust. For example, one can design a feedback which minimizes the norm of the transfer function from unwanted disturbances to the output. Another design of that type makes the feedback system maximally insensitive to parameter variations. One of the key ideas in robust control is the use of norms in the Hardy function space H_∞, for both signals and operators (transfer functions). A close connection between the minimum H_∞ norm solutions and the solutions of certain systems of matrix Riccati equations has been discovered.

Robust control theory is well understood for linear time invariant systems, and some results have been obtained for nonlinear systems. A link has been discovered between the game-theoretic approach to control problems with uncertainty and the linear and nonlinear robust control.

Stochastic Control

Stochastic control theory involves the study of control and recursive estimation problems in which the uncertainty is modeled by random processes. One of the most significant achievements of the linear theory

was the discovery of Kalman filtering algorithms and the separation principle of the optimal stochastic control. The principle states that under certain conditions the solution of the optimal stochastic control problem combines the optimal deterministic state feedback and the optimal filter estimate of the state vector, which are obtained separately from each other.

For nonlinear systems, Markov diffusions have become the tool of analysis. Stochastic optimal control conditions lead to certain nonlinear second order partial differential equations which may have no smooth solutions satisfying appropriate initial and boundary conditions. Weak solutions and viscosity solutions have recently been used to describe solutions to such optimal control problems.

Adaptive Control

One possible remedy against the uncertainty about the system and external signals is the use of adaptive feedback mechanism. During the system operation under a regular feedback, input and output signals can be processed to produce increasingly accurate estimates of system parameters which in turn can be used to adjust the regular feedback loop. Alternatively, the step of estimating the original system can be bypassed in favor of a direct tuning of the feedback controller to minimize the error. The control system built this way contains two feedback loops, one regular but with adjustable parameters, and one that provides the adjustment mechanism. Adaptive systems are inherently nonlinear.

The main theoretical issue is the question of stability of the adaptive feedback loop; stable adaptive feedback laws for certain classes of nonlinear systems have been discovered. In contrast, bursting phenomena, oscillations, and chaos have also been found in certain simple adaptive systems. Research efforts include the finding of robust adaptive control laws and at solving stochastic adaptive control problems for systems governed by some partial differential equations.

Intelligent Control

The term intelligent control is meant to describe control which includes decision making in uncertain environments, learning, self-organization, evolution of the control laws based on adaptation to new data, and to changes in the environment. An intelligent controller may deal with situations that require deciding which variables should be controlled, which models should be used, and which control strategy should be applied at any particular stage of operation. In some situations, no precise mathematical model of the system may exist, with the only information about the process being descriptive.

Intelligent control is a blend of control theory with artificial intelligence. In contrast to mathematical control theory, that uses precisely formulated models and control laws, intelligent control relies in many cases on heuristic models and rules. It is an area of research with few established paradigms. Tools of intelligent control includes expert systems, fuzzy set theory and fuzzy control algorithms, and artificial neural networks. Examples of systems where the intelligent control may become effective are autonomous robots and vehicles, flexible manufacturing systems, and traffic control systems.

Concluding Remarks

Among the main control theory challenges are: feedback control laws for nonlinear systems with many degrees of freedom, including systems governed by nonlinear partial differential equations (e.g., control of fluid flow); adaptive and robust control of such systems; control of systems based on incomplete models with learning and intelligent decision making; and feedback mechanisms based on vision and other non-traditional sensory data (SIAM 1988).

See

- ▶ Artificial Intelligence
- ▶ Calculus of Variations
- ▶ Dynamic Programming
- ▶ Mathematical Programming
- ▶ Neural Networks
- ▶ Nonlinear Programming
- ▶ Unconstrained Optimization

References

Anderson, B. D. O., & Moore, J. B. (1990). *Optimal control*. Englewood Cliffs, NJ: Prentice Hall.

Astrom, K., & Wittenmark, B. (1989). *Adaptive control*. Reading, MA: Addison-Wesley.

Fleming, W., & Soner, M. (1994). *Controlled markov processes and viscosity solutions*. New York: Springer Verlag.

Franklin, G. F., Powell, J. D., & Emami-Naeni, A. (2006). *Feedback control of dynamic systems* (5th ed.). New Jersey: Pearson Prentice Hall.

Green, M., & Limebeer, D. J. N. (1995). *Linear robust control*. Englewood Cliffs, NJ: Prentice Hall.

Lin, C. F. (1994). *Advanced control systems design*. Englewood Cliffs, NJ: Prentice Hall.

SIAM. (1988). *Future directions in control theory: A mathematical perspective, SIAM reports on issues in the mathematical*. Philadelphia: Sciences.

Sontag, E. D. (1998). *Mathematical control theory: Deterministic finite dimensional systems* (2nd ed.). New York: Springer Verlag.

Stoorvogel, A. (1992). *The H_∞ control problem*. London: Prentice Hall International.

Zabczyk, J. (1992). *Mathematical control theory: An introduction*. Boston: Birkhauser.

Control Variates

In stochastic or Monte Carlo simulation, a variance reduction technique whereby a simulated random variable with known expectation (the control variate) is used to construct a more precise estimator by combining it (usually linearly) with another more standard estimator.

See

- ▶ Monte Carlo Methods
- ▶ Monte Carlo Simulation
- ▶ Simulation of Stochastic Discrete-Event Systems
- ▶ Variance Reduction Techniques in Monte Carlo Methods

Controllable Variables

In a decision problem, variables whose values are determined by the decision process and/or decision maker. Such variables are also called decision variables.

See

- ▶ Decision Maker (DM)
- ▶ Decision Problem
- ▶ Mathematical Model

Convex Combination

A weighted average of points (vectors). A convex combination of the points $x_1, \ldots, x_k$ is a point of the form $x = \alpha_1 x_1 + \ldots \alpha_k x_k$, where $\alpha_1 \geq 0, \ldots, \alpha_k \geq 0$, and $\alpha_1 + \ldots \alpha_k = 1$.

Convex Cone

A cone that is also a convex set.

Convex Function

A function that is never above its linear interpolation. Mathematically, a function $f(x)$ is a convex over a convex set S, if or any two points x_1 and x_2 in S and for any $0 \leq \alpha \leq 1$,

$$f[\alpha x_1 + (1-\alpha)x_2] \leq \alpha f(x_1) + (1-\alpha) f(x_2).$$

Convex Hull

The smallest convex set containing a given set of points S. The convex hull of a given set S is the intersection of all convex sets containing S. The convex hull of a given set of points S is the set of all convex combinations of set of points from S. If the set S is a finite set of points in a finite-dimensional space, then the convex hull is a polyhedron.

See

- ▶ Convex Set
- ▶ Polyhedron

Convex Optimization

Yurii Nesterov
Université Catholique de Louvain (UCL),
Louvain-la-Neuve, Belgium

Introduction

Optimization problems arise naturally in many domains of Operations Research. Usually they come from some design or planning procedures facing to the limits on different resources (budget, raw materials, labor, time, etc.). The required amounts of these resources become the decision variables. It is convenient to represent them by a vector $x \in R^n$.

Very often, the results of the planning procedure can be characterized by certain functions $f_i(x)$, $i = 0, \ldots, m$, called the functional components of the problem. Choosing the most important characteristic, say f_0, as the objective function, the following is the standard formulation of the constrained optimization problem:

$$f^* = \min_{x \in Q} \{ f_0(x) : \; a_i \le f_i(x) \le b_i, i = 1, \ldots, m \}, \tag{1}$$

where $Q \subseteq R^n$ is a basic feasible set, and $[a_i, b_i] \subset R$ are the target intervals for different characteristics of the decision variables.

Clearly, the formulation (1) is very natural and very important for many areas of human activity. However, it can be shown that in its general form (1) this problem is numerically unsolvable. It is one of the most important results of the *Complexity Theory* for optimization problems, developed in Nemirovski and Yudin (1983). This theory studies the abilities of numerical methods in computing an approximate solution of optimization problems. Since the possibility of computing an exact solution is extremely rare in Nonlinear Numerical Analysis, the methods are treated as iterative procedures, which generate an answer by collecting some information on the particular problem instance.

The computational efforts of such methods are measured by the number of calls of oracle, the special unit which can compute the values and differential characteristics of functional components f_i at the requested point $x \in R^n$. It is assumed that the oracle is a Black Box, meaning that no information on its structure and intermediate computational results is available. At the same time, no bounds are introduced for the computational cost of iteration and for the volume of required memory. Nevertheless, it appears that the worst-case complexity bound for generating an approximate global solution to problem (1) with accuracy $\varepsilon > 0$ is of the order

$$O\left(\frac{1}{\varepsilon^n}\right) \tag{2}$$

calls of oracle. In this lower bound, ε represents the accuracy in estimating the optimal value of the objective function, and functional components of (1) are assumed to be Lipschitz continuous. Worst-case bound means that for each method from a reasonably wide class of optimization procedures there exists a very bad problem from our problem class, for which the number of calls is at least (2). Note that this bound destroys any hope for developing a reliable method for approximating a global solution to the general problem (1). Indeed, taking very moderate values for the parameters, say $\varepsilon = 0.01$ and $n = 20$, we get a computational cost which is intractable for any computer now and in a foreseen future.

Note that the main reason for the disastrous bound (2) is the ambitious intention to approach a global solution of the general problem (1). By stepping back and reducing the goal to finding a stationary point, for the unconstrained minimization problem

$$f^* = \min_{x \in R^n} f(x), \tag{3}$$

the stationary points satisfy the Fermat condition

$$\nabla f(x) = 0, \tag{4}$$

where ∇f denotes the gradient of function f. Thus, the goal now is to find $\bar{x} \in R^n$ with

$$\| \nabla f(\bar{x}) \| \le \varepsilon. \tag{5}$$

For function with Lipschitz continuous gradient,

$$\| \nabla f(x) - \nabla f(y) \| \le L \| x - y \|, \quad x, y \in R^n, \tag{6}$$

apply the simplest Gradient Method,

$$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k), \quad k \geq 0. \tag{7}$$

Then, after k iterations of the scheme,

$$\min_{0\leq i\leq k} \| \nabla f(x_i)\|^2 \leq \frac{L(f(x_0) - f^*)}{2(k+1)}.$$

Thus, the goal (5) can be reached in

$$O\left(\frac{1}{\varepsilon^2}\right) \tag{8}$$

iterations of method (7). As compared with (2), the estimate (8) does not depend on n. Thus, even for very large problems the goal (5) is reachable.

Since we are able to approach efficiently the stationary points of problem (3), the natural question is as follows:

What is the largest class of functions, for which the stationarity condition (4) *is a suffcient characterization of the global soution to* (3)*?*

Denoting this class of functions by $\mathcal{F}$, we could ask also for two other natural properties:

- If $f_i \in \mathcal{F}$ and $\alpha_i \geq 0$, then $\sum\limits_i \alpha_i f_i \in \mathcal{F}$.
- Any linear function belongs to $\mathcal{F}$.

Then, it can be shown, Section 2.1.1 in Nesterov (2004), that a differentiable function f belongs to $\mathcal{F}$ if and only if

$$f(y) \geq f(x) + \langle \nabla f(x), y - x\rangle, \quad x, y \in R^n, \tag{9}$$

where $\langle x, y\rangle \stackrel{\text{def}}{=} \sum\limits_{i=1}^{n} x^{(i)} y^{(i)}$. This is a definition of differentiable convex function on R^n.

This notion can be extended onto nondifferentiable functions defined onto the convex sets. A set $Q \subset R^n$ is called convex if

$$\begin{aligned} &x, y \in Q \quad \Rightarrow \\ &\quad x_\alpha \stackrel{\text{def}}{=} \alpha x + (1-\alpha) y \in Q, \ \alpha \in [0,1]. \end{aligned} \tag{10}$$

Function f is called convex if its epigraph is a convex set. That is

$$\begin{aligned} f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y), \\ x, y \in Q, \ \alpha \in [0,1]. \end{aligned} \tag{11}$$

Function f is called concave if $-f$ is convex.

Despite to the absence of usual differentiability and continuity, convex function possesses many interesting properties, especially at the interior points of its domain. At these points it is locally Lipschitz continuous and differentiable along any direction. Moreover, at any point x of its domain $\mathrm{dom} f$, there exists a special *set* of differential characteristics of this function called a subdifferential $\partial f(x) \subset R^n$. It is defined as follows:

$$f(y) \geq f(x) + \langle g_x, y - x\rangle, \quad x, y \in \mathrm{dom} f, \ g_x \in \partial f(x). \tag{12}$$

Subdifferential is a closed convex set, which is bounded for any interior point of $\mathrm{dom} f$. For differentiable functions, $\partial f(x) \equiv \{\nabla f(x)\}$. Convexity of functions is preserved by some natural operations (summation, multiplication by a positive constant, taking a maximum, etc.). All these operations are supported by corresponding operations with subdifferentials. Thus, in principal, the differential characteristics of convex functions are computable. Convex sets and convex functions are extensively studied in a special mathematical discipline called Convex Analysis, see Rockafellar (1970); Hiriart-Urruty and Lemarechal (1993).

The notion of convexity plays a central role in Operations Research and Optimization Theory. Using convex objects, we can write down the convex optimization problem:

$$f^* = \min_{x\in Q} \{ f_0(x) : \ f_i(x) \leq 0, \ i = 1, \ldots, m \}, \tag{13}$$

where Q is a closed convex set and all functions f_i, $i = 0, \ldots, m$, are convex. We will see below that this problem is generically tractable. It can be efficiently solved by different optimization methods.

Besides the convex optimization problems, there are two other important problem classes with convex structure.

Saddle point problems. In this setting, we need to find a solution of the following problem:

$$\min_{x\in Q_x} \max_{y\in Q_y} f(x, y), \tag{14}$$

where Q_x and Q_y are closed convex sets, and function $f(x, y)$ is convex in x and concave in y. For example,

we can write in this form a two-person zero-sum game. Note that optimization problem (13) is a particular case of problem (14):

$$f^* = \min_{x\in Q}\max_{y\in R^m_+}\left\{f_0(x) + \sum_{i=1}^m y^{(i)} f_i(x)\right\}.$$

For the saddle point problem (14), define a pair of primal-dual problems. Denote

$$f(x) = \max_{y\in Q_y} f(x,y), \quad \phi(y) = \min_{x\in Q_x} f(x,y).$$

Note that $f(x) \geq \phi(y)$ for all $x \in Q_x$, $y \in Q_y$. At the same time, under very mild assumptions there is zero duality gap:

$$f^* = \min_{x\in Q_x} f(x) = \max_{y\in Q_y} \phi(y).$$

Variational inequalities. Variational inequality problem (VI) is posed as follows:

$$\text{Find } x^* \in Q : \langle V(x^*), x - x^*\rangle \geq 0 \quad \forall x \in Q, \quad (15)$$

where Q is a closed convex set and $V : R^n \to R^n$. Point x^* is called the strong solution of VI. If $x_* \in Q$ and

$$\langle V(x), x - x_*\rangle \geq 0 \quad \forall x \in Q, \quad (16)$$

then x_* is called the weak solution of VI. For continuous operators, the sets of weak and strong solutions coincide. Not that the numerical schemes can approach only the set of weak solutions X_*, independently on existence of the strong ones. By definition, X_* is a closed convex set (may be empty).

The problem (15) has convex structure when the operator V is monotone:

$$\langle V(x) - V(y), x - y\rangle \geq 0 \quad \forall x, y \in Q. \quad (17)$$

Variational inequality problem with monotone operator is the most general (and most difficult) problem with convex structure. It includes, as a particular case, the saddle point problem (14).

An important example of monotone VI is the problem of finding the Nash equilibrium of a game with m players. Let $Q_i \subseteq R^{n_i}$, $i = 1, \ldots, m$, be the closed convex sets containing the feasible decision vectors of corresponding players. Assume that each player i has his own utility function $f_i(x_1, \ldots, x_m)$, which is convex in $x_i \in Q_i$, and jointly concave in all other variables $x_j \in Q_j$, $j \neq i$. The Nash equilibrium $x^* = (x_1^*, \ldots, x_m^*)$ is defined as follows:

$$x_i^* = \arg\min_{x_i\in Q_i} f_i(x_1^*, \ldots, x_i, \ldots, x_m^*), \quad i = 1, \ldots, m.$$

It can be shown that this point is a solution of corresponding VI with operator

$$V(x) = (\nabla_{x_1} f_1(x), \ldots, \nabla_{x_m} f_m(x)),$$
$$x = (x_1, \ldots, x_m) \in \prod_{i=1}^m Q_i.$$

This operator is monotone if function $\sum_{i=1}^m f_i(x)$ is convex.

From the modelling point of view, convexity is often a very natural property. Condition (10) implies that with two feasible decisions x and y, all intermediate variants x_α are feasible. Clearly, this assumption enormously facilitates the decision-making process. It is realized for a long time already, that even if the number of variables is relatively small, the problems with nonconvex or discrete feasible sets can be extremely difficult for human beings, e.g., "To be, or not to be?" Shakespeare (1602). For numerical methods, convexity is also a very favorable property.

Black-Box Optimization Methods

Nonsmooth Optimization

For explaining the main ideas of Black-Box optimization schemes, consider the simplest formulation of convex optimization problem,

$$\min_{x\in Q} f(x), \quad (18)$$

where $Q \subseteq R^n$ is a closed convex set, and f is a convex function defined on R^n. Black-box optimization methods approach the optimal solution of this problem by analyzing the answers of the oracle $(f(x_i), g_i \in \partial f(x_i))$ computed at the test points $\{x_i\}_{i=0}^\infty$. The simplest optimization strategy is

implemented in the (primal) Subgradient Method (Polyak 1967; Shor 1985):

$$x_{k+1} = \pi_Q(x_k - h_k g_k), \quad k \geq 0, \tag{19}$$

where $\pi_Q(x)$ is the Euclidean projection of x onto the convex set Q, and the apriori chosen step sized $\{h_k\}$ satisfy conditions

$$h_k > 0, \quad h_k \to 0, \quad \sum_{k=0}^{\infty} h_k = \infty.$$

Assuming that the subgradients of f are bounded on Q by constant M, and that $\| x_0 - x^* \| \leq R$, we can derive the optimal step size strategy for N-step process:

$$h_k = \frac{R}{M\sqrt{N+1}}, \quad k = 0, \ldots, N.$$

Then $\min\limits_{0 \leq k \leq N} f(x_k) - f^* \leq \frac{MR}{\sqrt{N+1}}$. Thus, in order to compute ε-solution of our problem, we need $\frac{M^2R^2}{\varepsilon^2}$ of calls of oracle. In accordance with Complexity Theory (Nemirovski and Yudin 1983), this efficiency estimate cannot be improved by the Black Box Methods working in a high-dimensional spaces (number of iterations never exceeds the dimension).

Unfortunately, the practical performance of the subgradient method coincides with its theoretical estimate, which is quite pessimistic. Therefore it is important to have numerical schemes which can accelerate on the particular problem instances. Note that by the Black Box Concept and inequality (12), after analyzing N test points, the full knowledge about the objective function is concentrated in the following inequality:

$$f(y) \geq f_N(x) \stackrel{\text{def}}{=} \max_{0 \leq i \leq N} [\, f(x_i) + \langle g_i, x - x_i \rangle], \quad x \in R^n. \tag{20}$$

Piece-wise linear function f_N is called the full model of our problem. It gives, for example, a computable lower bound for the optimal value of our problem:

$$f^* \geq f_N^* \stackrel{\text{def}}{=} \min_{x \in Q} f_N(x).$$

Note that $\hat{f}_N^* \stackrel{\text{def}}{=} \min\limits_{0 \leq k \leq N} f(x_k)$ gives us an upper bound for f^*. The models $f_N(x)$ are employed in so-called Bundle Methods, see Hiriart-Urruty and Lemarechal (1993). The most popular variant is the Level Method (Lemarechal et al. 1995):

$$Q_k = \{x \in Q : \; f_k(x) \leq \tfrac{1}{2}(\hat{f}_k^* + f_k^*)\},$$

$$x_{k+1} = \pi_{Q_k}(x_k), \quad k \geq 0.$$

The efficiency estimate for this method is the same as for the subragient scheme (the optimal one). However, its practical behavior usually is much better. Level Method can be also used for solving saddle point problems and variational inequalities.

For problem of moderate dimension, Complexity Theory provides us with lower complexity bound $O(n \ln \frac{1}{\varepsilon})$. Note that it has very weak dependence on ε. The methods which efficiency estimates depend polynomially on dimension and the logarithm of accuracy are called polynomial-time schemes.

In optimization, the methods which approach the above lower bound are based on idea of cutting planes. In accordance with (12), the optimal solution x^* satisfies the following condition:

$$\langle g_x, x - x^* \rangle \geq 0.$$

Therefore, after N iterations we know that

$$x^* \in \mathcal{L}_N \stackrel{\text{def}}{=} \{x \in Q : \; \langle g_k, x_k - x \rangle \geq 0, \; k = 0, \ldots, N\}.$$

The localization sets $\mathcal{L}_k$ can be used in different ways. The most straightforward strategy is implemented in the Method of Centers of Gravity (Newman 1960, Levin 1965):

$$x_{k+1} = \text{center\_of\_gravity}(\mathcal{L}_k), \quad k \geq 0.$$

It can be shown that this method has the optimal rate of convergence. However, its iteration is extremely expensive. An implementable version of this method is the famous Ellipsoid Method (Nemirovski and Yudin, 1983). It updates the outer

ellipsoidal approximations for the sets $\mathcal{L}_k$. For problem (18), the scheme is very simple:

$$\textbf{Initial settings}: R \geq \| x_0 - x^* \|, H_0 = R^2 \cdot I.$$

$$k\textbf{th iteration}: x_{k+1} = x_k - \frac{H_k g_k}{(n+1)\langle H_k g_k, g_k \rangle^{1/2}},$$

$$H_{k+1} = \frac{n^2}{n^2-1}\left(H_k - \frac{2H_k g_k g_k^T H_k}{(n+1)\langle H_k g_k, g_k \rangle}\right).$$

However, its efficiency estimate is not optimal: $O(n^2 \ln \frac{1}{\varepsilon})$. At this moment there exist several optimization methods with optimal efficiency estimate and reasonably small complexity of each iteration, see Nesterov (2004).

Smooth Optimization

Assume now that the objective function in problem (18) has Lipschitz-continuous gradient:

$$\| \nabla f(x) - \nabla f(y) \| \leq L \| x - y \|, \quad x, y \in Q. \quad (21)$$

The simplest scheme for solving this problem is the Primal Gradient Method:

$$x_{k+1} = \pi_Q\left(x_k - \frac{1}{L}\nabla f(x_k)\right), \quad k \geq 0.$$

Its rate of convergence is as follows: $f(x_k) - f^* \leq \frac{LR^2}{k+1}$. Thus, this scheme can compute ϵ-solution in $O(\frac{LR^2}{\varepsilon})$ iteration. Another possibility is to use Dual Gradient Method:

$$v_{k+1} = \arg\min_{v \in Q}\left\{\sum_{i=0}^{k}[f(v_i) + \langle \nabla f(v_i), v - v_i \rangle] + \frac{L}{2}\| v - v_0 \|^2\right\}, \quad k \geq 0.$$

Defining $x_k = \pi_Q\left(v_k - \frac{1}{L}\nabla f(v_k)\right)$, we get $\sum_{i=0}^{k}(f(x_k) - f^*) \leq \frac{L}{2}\| v_0 - x^* \|^2$. Finally, combining these two ideas, leads to the Fast Gradient Method (FGM):

$$v_k = \arg\min_{v \in Q}\left\{\sum_{i=0}^{k-1}\frac{i+1}{2}[f(y_i) + \langle \nabla f(y_i), v - y_i \rangle] + \frac{L}{2}\| v - x_0 \|^2\right\},$$

$$y_k = \frac{2}{k+2}v_k + \frac{k}{k+1}x_k, \quad x_{k+1} = \pi_Q(y_k - \frac{1}{L}\nabla f(y_k)), \quad k \geq 0. \quad (22)$$

It can be shown that $f(x_k) - f^* \leq \frac{2L\|x_0 - x^*\|^2}{(k+1)(k+2)}$, see Nesterov (2005). Thus, this method computes an ε-solution to problem (18) in $O(\frac{L^{1/2}R}{\varepsilon^{1/2}})$ iterations. Under assumption (21), this rate of convergence is optimal, Nemirovski and Yudin (1983). The first FGM was proposed in Nesterov (1983).

Second-Order Methods

If the second derivative of objective function is available, we can apply to problem (18) the second order schemes. Unfortunately, in this situation the classical Newton method does not allow the worst-case global complexity analysis. In order to get the full theoretical justification, we need to apply cubic regularization, Nesterov and Polyak (2006). Namely, let us assume that

$$\| \nabla^2 f(x) - \nabla^2 f(y) \| \leq K \| x - y \|, \quad x, y \in Q. \quad (23)$$

For problem (18) and (23), consider the following method:

$$x_{k+1} = \arg\min_{x \in Q}\left\{f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2}\langle \nabla^2 f(x_k)(x - x_k), x - x_k \rangle + \frac{K}{6}\| x - x_k \|^3\right\}.$$

It converges as $f(x_k) - f^* \leq \frac{27K\|x_0 - x^*\|^2}{2(k+3)^2}$, see Nesterov and Polyak (2006). Using the ideas of Fast Gradient Methods, it can be accelerated up to the rate $f(x_k) - f^* \leq \frac{14K\|x_0 - x^*\|^2}{k(k+1)(k+2)}$, $k \geq 1$, see Nesterov (2008).

Structural Optimization

For Convex Optimization, black-box Complexity Theory has a hidden drawback. Indeed, in order to

apply the corresponding schemes, we need to be sure that our problem is convex (otherwise, the methods do not work). However, the only reliable way for checking convexity is the examination of its structure. If the function is constructed from convex elements by appropriate operations, we conclude that it is convex. Thus, the structure is visible at the preparatory stage. However, later it is ignored by numerical schemes.

Several systematic ways of using the structure of nonlinear convex optimization problems have been developed. We give two most important examples.

Polynomial-Time Interior-Point Methods

This theory is based on the notion of self-concordant function (Nesterov and Nemirovski, 1994).

Definition 1. *Let f be a closed convex function with open domain. It is called self-concordant* (sc) if $f \in C^3(\mathrm{dom} f)$ and

$$D^3 f(x)[h,h,h] \le 2\left(D^2 f(x)[h,h]\right)^{3/2}, \quad x \in \mathrm{dom} f,\ h \in R^n,$$

where $D^k f(x)[h,\ldots,h]$ denotes kth differential of f at x along direction h.

The central role in the analysis of sc-functions play the local norms defined by Hessians:

$$\| h \|_x = \langle \nabla^2 f(x)h, h\rangle^{1/2}, \quad \| s \|_x^* = \langle s, [\nabla^2 f(x)]^{-1} s\rangle^{1/2},$$
$$x \in \mathrm{dom}\, f,\ s,h \in R^n.$$

Define the Dikin ellipsoid $W_r(x) = \{y : \| y - x \|_x \le r\}$. Then $W_r(x) \subset \mathrm{dom} f$ for any $x \in \mathrm{dom}\, f$ and $r \in [0,1)$. Inside the Dikin ellipsoid, all Hessians are proportional. This feature facilitates the convergence analysis of the damped Newton method

$$x_0 \in \mathrm{dom} f, \quad x_{k+1} = x_k - \frac{[\nabla^2 f(x_k)]^{-1}\nabla f(x_k)}{1 + \| \nabla f(x_k) \|_{x_k}^*}, \quad k \ge 0.$$

It can be proved that all iterations of this scheme are feasible. They either decrease the value of f by an absolute constant, or converge quadratically. The region of quadratic convergence of this scheme is described by inequality $\| \nabla f(x) \|_x^* < \frac{1}{2}$. An important characteristic of the set $\mathrm{dom} f$ is the analytic center $x_f^* = \arg\min_x f(x)$. Its existence is equivalent to boundedness of f from below. Its uniqueness implies nondegeneracy of the Hessian at any feasible point.

An important class of sc-functions is formed by self-concordant barriers (scb) defined by inequality

$$\langle \nabla f(x), h\rangle^2 \le v\langle \nabla^2 f(x)h, h\rangle, \quad x \in \mathrm{dom} f,\ h \in R^n.$$

The value v is called the parameter of scb. Using such a barrier, we can solve the standard optimization problem

$$\min_{x \in Q} \langle c, x\rangle, \quad Q = \mathrm{Cl}\,(\mathrm{dom} f). \tag{24}$$

For that, we form the central trajectory $x^*(t) = \arg\min_x \{t\langle c, x\rangle + f(x)\}$, $t \ge 0$, and follow it by the Newton method. This can be done approximately by updating the points in the Euclidean neighborhood of the central path $\{x : \| tc + \nabla f(x) \|_x^* \le \frac{1}{4}\}$ using a predictor-corrector scheme. It can find an ε-solution of problem (24) in $O(v^{1/2}\ln\frac{v}{\varepsilon})$ iterations.

It can be proved that for any convex set in R^n there exists a scb with the parameter proportional to n. However, for its computation it is necessary to evaluate n-dimensional volumes. Therefore, in practice scb are constructed by analyzing the structure of functional components. Important examples of scb are as follows:

$$Q = \{y \in R^m : \langle a_i, y\rangle \le b_i,\ i = 1,\ldots,m\},$$
$$f(x) = -\sum_{i=1}^m \ln(b_i - \langle a_i, x\rangle), \quad v = m,$$
$$Q = \{X = X^T \in R^{n\times n} : X \succeq 0\},$$
$$f(X) = -\ln\det X, \quad v = n.$$

The most efficient interior-point methods are constructed for optimization problems in conic form. They allow infeasible start, long steps and eventual local quadratic convergence. In practice, the number of iterations of such schemes is often proportional to $\ln v$.

Smoothing Technique

The idea of this approach consists in approximating the nosmooth function by a smooth one, which can be efficiently minimized by FGM (22). It appears that

for functions with explicit max-representation this can be done in a systematic way (Nesterov, 2005). Assume that

$$f(x) = \max_{u \in Q_d} \{\langle Ax, u\rangle - \phi(u)\},$$

where Q_d is a convex set and ϕ is a convex function. Let us choose a prox-function d of the set Q_d (it is strongly convex with parameter one and attains its minimum at the center u_0 with $d(u_0) = 0$). Define

$$f_\mu(x) = \max_{u \in Q_d} \{\langle Ax, u\rangle - \phi(u) - \mu d(u)\}, \quad \mu \geq 0. \quad (25)$$

Then $f(x) \geq f_\mu(x) \geq f(x) - \mu D$ with $D = \max_{u \in Q_d} d(u)$. On the other hand, by Danskin Theorem, f_μ has Lipschitz continuous gradient with constant $L_\mu = \frac{1}{\mu} \| A \|^2$, where

$$\| A \| = \max_{x,u} \{\langle Ax, u\rangle : \| x \| = 1, \| u \| = 1\}$$

(norms for x and u are different). Thus, choosing $\mu = \Omega(\varepsilon)$, an ε-solution of problem (18) can be found by minimizing f_μ over Q by a fast gradient scheme. It will need at most $O(\frac{1}{\varepsilon})$ iterations instead of $O(\frac{1}{\varepsilon^2})$ iterations for a black-box method. This difference is due to the change of the structure of the oracle.

Of course, the smoothing technique is applicable only if the computation (25) can be done in a closed form. One of the most important examples is as follows:

$$f(x) = \max_{1 \leq i \leq n} x^{(i)}, \quad Q_d = \Delta_n = \{u \geq 0 : \sum_{i=1}^n u^{(i)} = 1\},$$
$$A = I, \quad \phi(u) = 0,$$

$$\| u \| = \sum_{i=1}^n |u^{(i)}|, \quad d(u) = \sum_{i=1}^n u^{(i)} \ln u^{(i)},$$
$$f_\mu(x) = \mu \ln \left(\frac{1}{n} \sum_{i=1}^n e^{x^{(i)}/\mu} \right).$$

See

- Global Optimization
- Interior-Point Methods for Conic-Linear Optimization

References

Hiriart-Urruty, J., & Lemarechal, C. (1993). *Convex analysis and minimization algorithms*. Berlin: Springer-Verlag.

Lemarechal, C., Nemirovski, A., & Nesterov, Y. (1995). New variants of bundle methods. *Mathematical Programming, 69*, 111–147.

Levin, A. (1965). One algorithm for minimization of convex functions. *Soviet Mathematics-Doklady, 160*(6), 1244–1247 (in Russian).

Nemirovski, A., & Yudin, D. (1983). *Problem complexity and method efficiency in optimization*. New York: John Wiley & Sons.

Nesterov, Y. (2004). *Introductory lectures on convex optimization*. Boston, MA: Kluwer.

Nesterov, Y. (2005). Smooth minimization of non-smooth functions. *Mathematical Programming (A), 103*(1), 127–152.

Nesterov, Y. (2008). Accelerating the cubic regularization of Newton's method on convex problems. *Mathematical Programming, 112*(1), 159–181.

Nesterov, Y., & Nemirovski, A. (1994). *Interior point polynomial methods in convex programming: Theory and applications*. Philadelphia: SIAM.

Nesterov, Y., & Polyak, B. (2006). Cubic regularization of Newton's method and its global performance. *Mathematical Programming, 108*(1), 177–205.

Nesterov, Y. (1983). A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$. *Doklady AN SSSR* (translated as Soviet Math. Dokl.), **269** (3), 543–547.

Nesterov, Y. (2007). Gradient methods for minimizing composite functions. CORE DP 2007/76. Accepted by *Mathematical Programming*.

Newman, D. (1960). Location of the maximum of unimodal surfaces. *Journal of the ACM, 12*(3), 395–398.

Polyak, B. (1967). One general method for solving extremal problems. *Soviet Mathematics-Doklady, 174*(1), 33–36.

Rockafellar, R. (1970). *Convex analysis*. New York: Princeton University Press.

Shakespeare, W. (1602). The tragedie of Hamlet, prince of Denmarke.

Shor, N. (1985). Minimization methods for non-differentiable functions. Springer Ser. in Comp. Mathem. 3, Berlin: Springer-Verlag.

Convex Polyhedron

A set of points defined by the intersection of a finite number of linear equations and/or inequalities.

See

- Polyhedron

Convex Set

A set of points that contains the line segment connecting any two of its point. Mathematically, the set S is convex if for all $0 \leq \alpha \leq 1$ and for all x_1 and x_2 in S, the point $\alpha x_1 + (1-\alpha)x_2$ is also in S.

Convexity Rows

The constraints in the decomposition algorithm master problems that require solutions to be convex combinations of the extreme points of the subproblems.

See

- Dantzig-Wolfe Decomposition Algorithm

Convex-Programming Problem

A programming problem with convex objective function and convex inequality constraints. It is typically written as

$$\begin{aligned} &\text{Minimize } f(x) \\ &\text{subject to } g_i(x) \leq 0, \quad i = 1, ..., m, \end{aligned}$$

where the functions $f(x)$ and $g_i(x)$ are convex functions defined on Euclidean n-space.

See

- Convex Optimization
- Mathematical-Programming Problem
- Nonlinear Programming

CONWIP

CONstant WIP (Work in Process), corresponding to a pull-type production control system in which the number of parts in the system is kept fixed.

See

- Kanban
- Pull System

References

Spearman, M., Woodruff, D., & Hopp, W. (1990). CONWIP: A pull alternative to kanban. *International Journal of Production Research, 28*, 879–894.

Copula

A probability distribution function used to describe the dependence between random variables, which allows the joint cumulative distribution function CDF to be expressed in terms of the marginal CDFs and the copula. This representation enables the estimation of the marginals and the dependent behavior to be decoupled and the generation of the dependent random variables via the inverse transform method. Specifically, if $X_i \sim F_i$, $i = 1, \ldots, n$, where F_i are the marginal CDFs, the copula function is a mapping C: $[0,1]^n \rightarrow [0,1]$ given by (Nelsen 2010)

$$C(u_l, \ldots, u_n) = P(F(X_l) \leq u_l, \ldots, F(X_n) \leq u_n).$$

Thus, if the joint uniform random numbers $(U_1, \ldots, U_n)$ are generated according to C, the set of dependent random variates $(X_1, \ldots, X_n)$ can be generated by applying the corresponding inverse transform method to each component, i.e., $X_i = F_i^{-1}(U_i)$. The most well-known families of copulas are the Gaussian and the Archimedean.

See

- Inverse Transform Method
- Monte Carlo Methods
- Simulation of Stochastic Discrete-Event Systems
- Stochastic Input Model Selection

References

Nelsen, R. B. (2010). *An introduction to copulas* (2nd ed.). New York: Springer.

Corner Point

▶ Extreme Point

Corporate Strategy

Arnoldo C. Hax[1] and Nicolas S. Majluf[2]
[1]Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Pontificia Universidad Católica de Chile, Santiago, Chile

Introduction

A formal strategic planning process distinguishes three perspectives: corporate, business, and functional. These perspectives are different both in term of the nature of the decisions they address, as well as the organizational units and managers involved in formulating and implementing the corresponding action programs generated by the strategy formation process.

The corporate level deals with the tasks that cannot be delegated downward in the organization, because they need the broadest possible scope — involving the whole firm — to be properly addressed. The business level faces those decisions that are critical to establish a sustainable competitive advantage, leading toward superior economic returns in the industry where the business competes. The functional level attempts to develop and nurture the core competencies of the firm, the capabilities that are the sources of the competitive advantages.

This article deals exclusively with corporate strategic tasks (Hax and Majluf 1996). There are three different imperatives — leadership, economic, and managerial — that are useful to characterize these tasks, depending on whether the concern is with shaping the vision of the firm, extracting the highest profitability levels, or assuring proper coordination and managerial capabilities.

The Leadership Imperative

This imperative is commonly associated with the person of the CEO, who is expected to define a vision for the firm, and communicate it in a way that generates contagious enthusiasm.

The CEO's vision provides a sense of purpose to the organization, poses a significant but yet attainable challenge, and draws the basic direction to the pursuit of that challenge. Successful organizations invariable seem to have competent leaders who are able to define and transmit a creative vision, one that generates a spirit of success. In other words, success breeds success.

Hamel and Prahalad (1989) argued that the vision of the firm should carry with it an obsession that they refer to as Strategic Intent. It implies a sizable stretch for the organization that requires leveraging resources to reach seemingly unattainable goals.

Much has been written and said about leadership including the controversy on nature or nurture — whether leaders are born or made — and on the existence of common characteristics to describe successful leaders (Schein 1992; Kotter 1988). This literature is not reviewed here. Instead the concentration is on the economic and managerial imperatives of the corporate strategic tasks. Nonetheless, the set of corporate tasks that deal with the economic and managerial imperatives are the critical instruments to imprint the vision of the firm. The leadership capabilities are expressed and made tangible through the tasks that are discussed herein (Pfeffer 1992).

The Economic Imperative

This imperative is concerned with creating value at the corporate level. The acid test is whether the businesses of the firm are benefitting from being together, or if they would be better off as separate and autonomous units. From this point of view, the essence of corporate strategy is to assure that the value of the whole firm is bigger than the sum of the contributions of its businesses as independent units.

The economic imperative involves three central issues: the definition of the businesses of the firm; the identification and exploitation of interrelationships across those businesses, and the coordination of the business activities that allow sharing assets and skills (Porter 1987; Pearson 1989).

There are eight corporate tasks that are associated with the economic imperative of corporate strategy. The first one is the Environmental Scan at the Corporate Level, which allow us to start the reflection

of the firm's competitive position by a thorough understanding of the external forces that it is facing. One of the principal objectives of strategy is to seek a proper alignment between the firm and its environment. Therefore, it seems logical to start the corporate strategic planning process with a rigorous examination of the external environment.

The seven additional tasks imply critical strategic decisions seeking the attainment of corporate competitive advantages. They are mission of the firm, business segmentation, horizontal strategy, vertical integration, corporate philosophy, strategic posture of the firm, and portfolio management. The essence of these tasks are discussed next.

1. Environmental Scan at the corporate level — Understanding the external forces impacting the firm: The Environmental Scan provides an assessment of the distinct business opportunities offered by the geographical regions in which the firm operates. It also examines the general trends of the various industrial sectors related to the portfolio of businesses of the corporation. Finally, it describes the favorable and unfavorable impacts to the firm from technological trends, supply of human resources, as well as political, social, and legal factors. The output of the Environmental Scan is the identification of key opportunities and threats resulting from the impact of external factors.
2. The mission of the firm — Choosing competitive domains and the way to compete: The mission of the firm defines the business scope — products, markets, and geographical locations — as well as the unique competencies that determine its capabilities. The level of aggregation used to express this mission statement is very broad, because of the need to encompass all the critical activities and capabilities of the corporation.

   The mission of the firm defines the overall portfolio of businesses. It selects the businesses in which the firm will enter or exit, as well as the discretionary allocation of tangible and intangible resources assigned to them. The selection of a business scope at the level of the firm is often very hard to reverse without incurring in significant or prohibitive costs. The development of unique competencies shape the corporate advantage, namely, the capabilities that will be transferred across the portfolio of businesses.

   The mission of the firm involves two of the most essential decisions of corporate strategy: selecting the businesses of the firm, and integrating the business strategies to create additional economic value. Mistakes in these two categories of decisions could be painful, because the stakes that are assigned to the resulting bets are very high indeed.
3. Business segmentation — Selecting planning and organizational focuses: The mission of the firm defines its business scope, namely the products and services it generates, the markets it serves, and the geographical locations in which it operates. The business segmentation defines the perspectives or dimensions that will be used to group these activities in a way that will be managed most effectively. It adds planning and organizational focuses which are central for both the strategic analysis and the implementation of the business strategies. This concept is of great importance in the conduct of a formal strategic planning process, since the resulting businesses are the most relevant units of analysis in that process.
4. Horizontal strategy — Pursuing synergistic linkages across business units: One could argue that horizontal strategies are the primary sources for corporate advantage of a diversified firm. It is through the detection and realization of the existing synergy across the various businesses that significant additional economic value can be created. The value chain is the basic framework that is used to detect opportunities for sharing resources and activities across businesses (Porter 1985). The resulting degree of linkages among businesses determines their relative autonomy and independence.

   The mission of the firm defines the business scope; business segmentation organizes the businesses into planning and managerial units; horizontal strategies determines their degree of interdependence. Consequently, these tasks are highly linked. Moreover, the mission of the firm also defines the current and future corporate core competencies, which are the basis that supports the relationship among the various businesses, and the role to be played by horizontal strategy.
5. Vertical integration — Defining the boundaries of the firm: Vertical integration determines the breadth

of the value chain, as well as the intensity of each of the activities performed internally by the firm. It specifies the firm's boundaries, and establishes the relationship of the firm with its primary outside constituencies — suppliers. distributors, and customers.

The major benefits of vertical integration are realized through: cost reductions from economies of scale and scope; creation of defensive market power against suppliers and clients; and creation of offensive market power to profit from new business opportunities. The main deterrents of vertical integration are: diseconomies of scale from increases in overhead and capital investments; loss of flexibility; and administrative penalties stemming from more complex managerial activities (Stuckey and White 1993; Harrigan 1985; Walker 1988; Teece 1987).

6. Corporate philosophy — Defining the relationship between the firm and its stakeholders: The corporate philosophy provides a unifying theme and a statement of basic principles for the organization. First, it addresses the relationship between the firm and its employees, customers, suppliers, communities, and shareholders. Second, it specifies broad objectives for the firm's growth and profitability. Third, it defines the basic corporate policies; and finally, it comments on issues of ethics, beliefs, and rules of personal and corporate conduct.

   The corporate philosophy is the task that is most closely related to the leadership imperative, insofar as bringing a capability to articulate key elements of the CEO's vision.
7. Strategic posture of the firm — identifying the strategic thrusts, and corporate performance objectives: The strategic posture of the firm is a set of pragmatic requirements developed at the corporate level to guide the formulation of corporate, business, and functional strategies. The strategic thrusts characterize the strategic agenda of the firm. They identify all of the key strategic issues, and signal the organizational units responsible to respond to them. The corporate performance objectives define the key indicators used to evaluate the managerial results, and assign numerical targets as an expression of the strategic intent of the firm. The strategic posture captures the outputs of all of the previous tasks and use them as challenges to be recognized and dealt with in terms of action-driven issues.
8. Portfolio management — Assigning priorities for resource allocation and identifying opportunities for diversification and divestment: Portfolio management and resource allocation have always been recognized as responsibilities that reside squarely at the corporate level. As noted above, the development of core competencies shared by the various businesses of the firm constitute a critical source of corporate advantage. Those competencies are borne from resources that the firm should be able to nurture and deploy effectively, including: physical assets, like plant and equipment; intangible assets, like highly-recognized brands; and capabilities, like skills associated with product design and development.

   The heart of an effective resource allocation process is the capacity to create economic value. Sometimes, this value emerges from internal activities of the firm, other times it is acquired from external sources through mergers, acquisitions, joint ventures, and other forms of alliances. Even, on occasions, value can be created by divesting businesses that are not earning their cost of capital, i.e., they are destroying instead of adding value to the firm. Portfolio management deals with all of these critical issues.

   In the 1980s, most developed economies faced periods of stagnation which have forced firms to implement drastic restructuring policies. Restructuring leads to the realignment of physical assets (including divestment), human resources, and organizational boundaries of the various businesses with the intent of reshaping their structure and performance. Restructuring decisions are also part of portfolio management (Donaldson 1994).

The Managerial Imperative

This imperative is the major determinant for a successful implementation of corporate strategy. It involves two additional important corporate tasks: the design of the firm managerial infrastructure, and the management of its key personnel.

9. Managerial infrastructure — Designing and adjusting the organizational structure, managerial processes,

and systems in consonance with the culture of the firm to facilitate the implementation of strategy: Organizational structure and administrative systems constitute the managerial infrastructure of the firm. An effective managerial infrastructure is critical for the successful implementation of the strategies of the firm. Its ultimate objective is the development of corporate values, managerial capabilities, organizational responsibilities, and managerial processes to create a self-sustaining set of rules that allow the decentralization of the activities of the firm.

The term organizational architecture is commonly used to designate the design efforts that produce an alignment between the environment, the organizational resources, the culture of the firm, and its strategy (Nadler et al. 1992).

10. Human resources management of key personnel — selection, development, appraisal, rewards, and promotion: Regardless how large a corporation is, it will be always managed by a few key individuals. Percy Barnevik, the CEO of Asea Brown-Boveri (ABB), a successful global company, stated that one of ABB's biggest priority and crucial bottleneck is to create global managers. He immediately added, however, that a global company does not need thousands of them. At ABB, five hundred out of a total of fifteen thousand managers are enough to make ABB work well (Taylor 1991).

    Tom MacAvoy, the former President of Corning Glass-Works, used to talk, in a rather colorful way, about the need for one hundred centurions to run an organization. These are huge corporations, with operations in over one hundred countries. When it comes to identify the key personnel they need, the numbers are surprisingly small; yet, the process of identifying, developing, promoting, rewarding, and retaining them, is one of the toughest challenges that an organization faces.

The Fundamental Elements in the Definition of Corporate Strategy

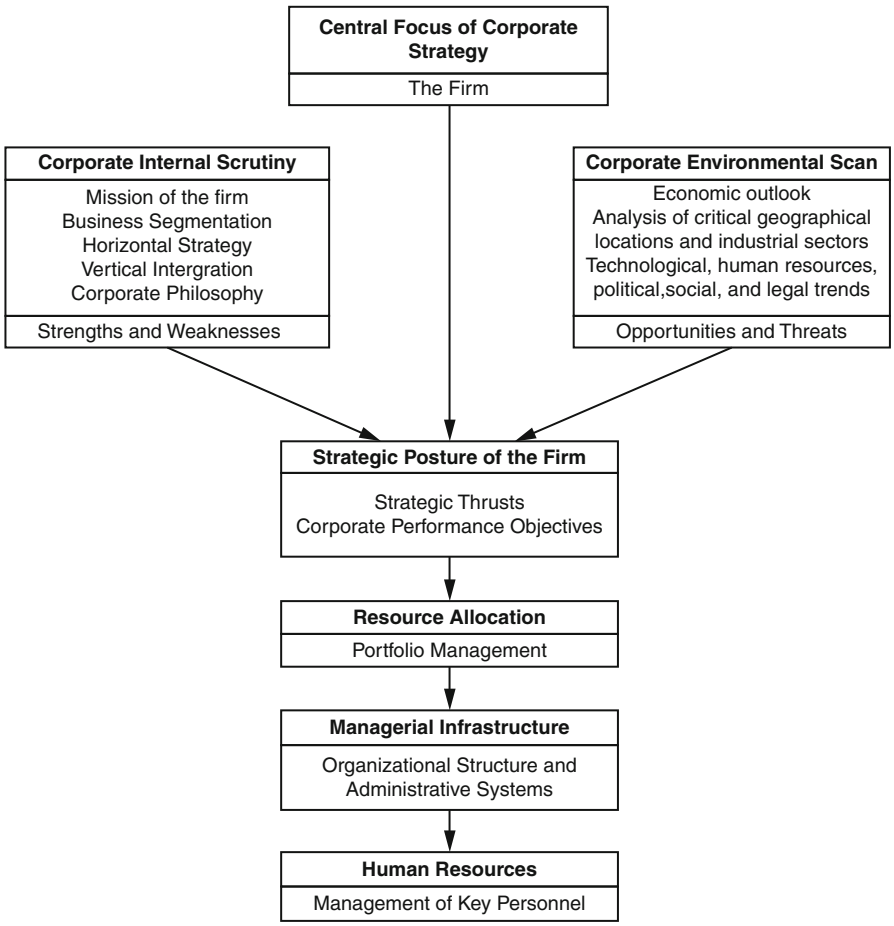
The corporate strategic tasks can be organized in a strategic planning framework, "The Fundamental Elements in the Definition of Corporate Strategy: The Ten Tasks" (Fig. 1).

The first element of the framework — The Central Focus of Corporate Strategy — consists in identifying the entity that is going to be part of the corporate strategic analysis. As opposed to the case of business strategy, where the unit of analysis is the Strategic Business Unit (SBU), corporate strategy can be applied at different levels in a large diversified organization. The amplest possible scope is the firm as a whole. There are circumstances, however, under which the scope of the analysis to a sector, group, or division of a given organization should be narrowed. These entities should encompass a number of different business units to be the subject of a meaningful corporate strategic analysis.

The next two elements of the framework are Corporate Environmental Scan and Corporate Internal Scrutiny. But, before addressing their collective tasks, it is important to note that throughout the corporate strategic analysis, existing conditions are contrasted with future ones. Thus, an underlying time frame is required to be spelled out at the beginning of the planning process.

In the case of the Corporate Environmental Scan, there are two different treatments of the future. When dealing with completely uncontrollable factors, there is a need to forecast their most likely trends to be able to understand their potential impacts. There are situations, however, in which the corporation would like to influence future events, especially when it can exercise some degree of control that will allow the future to be shaped to an advantage. By contrast, in all of the tasks that are part of the Internal Scrutiny, the future represents a state being directed at through a set of controllable decisions.

The Corporate Environmental Scan should be conducted first in the planning process, because it serves to frame the impacts resulting from the external environment. It has also the important role of transferring a common set of assumptions to the various businesses and functional managers of the firm, to serve as inputs in their own strategic planning efforts. It gives a sense of uniformity to the strategic planning thinking across all the key organizational units of the firm. This task culminates with the recognition of opportunities — the favorable impacts of the external environment which the corporation would like to seize — and threats — the unfavorable impacts which the corporation would like to neutralize.



Corporate Strategy, Fig. 1 The fundamental elements in the definition of corporate strategy: the ten tasks

The Corporate Internal Scrutiny captures the key actions and decisions the corporation has to address to gain a competitive position that is in line with the challenges generated by the external environment, and conducive to the development of a sustainable corporate advantage. This advantage is transferable to the various business units of the firm, and enhances its resources and capabilities. The tasks which are part of the Internal Scrutiny in our framework are:

- Mission of the Firm
- Business Segmentation
- Horizontal Strategy
- Vertical Integration
- Corporate Philosophy

In all of these decisions, the current state is contrasted with a desirable future one. The process then proceeds to define the challenges those changes generate for the formulation of corporate strategy. The Internal Scrutiny concludes with an overall statement of corporate strengths that the firm wishes to maintain and reinforce, as well as a statement of corporate weaknesses that the firm wishes to correct or eliminate.

The Corporate Environmental Scan and the Corporate Internal Scrutiny provide the basic inputs that will define the Strategic Posture of the firm. This task serves as a synthesis of the analysis conducted so far, and captures the strategic agenda of the firm. The strategic thrusts are a powerful expression of all of the issues that, from the perspective of the firm, need to be addressed to come out with an integrative strategy. The Corporate Performance Objectives define the key indicators that will be used to detect the operational and strategic effectiveness of the firm. The Strategic Posture is the essence of the formulation of the corporate strategy, and as such, it is a task that should receive the utmost attention. When properly conducted, the firm is able to frame the activities, responsibilities, and performance measurements that are critical for its superior strategic position.

The subsequent task, Resource Allocation and Portfolio Management, permits to backup the strategic actions implicit in the Strategic Posture of the firm with the necessary resources needed for their deployment. This leads to the realm of strategy implementation. These implementation efforts are reinforced strongly by the remaining two corporate tasks: Managerial Infrastructure and Human Resources Management of Key Personnel.

See

- ▶ Computational Organization Theory
- ▶ Organization

References

Donaldson, G. (1994). *Corporate restructuring, managing the change process from within*. Boston: Harvard Business School Press.

Hamel, G., & Prahalad, C. K. (1989). Strategic intent. *Harvard Business Review, 67*(3), 63–76.

Harrigan, K. R. (1985). *Strategic flexibility: A management guide for changing times*. Lexington, MA: Lexington Books.

Hax, A. C., & Majluf, N. S. (1996). *The strategy concept and process: A pragmatic approach* (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Kotter, J. P. (1988). *The leadership factor*. New York: Free Press.

Nadler, D. A., Gerstein, M. S., Shaw, R. B., & Associates. (1992). *Organizational architecture: Designs for changing organizations*. San Francisco: Josses-Bass.

Pearson, A. E. (1989). Six basics for general managers. *Harvard Business Review, 67*(4), 94–101.

Pfeffer, J. (1992). *Managing with power: Politics and influence in organizations*. Boston: Harvard Business School Press.

Porter, M. E. (1985). *Competitive advantage*. New York: Free Press.

Porter, M. E. (1987). From competitive advantage to corporate strategy. *Harvard Business Review, 65*(3), 43–59.

Schein, E. E. (1992). *Organizational culture and leadership* (2nd ed.). San Francisco: Josses-Bass.

Stuckey, J., & White, D. (1993). When and when not to vertically integrate. *Sloan Management Review, 34*(3), 71–83.

Taylor, W. (1991). The logic of global business: An interview with ABB's Percy barnevik. *Harvard Business Review, 69*(2), 90–105.

Teece, D. J. (1987). Profiting from technological innovations: Implications for integration, collaboration, licensing, and public policy. In D. J. Teece (Ed.), *The competitive challenge: Strategies for industrial innovations and renewal*. Cambridge, MA: Ballinger Publishing.

Walker, G. (1988). Strategic sourcing, vertical integration and transaction costs. *Interfaces, 19*(3), 62–73.

Cost Analysis

Stephen J. Balut[1] and Thomas R. Gulledge[2]
[1]Institute for Defense Analyses, Alexandria, VA, USA
[2]George Mason University, Fairfax, VA, USA

Introduction

Cost analysis is the process of estimating the individual and comparative costs of alternative ways of accomplishing an objective. The goal is not to forecast precisely accurate costs, but rather to reveal the extent to which one alternative costs more or less than another. A cost analysis is often conducted in conjunction with an effectiveness analysis to aid in the selection of one alternative over others.

Evolution

Cost analysis emerged as part of a broader initiative in the late 1940s and early 1950s to apply economic principles to the decision making process of the U.S. Department of Defense (DoD). A confluence of events following World War II resulted in a dramatic and enduring change in the way resource allocation decisions were made in public organizations. The development and evolution of cost-effectiveness analysis and cost analysis occurred nearly simultaneously and are closely related. Both types of analysis make use of operations research methods.

Operations research was invented and applied mainly by civilian scientists in support of the war effort. From its inception, operations research sought to "use scientific methods to get the most out of available resources" (Quade 1971). Immediately following the war, many of these scientists were retained by the Military Departments to apply newly developed quantitative methods to aid defense decisions. The forerunners of the RAND Corporation, the Institute for Defense Analyses (IDA), and the Center for Naval Analyses (CNA) were formed during this period.

After the war, separation of military responsibilities between the U.S. Armed Services broke down as a consequence of the rapid development of military technology and the different character of the military

threat (Smale 1967). The Services began competing for missions and disputes were settled via approval of budgets for new weapon systems. Competing systems were considered on the basis of cost-effectiveness. When equally effective weapon systems were compared, those estimated to cost the least won funding approvals. The analytical procedure applied to such decisions was first named weapon systems analysis, later shortened to systems analysis. The first documented systems analysis was accomplished in 1949 by the RAND Corporation and compared the B-52 to a turbo-prop bomber. The use of dollar costs as a proxy for real costs changed the basic systems analysis question from "Which weapon system is best for the job?" to "Given a fixed budget, which weapon system is most cost effective?" (Smale 1967; Novick 1988).

The birth of cost analysis as a separate activity occurred in the early 1950s and is attributed to Novick (1988), a cost analyst with the RAND Corporation. Novick pioneered weapon system cost analysis and is referred to as the father of cost analysis. Novick and his group at RAND are attributed with development of the fundamental building blocks of cost analysis. These include separation of total costs into cost elements, separation of one-time and recurring costs, development of cost estimating relationships, and development of conceptual costs or order-of-magnitude estimates used to compare future system proposals. Novick's group went on to invent parametric cost estimating, incremental costing, and "Total Force Costing" (Novick 1988; Hough 1989).

In the early 1960s, the Department of Defense established and implemented a centralized resource allocation process called the Planning, Programming and Budgeting System (PPBS). Under this system, future defense resources were allocated to missions in a systematic, rational manner using cost-effectiveness as the decision criterion. In 1961, a Systems Analysis Office was established within the Office of the Secretary of Defense (OSD) to help implement this new resource allocation procedure. In 1965, a Cost Analysis Division was established within the office of the Assistant Secretary of Defense, Systems Analysis. With this act, cost analysis gained a primary role in the examination of alternative force structures at the OSD level. Also in 1965, the PPBS system was extended to all federal agencies by President Lyndon Johnson.

The next few decades brought initiatives that strengthened the cost analysis capabilities of the DoD. The military departments established cost analysis offices at headquarters and major commands and staffed them, at least in part, with people trained and experienced in the methods of operations research. The DoD initiated systematic collection of cost information from defense contractors to provide defense cost analysts with records of cost experiences on major weapon system acquisitions. These records formed the bases of estimates of the costs of proposed systems at acquisition milestone decision points, strengthened the DoD's position during contract negotiations, and provided for DoD tracking of negotiated costs. In 1971, Deputy Secretary of Defense Packard instituted defense acquisition reforms that included establishment of the DoD Cost Analysis Improvement Group (Hough 1989), the requirement for independent parametric estimates for new systems acquisitions, formalization of cost analysis reviews at milestone decision points, and requirements for the military departments to improve their cost-estimating capabilities. As part of the Packard reform, cost was elevated to a principal design parameter with implementation of the "Design to Cost" initiative (Hough 1989). Ten years later, in 1981, Deputy Secretary of Defense Carlucci placed further demands on the DoD's cost analysis capabilities. He instituted the practice of "Multi-Year Procurement" based on benefit/risk analyses, "Budget to Most Likely or Expected Cost," budgeting more realistically for inflation, the use of economic production rates, the requirement to forecast business base at defense contractors' plants, increased efforts to quantify cost risk and uncertainty, and provision of greater incentives on design-to-cost goals by tying award fees to actual costs achieved in production.

Throughout the 1970s and 1980s, the practice of cost analysis continued to expand mainly in the public sector. The US government's cost analysis organizations grew in size by drawing people skilled in engineering, economics, operations research, accounting, mathematics, statistics, business, and related fields. Several focused educational programs were initiated to support this budding profession at military universities, including the Air Force Institute of Technology, the Naval Postgraduate School, and the Defense Systems Management College.

The 1990s brought a surge of activity in cost analysis with institutionalization of a Cost and Operational Effectiveness Analysis (COEA) as an integral part of the defense acquisition process. COEAs are required to be conducted and presented to the Defense Acquisition Executive at major milestone in the acquisition of a major weapon system.

Around the turn of the century, the DoD established a preference for the use of evolutionary acquisition strategies (DoD D5000.01) that promise to speed the delivery of advanced capabilities to warfighters while also providing follow-on improvements in capabilities as planned technological advances are achieved. Adoption of this approach provides cost analysts with the challenge of estimating the costs of systems that embody ultimate capabilities that cannot be fully defined at the beginning of the acquisition program.

Methods

Cost analysis is a sequential process: first identification, then measurement, and finally evaluation of alternatives. This involves the structuring and analysis of resource alternatives in a full planning context. In the case of defense, the size of the U.S. defense budget limits the dollars available to provide for the national defense. Monies spent on one mission/capability/weapon system are not available to spend on another. "Therefore, properly constructed cost estimates and cost analyses are essential because an accurate assessment of the cost of individual programs is the first necessary step towards understanding the comparative benefits of alternative programs and capabilities" (Smale 1967).

Economic costs are benefits lost and are often referred to as alternative costs or opportunity costs (Fisher 1970). An estimate of the economic cost of one choice, decision or alternative, within this context, is an estimate of the benefits that could otherwise have been obtained by choosing the best of the remaining alternatives. When constructed in this way, costs have the same dimension as benefits, and direct comparison is possible.

The following cost analysis concepts are briefly described here: the Work Breakdown Structure (WBS), Estimating Relationships (ER), and Cost Progress Curves. The treatment is not comprehensive in any sense and is provided to give those completely unfamiliar with the methods of cost analysis an idea of what is involved.

Work Breakdown Structure — Cost analysts break complex systems down into pieces before attempting to estimate their costs. A notion fundamental to this process is the Work Breakdown Structure (WBS) (U.S. Air Force Material Command 1993). The basic concept of a WBS is to represent an aircraft system, for example, as a hierarchical tree composed of hardware, software, facilities, data, services, and other work tasks. This tree completely defines the product and the work to be accomplished. It relates elements of work to each other and to the end product. Cost analysts usually estimate total systems costs as the sum of the costs of the individual elements of the WBS.

Estimating Relationships — Another tool that is fundamental to cost analysis is the estimating relationship (ER). In a broad sense, estimating relationships are transformation devices which permit cost analysts to go from basic inputs (e.g. descriptive information for some future weapon system) to estimates of the cost of output-oriented packages of military capability (Fisher 1970). More specifically, ERs are analytic devices that relate various categories of cost (e.g. dollars or physical units) to explanatory variables referred to as cost drivers. While taking many different forms, ERs are usually mathematical functions derived from empirical data using statistical analyses.

Cost Progress Curves — The basic notion of a learning curve is that, as a work procedure (e.g. sequence of steps/activities) is repeated, the person performing the procedure normally becomes better or more efficient at performing the procedure. The reduction in time or cost to perform the procedure is commonly attributed to learning. Cost analysts, who are more interested in reductions in cost, refer to this phenomenon as cost progress rather than learning.

The theory of cost progress curves states that as the total quantity of units (e.g. aircraft, wings, or fuselages) produced doubles, the cost per unit declines by some constant percentage. Wright (1936) empirically demonstrated the principle (Asher 1956). The standard mathematical model is a power function that relates manufacturing labor hours required to

produce a particular unit to the cumulative number of units produced. The functional form is simply:

$$C = aQ^b$$

where C is the number of hours required to produce unit Q, a is the labor hours required to produce the first unit, and b is a parameter that measures the amount of cost progress reflected in the data used to estimate the model parameters. The form is a hyperbolic function that is linear in logarithmic space. The characteristic of linearity in logarithmic space and the ease of application account for the general acceptance and popularity of the cost progress curve among cost analysts. The cost progress curve is applied widely by defense cost analysts when estimating the costs of alternative force sizes and compositions.

Professional Organizations

As cost analysis evolved over the past few decades, a number of professional organizations were formed to further advance cost analysis and related professional activities. The Cost-Effectiveness Technical Section of the Operations Research Society of America (now the Institute for Operations Research and the Management Sciences–INFORMS) was formed in 1956 to provide for the exchange of experiences in conducting such analyses. This organization has since changed its name to the Military Application Section (MAS) of INFORMS.

The National Estimating Society (NES) was formed in 1978. This organization's focus was on cost estimating from the perspective of the private sector. The formation of the Institute of Cost Analysis (ICA) in 1981 was referred to as the most significant event of the decade for DoD cost analysts (Hough 1989). ICA was dedicated to the furtherance of cost analysis in the public and private sectors. Both ICA and NES established programs under which the technical competence of members were certified, leading to a designation of Certified Cost Analyst or Certified Cost Estimator. ICA and NES subsequently merged to form the Society of Cost Estimating and Analysis (SCEA). SCEA continues the certification process by conferring the "Certified Cost Estimator/Analyst" designation to those who pass a qualifying examination.

See

- ▶ Center for Naval Analyses
- ▶ Cost-Effectiveness Analysis
- ▶ RAND Corporation

References

Asher, H. (1956). *Cost-quantity relationships in the airframe industry, R-291*. Santa Monica, CA: The RAND Corporation.

DoD Directive 5000.01 (2003). *The defense acquisition system*. Washington, DC: Department of Defense.

Fisher, G. H. (1970). *Cost considerations in systems analysis, R-490-ASD*. Santa Monica, CA: The RAND Corporation.

Hough, P. G. (1989). *Birth of a profession: Four decades of military cost analysis*. Santa Monica, CA: The RAND Corporation.

Novick, D. (1988). *Beginnings of military cost analysis: 1950–1961, P-7425*. Santa Monica, CA: The RAND Corporation.

Quade, E. S. (1971). *A history of cost-effectiveness analysis, Paper P-4557*. Santa Monica, CA: The RAND Corporation.

Smale, G. F. (1967). *A commentary on defense management*. Washington, DC: Industrial College of the Armed Forces.

U.S. Air Force Materiel Command (1993). *Work breakdown structures for defense material items*. Military Standard 881B.

Wright, T. P. (1936). Factors affecting the cost of air-planes. *Journal of Aeronautical Sciences, 3*, 122–128.

Cost Coefficient

In a linear programming problem, the generic name given to the objective function coefficients.

Cost Range

- ▶ Ranging
- ▶ Sensitivity Analysis

Cost Row

The row in a simplex tableau that contains the reduced costs of the associated feasible bases.

See

- ▶ Simplex Method (Algorithm)

Cost Slope

The rate of cost change per unit of time duration of a project's work item.

See

► Network Planning

Cost Vector

In a linear-programming problem, a row vector $\boldsymbol{c}$ whose components are the objective function coefficients of the problem.

See

► Cost Row

Cost-Effectiveness Analysis

Norman Keith Womer
University of Missouri-St Louis, St. Louis, MO, USA

Introduction

Cost effectiveness analysis is a practical way of assessing the usefulness of public projects. The history of the subject can be traced to Dupuit's classic 1844 paper, "On the Measurement of the Utility of Public Works." The technique has been a mainstay of the Army Corps of Engineers since 1902. Recent variations of the technique have been labeled cost effectiveness analysis, cost benefit analysis, systems analysis, or merely analysis. It has been extensively applied to projects in defense, transportation, irrigation, waterways, and housing. Cost effectiveness analysis is required by law and regulation throughout the federal government to decide among certain alternative policies and projects. It has been recently required in federal regulations designed to protect human health, safety, or the environment. Despite this fact, the practice of cost effectiveness analysis is subject to criticism. Robert Dorfman (1996) declared, "Three prominent shortcomings of benefit-cost analysis as currently practiced are (1) it does not identify the population segments whom the proposed measure benefits or harms, (2) it attempts to reduce all comparisons to a single dimension, generally dollars and cents, and (3) it conceals the degree of inaccuracy or uncertainty in its estimates."

Cost effectiveness analysis (CEA) is the process of using theory, data, and models to examine a problem's relevant objectives and alternative means of achieving them. It is used to compare the costs, benefits, and risks of alternative solutions to a problem and to assist decision makers in choosing among them. The differences between cost effectiveness analysis and the discipline of operations research itself are subtle and, in some treatments, merely a matter of emphasis (see the discussion in Quade 1971). The convention adopted here is that operations research is a body of knowledge that includes all of the tools and methods that might be used in any study, while cost effectiveness analysis is a particular application of models and methods to a choice problem.

Sometimes CEA is portrayed as the combination of the difficult problem of measuring effectiveness with the rather mundane problem of cost estimation. In fact, cost measurement is an important issue. Cost effectiveness analysis provides a tool for effective resource allocation only when all the resource implications associated with each alternative — both direct and indirect — are included in the analysis. The opportunity cost of a proposed allocation of resources is the value of those resources in their best alternative use. The very concept of opportunity cost therefore requires knowledge of the goals and objectives, measures of effectiveness, the other alternatives and constraints of the organization. That is, to employ this basic concept of cost, a careful analysis of the problem must be accomplished.

Therefore, CEA must focus on the process of modeling both cost and effectiveness to develop relevant measures that shed light on the problem under study. Ultimately, CEA consists of methods for evaluating vectors of measures. In the process, CEA must grapple with issues like the scale of operations, risk, uncertainty, timing, and actions of other players.

The Role of Models

Figure 1, adapted from Quade (1971), portrays the elements of CEA. Models are used in CEA to aid in the evaluation of alternatives. These models often take the form of equations that relate the physical description of alternative systems to various impacts of their production and use. The models may concern the acquisition of the systems, their operation, or various circumstances associated with applying the system in an environment.

There are many assumptions in any analysis. One important class of assumptions that is often left unstated in CEAs concerns the behavior of key players in the process. Traditionally, CEAs have been based on rather mechanical models that relate a system's physical characteristics (e.g., weight and speed) to production cost. Any reference to behavior has often been confined to vague statements about efficiency. In fact, costs and benefits result only from actions. Thus, the motivation to act is an important part of modeling costs and benefits. Unfortunately, these behavioral assumptions are often not stated explicitly. Instead, they are frequently imbedded in detailed computer simulations that attempt to emulate the simultaneous operation of complex systems in realistic environments.

Incommensurable Impacts

The output of a suite of models may be a rather long list of measured system impacts. Some of the system impacts are measurable in units of effectiveness or costs, while others are external to our frame of reference. Generally, each of the impacts will be measurable in units that are unique to that impact, for example, number of lives lost, replacement cost of lost equipment, number of minutes of error free transmission accomplished, etc. Choice requires not only the objective consideration of the measurable impacts, but also the consideration of the often immeasurable externalities. As a result, it is important that the analyst carefully report both impact measures and their accuracy and those impacts that remain unmeasured. Choice also requires the explicit use of a criterion that evaluates the impacts and their relation to the choice problem at hand.

The Analyst and the Decision Maker

In doing analysis, the first and most important issue is to understand the decision maker's problem. Answering the question "What is the problem?" often requires understanding both the organization for which the analysis is performed and the physical system or structural change that is under study. The problem may be stated in different forms at different points in time and at different levels in the organization. Thus, understanding the problem requires understanding the objectives of the entire organization.

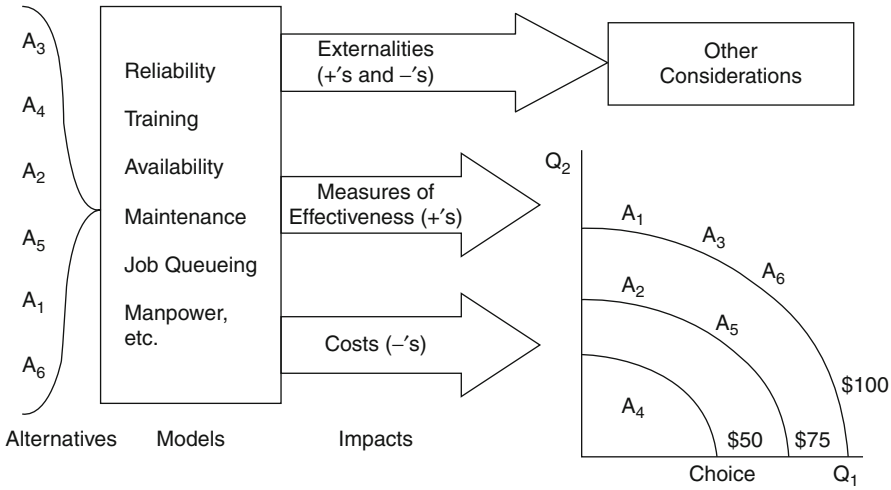
For example, consider the problem of analyzing the cost of a mission currently assigned to an aircraft system. What is the problem? Some candidates are:

- Should the existing system be replaced?
- What design should be chosen?
- Who should produce the system?
- How should the mission be performed?
- Is the mission affordable?

Often analysis is done with reference to one of these problems and then later the same study is applied to a different problem. Clearly, the alternatives, the risks, the objectives, and the cost are not independent of the problem being addressed.

Whose problem is this? It is the analyst who must choose the techniques, collect the data, model the processes, and measure the costs and outputs. It is the analyst who must justify the choices made in the particular context of the problem being addressed. Thus, it is the analyst who must be able to answer the question, "What is the Problem?"

If the role of the analyst is so large, what is left for the decision maker? The decision maker must also understand the problem and judge the value of the analysis. The decision maker must examine the completeness of the alternatives, evaluate the assumptions, examine the measurement of the impacts, and determine if risks are adequately addressed. All of these tasks are important. But the most important task of the decision maker is the task of evaluating the relative importance of the various positive and negative impacts. This includes not only the impacts that are internal to the organization but also the externalities. Evaluating the impacts also means dealing with their risks and uncertainties. The decision maker's values also include his or her attitudes toward risk. It is in this effort that the decision maker's role is uniquely different from that of the analyst.



Cost-Effectiveness Analysis, Fig. 1 The elements of cost-effectiveness analysis

Once the impacts have been evaluated, then choice can be merely a matter of adding them up and comparing the weighted impacts of each of the alternatives.

Criteria

Cost-benefit ratios — Cost effectiveness analysis often is implemented by classifying each impact of a system as either a cost or a benefit. Common units are then found for costs and for benefits and the discounted present value of each is calculated. Alternatives are compared by the ratio of these two measures.

Using a cost-benefit ratio to choose among alternatives presents several problems. Often, this approach leaves out relevant measures (i.e., treats them as the externalities depicted in Fig. 1) because those impacts cannot be evaluated in units that are comparable to the main impacts. Choosing units for the main impacts involves subjective decisions that trade-off relative measures of merit. For example, lives lost must be compared to visual pollution or environmental impact must be valued relative to economic loss. The person who determines common units for such diverse measures of merit is no longer playing the role of an analyst. That person is acting as the decision maker.

The alternative is to leave the various measures of merit uncombined. But a major problem occurs with ratio analysis when the analysis must consider multiple inputs and multiple outputs. Several ratios may be constructed but then it is not clear how these multiple ratios should be combined to determine the overall value of an alternative. Cost-benefit ratios provide the decision maker with little guidance on how to proceed in this case.

Another problem with ratio analysis is the constant returns to scale assumption that is implicit in calculating a cost-benefit ratio. By displaying the results in ratio form, the analyst implies that if the system is expanded or contracted the costs and the benefits both change proportionately. Unfortunately, the world is replete with examples of alternatives that violate such proportionality rules. The use of the incremental cost-effectiveness ratio recognizes this problem. Finally, ratio analysis does not lend itself to explicit treatments of risk and uncertainty, see Conigliani and Tancredi (2009).

Production functions — The production function approach to CEA can deal with variable returns to scale and with other nonlinearities in technology. Numerous estimates of costs and benefits for various alternatives at different scale levels are used to fit a nonlinear production function by regression. This technique can deal with several measures of input and can therefore overcome some of the difficulties of the cost-benefit ratio. Production functions can also incorporate risk described with random variables. But, the multiple regression production function also has some drawbacks. First, the use of regression tends to measure efficiency relative to average performance instead of best performance. That is, all the observations are pooled to fit the production function, a measure of average efficiency, then each alternative is compared to that average measure. Also, multiple regression requires that a single indicator for output

be used. Thus, multiple outputs must be combined into a single effectiveness indicator, similar to ratio analysis. This type of problem is especially severe in non-profit and governmental organizations where prices for outputs are unavailable or incomplete. Charnes and Cooper (1985) also criticized regression's lack of ability, "in identifying the underlying sources and amounts of inefficiencies."

Data envelopment analysis — Data envelopment analysis (DEA) provides an efficiency measure that offers some aid for the criterion problem. This linear-programming based measure has its origin in linear production theory. Golany (1988) pointed out that "DEA is quickly emerging as the leading method for efficient evaluation, in terms of both the number of research papers published and the number of applications to real world problems."

DEA is a procedure that has been designed specifically to measure relative efficiency in situations in which there are multiple measures of merit and there is no obvious objective way of aggregating measures of merit into a meaningful index of productive efficiency. Compared to regression, which averages the aggregate impact of a system, DEA is an extremal method. DEA calculates the efficiency of each alternative by comparing (via mathematical programming models) an alternative's measures of merit with the measures of merit of the other alternatives. Each alternative's measures of merit are weighed as favorably as possible. If the alternative is inefficient, DEA indicates which of its measures of merit imply its inefficiency. Also, DEA does not require the parametric specification of a production function; it derives an estimate of the production function directly from the observed data on elements of cost and effectiveness that are model outputs. DEA has been used to measure the productivity and efficiency of many organizations. It has been particularly useful for public sector organizations where market prices of outputs are not available. DEA has the potential to be extremely helpful in developing criteria in cost effectiveness analyses.

Advances — Contributions to the literature on CEA make explicit use of methods for analyzing risk and uncertainty, Conigliani and Tancredi, (2009); of dynamic models some using Markov models, Soares and Castro, (2010); and others using computable general equilibrium models, Löschel and Otto (2009).

Examples. Cost effectiveness analyses have been conducted in support of (and in opposition to) numerous significant national decisions. For example, the study of alternative delivery systems that resulted in the choice of the space shuttle, the series of studies on the Anti-Ballistic Missile, and the studies for and against the breakup of AT&T are classic studies that illustrate both the power and the fragility of this important concept.

See

- ▶ Cost Analysis
- ▶ Data Envelopment Analysis
- ▶ Measure of Effectiveness (MOE)
- ▶ Multi-Criteria Decision Making (MCDM)
- ▶ Opportunity Cost

References

Charnes, A., & Cooper, W. W. (1985). Preface to topics in data envelopment analysis. *Annals Operations Research, 2*, 59–94.

Charnes, A., Cooper, W. W., & Sueyoshi, T. (1988). A goal programming/constrained regression review of the bell system breakup. *Management Science, 34*, 1–26.

Conigliani, C., & Tancredi, A. (2009). A Bayesian model averaging approach for cost-effectiveness analyses. *Health Economics, 18*, 807–821.

Dorfman, R. (1996). Why benefit-cost analysis is widely disregarded and what to do about it. *Interfaces, 26*(5), 1–6.

Dupuit, J. (1844). *De la Mesure de l'utilité des travaux publics*. Reprinted in Jules Dupuit, De l'utilitéet de sa mesure, Torino, la Roforma sociale, 1933.

Evans, D. S., & Heckman, J. J. (1983). Natural monopoly. In D. S. Evans (Ed.), *Breaking up bell* (pp. 127–156). New York: North Holland.

Evans, D. S., & Heckman, J. J. (1988). Natural monopoly and the bell system: Response to Charnes, Cooper and Sueyoshi. *Management Science, 34*, 27–38.

Golany, B. (1988). An interactive MOLP procedure for the extension of DEA to effectiveness analysis. *Journal of the Operational Research Society, 39*, 725–734.

Gregory, W. H. (1973). NASA analyzes shuttle economics. *Aviation week and space technology*, Sept 24, 1973.

Heiss, K. P., & Morgenstern, O. (1971). *Factors for a decision on a new reusable space transportation system*. Memorandum for Dr. James C. Fletcher, Administrator NASA, Mathematica, Princeton, NJ.

Löschel, A., & Otto, V. M. (2009). Technological uncertainty and cost effectiveness of CO_2 emission reduction. *Energy Economics, 31*, S4–S17.

Operations Research Society of America. (1971). Guidelines for the practice of operations research. *Operations Research, 19*, 1123–1258.

Quade, E. S. (1964). *Analysis of military decisions*. Santa Monica, CA: United States Air Force Project Rand, R–387-PR.
Quade, E. S. (1971). *A history of cost-effectiveness*. Santa Monica, CA: United States Air Force Project Rand, P-4557.
Soares, M. O., & Castro, L. C. (2010). Simulation or cohort models? Continuous time simulation and discretized Markov models to estimate cost-effectiveness. *CHE Research Paper Centre for Health Economics*, Alcuin College, University of York, York, UK.
Sueyoshi, T. (1991). Estimation of stochastic frontier cost function using data envelopment analysis: An application to the AT&T divestiture. *Journal of the Operational Research Society, 42*, 463–477.

COV

▶ Coefficient of Variation

Covering Problem

▶ Set-Covering Problem

Coxian Distribution

A probability distribution whose Laplace-Stieltjes transform may be written as the quotient of two polynomials (i.e., a rational function). All Coxian distributions have a phase-type formulation which may include fictitious stages.

See

▶ Queueing Theory

CPM

Critical path method.

See

▶ Critical Path Method (CPM)
▶ Network Planning
▶ PERT
▶ Research and Development

CPP

▶ Chinese Postman Problem

Cramer's Rule

A formula for calculating the solution of a nonsingular system of linear equations. Cramer's rule states that the solution of the $(n \times n)$ nonsingular linear system $\boldsymbol{Ax} = \boldsymbol{b}$ is $x_i = det\ \boldsymbol{A_i(b)}/det\ \boldsymbol{A}$, $i = 1, \ldots, n$, where $det\ \boldsymbol{A}$ is the determinant of A, and $det\ \boldsymbol{A_i(b)}$ is the determinant of the matrix obtained by replacing the ith column of $\boldsymbol{A}$ by the right-hand side vector $\boldsymbol{b}$. This rule is inefficient for numerical computation and its main use is in theoretical analysis.

See

▶ Matrices and Matrix Algebra

Crash Cost

The estimated cost for a job (project) based on its crash time.

See

▶ Network Planning

Crash Time

The minimal time in which a job may be completed by expediting the work.

See

▶ Network Planning

Crew Scheduling

The determination of the temporal and special succession of the activities of staff personnel, as, for example, in an airlines, train, factory, etc. Such problems are often modeled as mathematical programs.

See

► Airline Industry Operations Research

Crime and Justice

Arnold Barnett[1], Jonathan P. Caulkins[2] and Michael D. Maltz[3]
[1]Massachusetts Institute of Technology, Cambridge, MA, USA
[2]Carnegie Mellon University, Pittsburgh, PA, USA
[3]University of Illinois at Chicago, Chicago, IL, USA

Introduction

Ever since the publications of President Johnson's Commission on Law Enforcement and Administration of Justice (Government Printing Office 1967a, b), OR/MS professionals have investigated just about all facets of the U.S. national, state, and local aspects of crime and justice. There results have had a major influence all out of proportion to their numbers; OR/MS scholars have transformed the way many decision makers think about problems of crime and punishment. Of particular importance is the research of Blumstein and Larson (1969) on the total criminal justice system.

The OR/MS contribution pervades quantitative discussions about crime and justice systems. It has generated a more precise and transparent description of the crime problem than had hitherto been available. It has achieved uneven but sometimes magnificent successes in both identifying and implementing crime-reduction strategies. And it has enhanced the scientific rigor with which criminal justice policy experiments are analyzed and interpreted.

It is not commonly known that some of the most frequently used tools of OR/MS were developed because of crime and justice problems. In the early 19th century, France began to amass statistics on the operation of the criminal justice system (Daston 1988), and the richness of these data led statisticians to devise new techniques to analyze them. Stigler (1986) describes how Simeon Denis Poisson developed the statistical distribution that bears his name – arguably the union label of the OR/MS profession – while modeling conviction rates in French courtrooms. Similarly, Hacking (1990) shows how Poisson developed the law of large numbers by modeling the reliability of jurors in criminal trials.

As noted, the application of OR to crime and justice began in the mid-1960s, when operations researchers and systems analysts on the President's Crime Commission directed their talents to the science and technology aspects of the criminal justice system (Government Printing Office 1967b). Since then, the application of OR/MS ideas in this area has burgeoned (Maltz 1994). Some of the more salient roles played by OR/MS in this field are discussed in this article, especially how OR/MS has been used in analyzing crime statistics, offender behavior, and criminal justice system dynamics. Also described here are how queueing models and optimization techniques have been applied in criminal justice contexts, how OR/MS has caused (some) criminologists to rethink some of their conclusions, the growing role of Geographic Information Systems (GIS) in criminal justice, and how OR/MS is pioneering the extension of quantitative analysis to model offenders who do not fit the traditional street offender mold.

Homicide

In discussing crime, it is natural to start with the most serious offense – murder. Led by the FBI, those assessing homicide patterns had thought it sufficient to consider annual murder rates, expressed in killings per 100,000 citizens per year. The calculated rates had a reassuring quality about them: if 50 per 100,000 citizens were murdered last year, then the other 99,950 were not murdered. Thus, after Detroit had precisely that murder rate in 1973, *The New York Times* reported that "If you live in Detroit, the odds are 2000–1 (i.e., 99,950–50) that you will not be killed by

one of your fellow citizens. Optimists searching for perspective in the city's murder statistics insist that these odds are pretty good."

But some OR/MS scholars raised a question: why measure homicide risk per year as opposed to (say) per day, per month, or per decade? Given that an urban resident has a lifetime danger of being murdered, that would seem the natural time frame over which to measure the risk. And at an annual risk of 1 in 2000, a person with a natural lifespan of 70 years would face a lifetime murder risk of 1 in 28 (!!). Refinements of this raw calculation leave its result virtually unchanged (Barnett et al. 1975; 1980; Barnett and Schwartz 1989).

The idea of estimating lifetime risk of murder has come into general use: detailed projections appeared in the 1981 FBI *Uniform Crime Reports*, and such forecasts have since been incorporated into the actuarial projections of the (U.S.) Centers for Disease Control. There is now widespread awareness that homicide is not a tragic, rare phenomenon, but instead a critical public health problem.

Parallel reasoning is likewise being applied to the lifetime risk of imprisonment. For example, even though the probability of being sent to prison on any given day is small, Bonczar and Beck (1997) estimate that 5.1% of all persons in the U.S. and 28.5% of black males will serve time in prison at some point in their lives.

Offender Behavior

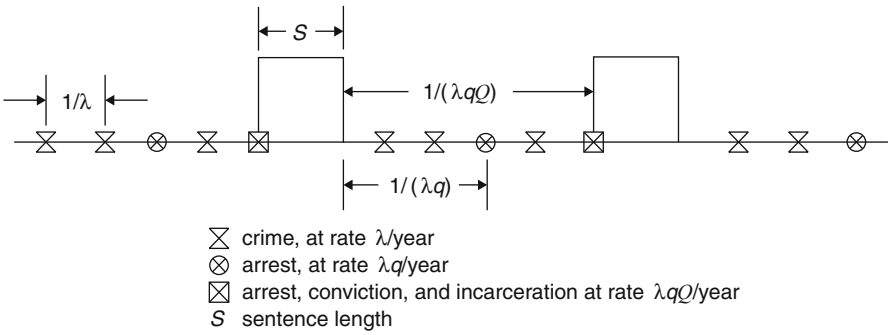
From the standpoint of public policy, it makes a great deal of difference whether existing crimes are committed by relatively few individuals who all offend frequently or by a large number who all offend rarely. OR/MS researchers have taken part in efforts to estimate the total number of offenders, some of whom may never be apprehended for their crimes (Greene and Stollmack 1981; Greene 1984). Of course, the offender population is highly diverse in terms of both frequency of criminal activity and types of crime committed (Chaiken and Chaiken 1982). A major OR/MS contribution to criminal justice has been in creating succinct models that can characterize both individual criminal behavior and the variation of that behavior across offenders.

Most offenders do not commit crimes according to some deterministic schedule. The exact nature of their crime-generation process is by and large unknown, but it is generally safe to say that the aggregate crime commission by a group of offenders can be modeled by the Poisson distribution. This distribution plays the same role in aggregating point processes that the normal distribution plays in aggregating continuous processes.

In highly influential work, Shinnar and Shinnar (1975) proposed a simple but insightful model of the crime and punishment process. The authors assumed that an active offender commits crimes at a Poisson rate λ per year over a career of length T years. If not arrested, the offender would commit on average λT career crimes. But things change if, like Shinnar and Shinnar, it is assumed that the offender's probability of arrest for each crime is q, that the probability of imprisonment given arrest is Q, and that the average sentence length per prison term is S. If career length T is long relative to sentence length S, then steady-state arguments imply that, under the revised scenario, the offender is free on average for only $1/\lambda qQ$ years between successive imprisonments (Fig. 1). Thus, because of detention, the offender is free and active only for fraction $(1/\lambda qQ)/(1/\lambda qQ + S)$ of the offender's career rather than for all of it. It follows that incapacitation has reduced the offender's total number of offenses by the proportion $S/(S + 1/\lambda qQ)$ compared to the number in a world free of punishment.

There are some gross simplifications in this model (for example, the career length T is assumed independent of the punishment policy in place). But it encapsulates in one equation the effects of all primary elements of the criminal justice system: the offender (via crime commission rate λ), the police (arrest probability q), the courts (chance of imprisonment given arrest, Q), and the correctional system (sentence length S). The model also provides guidance to those exploring empirical data (e.g., offenders' arrest, sentencing, and conviction records) about which quantities were especially worth trying to estimate.

OR/MS professionals like Blumstein and his colleagues worked to flesh out the description of the individual criminal career (Blumstein et al. 1986). They estimated key parameters like the proportion of citizens who participated at some time in criminal behavior, the frequency of crime-commission during the career, the degree to which offenders specialize by crime-type, and the duration of the criminal career.



Crime and Justice, Fig. 1 A (deterministic) criminal career

A simple model of the career might summarize it with four parameters: P, the fraction of individuals in a birth cohort who initiate criminal careers; A, the age of onset for the career; λ, the average annual crime commission rate while free and active, and ρ, the annual probability that the career ends. A macroscopic model could reflect diversity among offenders by assigning a population-wide distribution to each of these parameters, as does the model in Blumstein et al. (1993). (Other distributions about offense type would fill out the description; e.g., Chaiken and Chaiken 1982). Interestingly, offenders who differ greatly on some parameters may be quite similar on others. In a cohort of London multiple-offenders, for example, individuals appear to differ far more in their λ-values than their ρ -values (Barnett et al. 1987). Thus, their career lengths may diverge far less than do intensities of activity during their careers.

A common problem in criminal justice policy analysis has been estimating the effect of enhanced sentences on both crime rates and criminal justice spending. For example, Greenwood et al. (1994) used these Poisson models of criminal offending to predict the consequences of full-scale implementation of California's "Three-Strikes and You're Out" law, and Greenwood et al. (1999) use the model to help identify ways in which actual implementation deviated from that bench-mark.

As is shown in Fig. 1, many offenders continue to commit offenses despite their having been in correctional institutions. But not all do, and the extent of recidivism (commission of additional offenses) is an important concern in criminal justice research. OR/MS researchers have devised, calibrated, and tested probabilistic models that assess the likelihood that given offenders with given past records will again commit crimes within particular future time periods (Stollmack and Harris 1974; Harris et al. 1981; Maltz 1984; Ellerman et al. 1992). These flexible and mathematically-rich techniques allow frequent updating of the prognoses for particular individuals.

The Criminal Justice System (CJS)

With mathematical models, OR/MS professionals described the idea that the CJS – composed of police, courts, and corrections — is, in fact, a system (Government Printing Office 1967b), within which policy shifts in one component generally have consequences for the others. An increase in arrests aimed at reducing crime, for example, can first clog the courts and then overcrowd jails and prisons which, in turn, may be required to reduce surging inmate populations by instituting early release programs for those incarcerated. One of the earliest models to incorporate such feedback effects was the Justice System Interactive Model (JUSSIM), (Belkin et al. 1972); subsequent efforts include Cassidy (1985) and Morgan (1985). JUSSIM has since been updated and software written for personal computers by the U.S. Department of Justice to permit its widespread use (Institute for Law and Justice 1991).

While it is hoped that the CJS provides a fair and cost-effective way to reduce crime, there is a continuing national debate about whether this goal can be achieved. The aims of the CJS are to deter potential offenders from committing crime, to incapacitate those who have been convicted by imprisoning them, and to rehabilitate past offenders so that they are harmless in the future. Of course, the system might induce undesirable changes in criminal behavior, such as brutalization under

which an offender released from prison is more violent than ever before. Statistical investigations by OR/MS researchers have tried to estimate various net effects of the CJS on crime levels (Blumstein et al. 1978, 1986), as well as to assess the realism of specific attempts to estimate such effects from aggregate data (Barnett 1981).

Over the past many years, the number of people under criminal justice supervision in the U.S. has grown dramatically. Although crime rates have fallen, they were rising during much of the build-up in incarceration and, even if increased incarceration were contributing substantially to declining crime rates, it is clearly an expensive way to suppress crime, in both budgetary and human terms. Hence, there has been considerable interest in creating systems models that embrace not only the CJS, but also broader sets of interventions in a manner that allows different classes of interventions to be compared. The goal is to determine whether spending more money on violence prevention or drug treatment programs, for example, might be a more cost-effective way to reduce crime than would be spending more money on prisons and jails.

These analyses show that violence prevention programs are a promising alternative (Greenwood et al. 1996) and that prevention interventions offer a broad array of benefits. They also, however, find that there is a great deal of uncertainty associated with estimates of prevention's cost-effectiveness, even though many, many individual prevention programs have been evaluated (Caulkins et al. 1999). Importantly, the systems framework identifies the sources of the uncertainty and highlights why past evaluations have not been more informative. Some of the reasons are pedestrian, such as never reporting program costs or a focus on showing statistical significance of effects rather than estimating their magnitude. Others are more insightful, such as the fact that traditional evaluations often only consider effects on program participants, even though indirect effects on those not actually in the program are, in some cases, larger in aggregate.

Queueing Models

While everyone recognizes that crime rates vary from neighborhood to neighborhood and by time of day, OR/MS analysts have built probabilistic models that allow exploration of the consequences of such heterogeneity, especially with respect to practical issues of police deployment and staffing of 911 emergency centers (Larson 1972; Kolesar et al. 1975; Chelst 1978). From such models and OR/MS insights into queueing theory have come an unpleasant realization: randomness in the arrival times of calls for service can cause surprisingly large delays in responding to them. Getting six calls randomly distributed over a one-hour period, for example, can yield much slower responses than getting six calls spaced exactly ten minutes apart.

Queueing theory has been applied and extended in developing improved allocation methods for police patrol resources. Such formulations as the hypercube queueing model (Larson 1974; Larson and Odoni 1981; Larson and Rich 1987) and RAND's Patrol Car Allocation Model (Chaiken and Dormont 1978) have depicted with great accuracy how particular police response strategies affect mean response times, workload imbalance across officers, and a host of other performance measures. The models, which are used by many U.S. cities to set police dispatching strategies, allow the user to vary the number of patrol cars and the deployment rules, and then to observe on a computer screen the performance statistics under each scenario. Other OR/MS developments allow the user to set priorities in responding to calls for service and to analyze sending multiple vehicles to incidents (Green and Kolesar 1984). A review of this work is provided in Swersey (1994).

Optimization

Optimization, one of the strongest OR/MS specialties, has played a relatively small role in the profession's contribution to criminal justice. For example, limited success has attended OR/MS efforts to suggest optimal punishment policies. Under particular assumptions about crime-commission processes and their sensitivity to the sentencing strategy in place, Blumstein and Nagin (1978) and Barnett and Lofaso (1986) have worked out optimal allocations of prison space. But the verification of such assumptions — let alone the estimation of key model parameters — has not gone far enough that such models are taken very seriously. Associated attempts to estimate how prison populations vary with changes

in demography and sentencing policy have yielded prison population forecasts that do not immediately demonstrate the practicality of the models (Blumstein et al. 1980; Barnett 1987).

Perhaps the most famous OR/MS proposal for optimal prison sentencing was Greenwood's selective incapacitation scheme in which heavy sentences would be imposed on offenders with at least four of seven high-risk characteristics (Chaiken and Rolph 1980). But data analyses revealed difficulties with implementing such policies (Chaiken and Chaiken 1982; Greenwood and Turner 1987), including a high rate of false positives (people incarcerated to prevent projected future crimes that would never have occurred were they free). These false positives raise controversies about sentencing by conjecture and yield smaller crime-reduction benefits in practice than the strategy can achieve in theory.

A Sense of Ambiguity

Sometimes OR/MS people have contributed to criminal justice research less by what they said than by what they didn't say. OR/MS scholars approach data with a sense of ambiguity: an awareness that a particular empirical pattern is often consistent with a broad range of possibilities. Thus, they have usefully called out "not so fast!" when the most obvious interpretation of certain data was being treated as the only viable one. Four examples of such rescue activities are described below.

One case concerns the Kansas City Preventive Patrol Experiment conducted in the early 1970s. When not responding to calls to service, patrol cars drive randomly through their districts; in theory, such preventive patrol reduce crimes because would-be offenders realize that, even if their victims cannot contact the police, a patrol car might reach the crime scene purely by chance. That theory was called into doubt after Kansas City, in a prearranged experiment, acted to increase preventive patrol sharply in some beats and virtually eliminate it in others. When neither beat-by-beat crime rates nor citizen perceptions about police presence changed visibly during the (unannounced) experiment, some people saw preventive patrol as having lost any rationale.

arson (1975), however, demonstrated with detailed calculations that actual conditions during the Kansas City experiment were quite different from the anticipated ones. Patrol cars from high-activity beats were spending much of their time responding to calls for service from low-activity ones, which had been deprived of all local police vehicles. The upshot was that there was a great deal of police-car movement — often with sirens screaming — in the districts supposedly without preventive patrol, and surprisingly little increase in patrol in the districts supposedly saturated with it. Perhaps, Larson argued, the reason crime rates and citizen perceptions did not change was that police activity itself had not meaningfully changed.

A second example concerned the relationship, well-known to criminologists, between arrests and age. The graph of arrests vs. age is unimodal, reaching a peak in the late teens and then dropping off steadily and sharply. Given this curve, some people argued that it was not cost-effective to give long sentences to offenders convicted at age 30; such offenders, it was contended, were already far less active than at their primes and were unlikely to do much harm even if left on the streets.

But Blumstein et al. (1982) pointed out that such an analysis was vulnerable to a variant of the well-known ecological fallacy: Even if arrests in the aggregate were dropping rapidly with age, it did not follow that individual offenders exhibited this pattern. Having studied longitudinal data about individual offenders, they found that the drop in arrests with age reflected not less activity per year among active offenders, but rather a growing fraction of offenders who had retired from criminal activity. Statistically, an individual convicted at age 30, presumably still active at that age, would be expected, if allowed to go free, to commit as many crimes over the next several years as someone several years younger.

While citizens in the U.S. were debating in the mid-1970s whether to restore the death penalty, several economists came forth with historically-based analyses that purported to weed out extraneous factors and estimate how each execution affects the overall homicide level. The model, whose findings were cited by the U.S. Supreme Court, purported to show that each execution deterred eight homicides. But Barnett (1981) wondered whether the econometric models being used had sufficient explanatory power to fulfill their ambitious goals. Arguing that homicide levels were subject to roughly Poisson-level statistical

noise, he proposed a test of how well the econometric models could forecast state-by-state homicide levels in the very data sets used to calibrate them. The test results indicated that the predictions from all the models suffered large systematic errors of unknown cause and that, indeed, the errors were far larger than any reasonable estimate of the size of the effect the models sought to measure. Thus, Barnett concluded, the analyses were not sensitive enough to answer the question that motivated them.

A final example of ambiguity arises from a study that concluded that juvenile detention acts to reduce delinquency. The study found that an average Illinois male youth sent to a reformatory, though not cured of criminal activity by his stay there, was arrested far fewer times per month after his release than just prior to his detention. The decline was interpreted as the post-release suppression effect on incarceration: getting tough works.

Maltz and Pollock (1980), however, saw another possibility, tied to the phenomenon called regression-to-the-mean. Even if a youth commits crime at a steady frequency and has an unchanging probability of arrest per offense, varying luck in the arrest lottery will cause his observed arrest rate to fluctuate from month to month. But the authorities are especially likely to send him to a reformatory after an upsurge of arrests — that is, at a peak of the fluctuating pattern. Thus, even if the reformatory has no effect on his underlying pattern of criminal behavior, his post-detention arrests would likely fall in frequency compared to his unluckily high predetention levels. Tierney (1983) proposed a revision in their analysis that modified its result, but the work still showed that the suppression effect was quite possibly just an illusion.

These four examples show one of the primary assets of OR/MS thinking as applied to a field so data rich as the CJS. Although data may exhibit certain aggregate patterns, these patterns need not illuminate what is happening at the more detailed level that is, quite often, the appropriate focus of policy analysis. OR/MS analysts should never forget the importance of studying a problem's molecular structure.

Geographical Analyses

Location is a key attribute of crime. Crime is the result of a convergence in time and space of criminals and victims, in the absence of guardians. Land use types, physical geography, the built environment, population characteristics, and police resource allocations and tactics are among the location-based factors influencing crime. Until recently, there was limited capacity for analysts to consider geographical factors explicitly. (Even a panel data set with city-level information is aggregated both in the sense that all events within a city are pooled and in the sense that information about the relative proximity of different cities is usually ignored). The advent and development of geographic information systems (GIS) has changed this situation. GIS is an information technology that geocodes information and processes it spatially in order to facilitate analysis.

Geocoding takes three forms. GIS uses 1) rule bases and scoring to match text of crime incident street addresses with street addresses that already have locations (latitude/longitude coordinates or their flat projections) to place points on maps, 2) global positioning system instruments to read world coordinates of a point, and 3) spatial overlay of map boundary layers (e.g., police precincts, patrol beats, or uniform grid cells) on crime incident points to classify the points' area membership.

Spatial processing includes 1) integrating crime incidents and other factors affecting crime through location, 2) spatial overlay of points and reapportionment of area-based statistics to a consistent crime space/time series (e.g., counts of persons aged 14 to 25 and monthly robberies of persons by spatial grid cell), 3) proximity determination using spatial queries (e.g., all drug arrests within 1000 feet of schools), and 4) connectivity of streets for routing.

The geocoded and processed information can be used for a wide variety of decision support roles. Crime maps for analysis or communication (Maltz et al. 1991; Brantingham and Brantingham 1998) are among the most familiar, but crime forecasting (Foster and Gorr 1986; Gorr and Olligschlaeger 1994; Olligschlaeger 1998) is growing in importance, particularly in conjunction with computer statistics (COMSTAT) police management systems that need counter-factuals against which actual performance can be compared. Likewise, GIS can provide the real time data and detailed information about street networks that is needed to implement effectively some of the OR models described above. Examples

of these applications include patrol resource allocation and the design of administrative boundaries to balance workloads (Koper 1995), as well as queueing applications for routing emergency response (Green 1984; Larson and Rich 1987; Larson and McKnew 1982).

Consensual Crime

Originally the modern application of OR/MS efforts in crime and justice focused on the actions and consequences of the typical street offender who robbed, burglarized, assaulted, or killed strangers. OR/MS methods such as process control charts (Anderson and Diaz 1996), data envelopment analysis (Thanassoulis 1995), and simulation (Larson et al. 1993) continue to be applied in innovative ways to address these problems. But other OR/MS methods have also been finding their way into analyses of consensual crimes including corruption and drug trafficking.

Corruption is widespread among societies and institutions and has stimulated a small but vibrant literature at the intersection of management science and economics. An early standard reference on corruption is Rose-Ackerman (1978) who studied the economics of the supply and demand of bribes. These ideas have been applied to analyses of tax evasion (Chander and Wilde 1992), the distribution of bribes within a hierarchical bureaucracy (Hillman and Katz 1987), and a range of other situations reviewed by Andvig (1991) and Shleifer and Vishny (1993).

Understandably, researchers often approach corruption in game-theoretic terms. Static analyses are common (e.g., Basu et al. 1992; Mookherjee and Png 1995; Marjit and Shi 1998), in part because of their relative tractability, but some of the most exciting developments have involved dynamic optimization. For example, Dawid and Feichtinger (1996a) dynamically extend Akerlof's (1980) model. They find that, unless corruption is the globally dominant strategy, the solution is like that described by a so-called Schelling diagram (Schelling 1973; cf. also Andvig 1991). There are two locally stable equilibria, one where everyone is corrupt and this corruption is accepted, and another where the whole population is honest and corruption is uniformly condemned. The only intermediate equilibrium is unstable.

Likewise, Lui (1986) considers the impact of exogenous corruption deterrence on the (stationary) level of corruption and views anti-corruption campaigns as efforts to shift from an unfavorable to a favorable equilibrium. Feichtinger and Wirl (1994) endogenize these episodes of crusades against corruption. Antoci and Sacco (1995) use replicator dynamics to describe the changing behavior of a population where each individual can decide in each period whether the individual will act honestly or be corrupt. Bicchieri and Rovelli (1995) model the exchange of bribes as a system in which there are two types of players who play a sequence of repeated prisoner's dilemma games with randomly chosen opponents. Wirl et al. (1997) consider interaction between a corrupt politician and an investigative journalist in a differential game and calculate the open-loop Nash equilibrium, which generates interesting insights into the non-cooperative dynamic interaction of crime and enforcement. [Dawid and Feichtinger (1996b) extend the analysis for a similar model to a feedback Nash equilibrium]. Bicchieri and Duffy (1997) demonstrate how corruption can become cyclical under the assumption that politicians, in order to be reelected, have to compensate voters through material incentives.

Whereas corruption has been a problem through the ages, illicit drugs have risen to prominence during the last half of the 20th century. But the OR/MS community and OR/MS tools have already played a prominent role in this new area. Some of the applications have looked specifically at the relationship between drug use and predatory crime (Powers et al. 1991), but many focus explicitly on the production, distribution, sale, and consumption of the drugs themselves. Interdiction activities have received particular attention (Caulkins et al. 1993; Washburn and Wood 1995), perhaps because of the prominent role of the military in that sphere. But production in source countries (Kennedy et al. 1993), domestic distribution networks (Caulkins 1997), managing local enforcement operations (Caulkins 1993a; Naik et al. 1996; Baveja et al. 1997), and drug testing (Lattimore et al. 1996; Meyer and Savory 1997; Kaushal et al. 1998) have also been active research areas.

Just as the Poisson model of offending has been a workhorse in the analysis of predatory offenders, Markov models of drug demand (Everingham et al. 1995) and models of drug markets provide the framework for systems analyses that compare the

effectiveness of different drug control interventions. Typical findings include, for example, that treating heavy users is cost-effective (Rydell et al. 1996), and that mandatory minimum sentences are not (Caulkins 1993b; Caulkins et al. 1997).

The finding that treatment is cost-effective illustrates the importance of choosing the right objective function. Treating heavy users performs miserably if the performance measure is proportion of people treated who are abstinent two years later, but it dominates other available interventions when the performance measure is kilograms of cocaine consumption averted per million dollars spent. True, most people quickly drop out of treatment. They make relapse rates look awful and they contribute nothing to the numerator of a measure such as consumption averted per million dollars. They also, however, do not contribute much to the denominator. The system simply cannot waste that much money on someone who only stays in the program for a few days. Furthermore, relapse measures completely ignore the benefits of reductions in use while someone is in treatment. It turns out that even if 100% of heavy users relapsed, treatment would still be a cost-effective way to reduce drug use just on the basis of the in-treatment effect.

Dynamic models that examine how policies should vary over the course of a drug epidemic are an area of particular interest. Systems dynamics models (Homer 1993) take a descriptive approach to this issue, but an optimal control framework can also yield interesting insights. For example, Tragler et al. (2000) show that detecting the onset of a drug epidemic quickly is valuable because total costs are much lower if control begins early. They also show that sharp price declines, such as those observed in the 1980s for cocaine in the U.S., do not necessarily imply a policy failure; indeed it can be optimal to have such declines. Likewise Behrens et al. (1999) show that it is rarely optimal to advocate greater spending on demand-side interventions generally. Both prevention and treatment can play an important role in drug control, but probably not at the same time. Their comparative advantages come at different stages of an epidemic.

Concluding Remarks

It is not easy to quantify the overall OR/MS contribution to public safety and crime control. OR/MS research has brought about a deeper understanding of the crime problem and how it affects and is affected by the criminal justice system. Crime, however, has such deep psychological, cultural, economic, and social roots that there are limits to what mathematical models can be expected to accomplish on their own. But, likewise, there are limits to what less quantitative perspectives can accomplish on their own. Crime and justice are truly multi-disciplinary problems that are best addressed by multi-disciplinary collaborations, with OR/MS being an integral part of that collaboration.

See

- ▶ Emergency Services
- ▶ Hypercube Queueing Model
- ▶ Program Evaluation
- ▶ Public Policy Analysis
- ▶ Queueing Theory

References

Akerlof, G. A. (1980). A theory of social custom, of which unemployment may be one consequence. *Quarterly Journal of Economics, 94*, 749–775.

Anderson, E. A., & Diaz, J. (1996). Using process control chart techniques to analyze crime rates in Houston, Texas. *Journal of the Operational Research Society, 47*(7), 871–882.

Andvig, J. C. (1991). The economics of corruption: A survey. *Studi Economici, 43*(1), 57–94.

Antoci, A., & Sacco, P. L. (1995). A public contracting evolutionary game with corruption. *Journal of Economics, 61*(2), 89–122.

Barnett, A. (1981). The deterrent effect of capital punishment: A test of some recent studies. *Operations Research, 29*, 346–370.

Barnett, A. (1987). Prison populations: A projection model. *Operations Research, 35*, 18–34.

Barnett, A., Blumstein, A., & Farrington, D. P. (1987). Probabilistic models of youthful criminal careers. *Criminology, 30*, 83–108.

Barnett, A., Essenfeld, E., & Kleitman, D. J. (1980). Urban homicide: Some recent developments. *Journal of Criminal Justice, 8*, 379–385.

Barnett, A., Kleitman, D. J., & Larson, R. C. (1975). On urban homicide: A statistical analysis. *Journal of Criminal Justice, 3*, 85–110.

Barnett, A., & Lofaso, A. J. (1986). On the optimal allocation of prison space. In A. J. Swersey & E. Ignall (Eds.), *Delivery of urban services* (TIMS series in the management sciences, Vol. 22, pp. 249–268). Amsterdam: Elsevier-North Holland.

Barnett, A., & Schwartz, E. (1989). Urban homicide: Still the same. *Journal of Quantitative Criminology, 5*, 83–100.

Basu, K., Bhattacharya, S., & Mishra, A. (1992). Notes on bribery and the control of corruption. *Journal of Public Economics, 48*, 349–359.

Baveja, A., Caulkins, J. P., Liu, W., Batta, R., & Karwan, M. H. (1997). When haste makes sense: Cracking down on street markets for illicit drugs. *Socio-Economic Planning Sciences, 31*, 293–306.

Behrens, D. A., Caulkins, J. P., Tragler, G., & Feichtinger, G. (1999). Optimal control of drug epidemics: Prevent and treat – But not at the same time. *Management Science, 46*, 333–347.

Belkin, J., Blumstein, A., Glass, W., & Lettre, M. (1972). JUSSIM: An interactive computer program and its uses in criminal justice planning. In G. Cooper (Ed.), *Proceedings of international symposium on criminal justice information and statistics systems* (pp. 467–477). Sacramento, CA: Project SEARCH.

Bicchieri, C., & Duffy, J. (1997). Corruption cycles. *Political Studies, 45*, 477–495.

Bicchieri, C., & Rovelli, C. (1995). Evolution and revolution: The dynamics of corruption. *Rationality and Society, 7*, 201–224.

Blumstein, A. (2002). Crime modeling. *Operations Research, 50*(1), 16–24.

Blumstein, A. (2007). An OR missionary's visits to the criminal justice system. *Operations Research, 55*(1), 14–23.

Blumstein, A., Canela-Cacho, J. A., & Cohen, J. (1993). Filtered sampling from populations with heterogeneous event frequencies. *Management Science, 39*, 886–899.

Blumstein, A., Cohen, J., & Hsieh, P. (1982). *The durations of adult criminal careers. Final report to national institute of justice*. Pittsburgh, PA: Carnegie-Mellon University.

Blumstein, A., Cohen, J., & Miller, H. (1980). Demographically disaggregated projections of prison populations. *Journal of Criminal Justice, 8*, 1–25.

Blumstein, A., Cohen, J., & Nagin, D. (Eds.). (1978). *Deterrence and incapacitation: Estimating the effects of criminal sanctions on crime rates*. Washington, DC: National Academy of Sciences.

Blumstein, A., Cohen, J., Roth, J. A., & Visher, C. (1986). *Criminal careers and "career criminals." vols. I and II*. Washington, DC: National Academy of Sciences.

Blumstein, A., & Larson, R. (1969). Models of a total criminal justice system. *Operations Research, 17*(2), 199–232.

Blumstein, A., & Nagin, D. (1978). On the optimum use of incarceration for crime control. *Operations Research, 26*, 383–405.

Bonczar, T. P., & Beck, A. J. (1997). *Lifetime likelihood of going to state or federal prison*. Washington, DC: National Institute of Justice.

Brantingham, P. L., & Brantingham, P. J. (1998). Mapping crime for analytic purposes: Location quotients, counts, and rates. In D. Weisburd & T. McEwen (Eds.), *Crime mapping, crime prevention, crime prevention studies 8*. New York: Criminal Justice Press.

Cassidy, R. G. (1985). Modelling a criminal justice system. In D. P. Farrington & R. Tarling (Eds.), *Prediction in criminology*. Albany, NY: State University of New York Press.

Caulkins, J. (1993a). Zero-tolerance policies: Do they inhibit or stimulate illicit drug consumption? *Management Science, 39*, 458–476.

Caulkins, J. (1993b). Local drug markets' response to focused police enforcement. *Operations Research, 41*, 843–863.

Caulkins, J. P. (1997). Modeling the domestic distribution network for illicit drugs. *Management Science, 43*, 1364–1371.

Caulkins, J. P., Crawford, G., & Reuter, P. (1993). Simulation of adaptive response: A model of drug interdiction. *Mathematical and Computer Modelling, 17*(2), 37–52.

Caulkins, J. P., Rydell, C. P., Everingham, S. S., Chiesa, J., & Bushway, S. (1999). *An ounce of prevention, a pound of uncertainty: The cost-effectiveness of school-based drug prevention program* (Technical report). Santa Monica, CA: RAND Corporation.

Caulkins, J. P., Rydell, C. P., Schwabe, W. L., and Chiesa, J. (1997). Mandatory minimum drug sentences: Throwing away the key or the taxpayers' money? (Report MR-827-DPRC). Santa Monica, CA: RAND Corporation.

Chaiken, J. M., & Chaiken, M. R. (1982). *Varieties of criminal behavior* (Report R-2814-NIJ). Santa Monica, CA: Rand Corporation.

Chaiken, J. M., & Dormont, P. (1978). A patrol car allocation model: Background, capabilities, and algorithms. *Management Science, 24*, 1280–1300.

Chaiken, J. M., & Rolph, J. (1980). Selective incapacitation strategies based on estimated crime rates. *Operations Research, 28*, 1259–1274.

Chander, P., & Wilde, L. (1992). Corruption in tax administration. *Journal of Public Economics, 49*, 333–349.

Chelst, K. (1978). An algorithm for deploying a crime-directed (tactical) patrol force. *Management Science, 24*, 1314–1327.

Cormican, K. J., Morton, D. P., & Wood, R. K. (1998). Stochastic network interdiction. *Operations Research, 46*, 184–197.

Daston, L. (1988). *Classical probability in the enlightenment*. Princeton, NJ: Princeton University Press.

Dawid, H., & Feichtinger, G. (1996a). On the persistence of corruption. *Journal of Economics, 64*(2), 177–193.

Dawid, H., & Feichtinger, G. (1996b). Optimal allocation of drug control efforts: A differential game analysis. *Journal of Optimization Theory and Applications, 91*, 279–297.

Ellerman, P., Sullo, P., & Tien, J. M. (1992). An alternative approach to modeling recidivism using quantile residual life functions. *Operations Research, 40*, 485–504.

Everingham, S., Rydell, C. P., & Caulkins, J. P. (1995). Cocaine consumption in the US: Estimating past trends and future scenarios. *Socio-Economic Planning Sciences, 29*, 305–314.

Feichtinger, G., & Wirl, F. (1994). On the stability and potential cyclicity of corruption in governments subject to popularity constraints. *Mathematical Social Sciences, 28*, 113–131.

Foster, S. A., & Gorr, W. L. (1986). An adaptive filter for estimating spatially-varying parameters: Application to modeling police hours spent in response to calls for service. *Management Science, 32*, 878–889.

Gorr, W. L., & Olligschlaeger, A. M. (1994). Weighted spatial adaptive filtering: Monte Carlo studies and application to illicit drug market modeling. *Geographical Analysis, 26*(1), 67–87.

Government Printing Office. (1967a). *The challenge of crime in a free society*. Washington, DC: President's Commission on Law Enforcement and Administration of Justice.

Government Printing Office. (1967b). *Task force report: Science and technology*. Washington, DC: President's Commission on Law Enforcement and Administration of Justice.

Green, L. (1984). A multiple dispatch queueing model of police patrol operations. *Management Science, 30*, 653–664.

Green, L., & Kolesar, P. (1984). A comparison of multiple dispatch and M/M/C priority queueing models of police patrol. *Management Science, 30*, 665–670.

Greene, M. A. (1984). Estimating the size of the criminal population using an open population approach. *Proceedings American Statistical Association*, *Survey Methods Research Section*, pp. 8–13.

Greene, M. A., & Stollmack, S. (1981). Estimating the number of criminals. In J. A. Fox (Ed.), *Models in quantitative criminology* (pp. 1–24). New York: Academic.

Greenwood, P. W., & Abrahamse, A. F. (1981). *Selective incapacitation* (Report R-2815-NIJ). Santa Monica, CA: Rand Corporation.

Greenwood, P. W., Everingham, S. S., Chen, E., Abrahamse, A. F., Merritt, N., & Chiesa, J. (1999). *Three strikes revisited*: *An early assessment of implementation and effects* (Technical report). Santa Monica, CA: RAND Corporation.

Greenwood, P. W., Model, K. E., Rydell, C. P., & Chiesa, J. (1996). *Diverting children from a life of crime*: *Measuring the costs and benefits* (Report MJ-699-UCB/RC/F). Santa Monica, CA: RAND Corporation.

Greenwood, P. W., Rydell, C. P., Abrahamse, A. F., Caulkins, J. P., Chiesa, J. R., Model, K. E., & Klein, S. P. (1994). *Three strikes and you're out*: *Estimated benefits and costs of California's new mandatory-sentencing law* (Report MR-509-RC). Santa Monica, CA: RAND Corporation.

Greenwood, P. W., & Turner, S. (1987). *Selective incapacitation revisited*: *Why the high-rate offenders are hard to predict* (Report R-3397-NIJ). Santa Monica, CA: Rand Corporation.

Hacking, I. (1990). *The taming of chance*. England: Cambridge University Press.

Harris, C. M., Kaylan, A. R., & Maltz, M. D. (1981). Recent advances in the statistics of recidivism measurement. In J. A. Fox (Ed.), *Models of quantitative criminology* (pp. 61–79). New York: Academic Press.

Hillman, A. L., & Katz, E. (1987). Hierarchical structure and the social costs of bribes and transfers. *Journal of Public Economics, 34*, 129–142.

Homer, J. B. (1993). A system dynamics model of national cocaine prevalence. *System Dynamics Review, 9*(1), 49–78.

Institute for Law and Justice. (1991). *CJSSIM: Criminal justice system simulation model: Software and user manual*. Alexandria, VA: Institute for Law and Justice.

Karoly, L. A., Greenwood, P. W., Everingham, S. S., Houbé, J., Kilburn, M. R., Rydell, C. P., Sanders, M., & Chiesa, J. (1998). *Investing in our children*: *What we know and don't know About the costs and benefits of early childhood*, *interventions* (MR-898-TCWF). Santa Monica, CA: RAND Corporation.

Kaushal, C., Baker, J. R., & Lattimore, P. K. (1998). A decision support system for partial drug testing. *Decision Support Systems, 23*, 241–257.

Kennedy, M., Reuter, P., & Riley, K. J. (1993). A simple economic model of cocaine production. *Mathematical and Computer Modelling, 12*(2), 19–36.

Kolesar, P. J., Rider, K. L., Crabill, T. B., & Walker, W. W. (1975). A queuing-linear programming approach to scheduling police patrol cars. *Operations Research, 23*, 1045–1062.

Koper, C. S. (1995). Just enough police presence: Reducing crime and disorderly behavior by optimizing patrol time in crime hot spots. *Justice Quarterly, 12*, 649–672.

Larson, R. C. (1972). *Urban police patrol analysis*. Cambridge, MA: MIT Press.

Larson, R. C. (1974). A hypercube queueing model for facility location and redistricting in urban emergency services. *Journal of Computers and Operations Research, 1*, 67–95.

Larson, R. C. (1975). What happened to patrol operations in Kansas city?: A review of the Kansas city preventive patrol experiment. *Journal of Criminal Justice, 3*, 267–297.

Larson, R. C., Cahn, M. F., & Shell, M. C. (1993). Improving the New York city arrest-to-arraignment system. *Interfaces, 23*(1), 76–96.

Larson, R. C., & McKnew, M. A. (1982). Police patrol-initiated activities within a systems queueing model. *Management Science, 28*, 759–774.

Larson, R. C., & Odoni, A. R. (1981). The hypercube queueing model. In *Urban operations research* (pp. 292–335). Englewood Cliffs, NJ: Prentice-Hall.

Larson, R. C., & Rich, T. (1987). Travel time analysis of New York city police patrol cars. *Interfaces, 17*(2), 15–20.

Lattimore, P. K., Baker, J. R., & Matheson, L. A. (1996). Monitoring drug use using Bayesian acceptance sampling: The Illinois experiment. *Operations Research, 44*, 274–285.

Lui, F. T. (1986). A dynamical model of corruption deterrence. *Journal of Public Economics, 31*, 215–236.

Maltz, M. D. (1984). *Recidivism*. Orlando, FL: Academic Press.

Maltz, M. D. (1994). Operations research in studying crime and justice: Its history and accomplishments. In S. M. Pollock, A. Barnett, & M. Rothkopf (Eds.), *Operations research and public systems*. Amsterdam: Elsevier.

Maltz, M. D., Gordon, A. C., & Friedman, W. (1991). *Mapping crime in its community setting: Event geography analysis*. New York: Springer.

Maltz, M. D., & Pollock, S. M. (1980). Artificial inflation of a delinquency rate by a selection artifact. *Operations Research, 28*, 547–559.

Marjit, S., & Shi, H. L. (1998). On controlling crime with corrupt officials. *Journal of Economic Behavior and Organization, 34*(1), 163–172.

Meyer, J. L., & Savory, P. A. (1997). Selecting employees for random drug tests at union pacific railroad. *Interfaces, 27*(5), 58–67.

Mookherjee, D., & Png, I. P. L. (1995). Corruptible law enforcers: How should they be compensated? *The Economic Journal, 105*, 145–159.

Morgan, P. M. (1985). *Modelling the criminal justice system*. Home office research and planning unit paper 35. London: Home Office.

Naik, A. V., Baveja, A., Batta, R., & Caulkins, J. P. (1996). Scheduling crackdowns on illicit drug markets. *European Journal of Operational Research, 88*, 231–250.

Olligschlaeger, A. M. (1998). Artificial neural networks and crime mapping. In D. Weisburd & T. McEwen (Eds.), *Crime mapping, crime prevention, crime prevention studies 8*. New York: Criminal Justice Press.

Powers, K., Hanssens, D. M., & Hser, Y. I. (1991). Measuring the long-term effects of public policy: The case of narcotics use and property crime. *Management Science, 37*, 627–644.
Rose-Ackerman, S. (1978). *Corruption: A study in political economy*. New York: Academic Press.
Rydell, C. P., Caulkins, J. P., & Everingham, S. (1996). Enforcement or treatment: Modeling the relative efficacy of alternatives for controlling cocaine. *Operations Research, 44*, 687–695.
Schelling, T. C. (1973). Hockey helmets, concealed weapons, and daylight saving: A study of binary choices with externalities. *Journal of Conflict Resolution, 17*, 381–428.
Shinnar, R., & Shinnar, S. (1975). The effects of the criminal justice system on the control of crime: A quantitative approach. *Law and Society Review, 9*, 581–611.
Shleifer, A., & Vishny, R. W. (1993). Corruption. *Quarterly Journal of Economics, 108*, 599–617.
STIF: Science and Technology Task Force. (1967). *Task force report: Science and technology. President's commission on law enforcement and the administration of justice*. Washington, DC: US Government Printing Office.
Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Cambridge, MA: The Belknap Press of Harvard University Press.
Stollmack, S., & Harris, C. (1974). Failure-rate analysis applied to recidivism data. *Operations Research, 22*, 1192–1205.
Swersey, A. J. (1994). The deployment of police, fire, and emergency medical units. In S. M. Pollock, A. Barnett, & M. Rothkopf (Eds.), *Operations research and public systems*. Amsterdam: Elsevier.
Thanassoulis, E. (1995). Assessing police forces in England and Wales using data envelopment analysis. *European Journal of Operational Research, 87*, 641–658.
Tierney, L. (1983). A selection artifact in delinquency data revisited. *Operations Research, 31*, 852–865.
Tragler, G., Caulkins, J. P., & Feichtinger, G. (2000). Optimal dynamic allocation of treatment and enforcement in illicit drug control. *Operations Research, 49*, 352–362.
Washburn, A., & Wood, K. (1995). Two-person zero-sum games for network interdiction. *Operations Research, 43*, 243–252.
Wirl, F., Novak, A., Feichtinger, G., & Dawid, H. (1997). Indeterminacy of open-loop Nash equilibria: The ruling class versus the tabloid press. In H. G. Natke & Y. Ben-Haim (Eds.), *Uncertainty: Models and measures* (pp. 124–136). Berlin: Akademie-Verlag.

Criterion Cone

► Multiobjective Programming

Criterion Space

► Multiobjective Programming

Criterion Vector

► Multiobjective Programming

Critical Activity

A project work item on the critical path having zero float time.

See

► Critical Path
► Critical Path Method (CPM)
► Network Planning

Critical Path

The longest continuous path of activities through a project network from beginning to end. The total time elapsed on the critical path is the shortest duration of the project. The critical path will have zero float time, if a date for completion has not been specified. Any delay of activities on the critical path will cause a corresponding delay in the completion of the project. It is possible to have more than one critical path.

See

► Critical Path Method (CPM)
► Network Planning

Critical Path Method (CPM)

A project planning technique that is used for developing strategy and schedules for an undertaking using a single-time estimate for each activity of which the project is comprised. In its basic form, i.e., concerned with determining the critical path, that is, the longest sequence of activities through the project

network from beginning to end. CPM arose from a jointly sponsored venture of E.I. du Pont de Nemours and Company and the Sperry-Rand Corporation (Kelley 1961).

See

- ▶ Network Planning
- ▶ Program Evaluation and Review Technique (PERT)
- ▶ Project Management

References

Kelley, J. E. (1961). Critical-path planning and scheduling: Mathematical basis. *Operations Research, 9*, 296–320.csm.

Critical Systems Thinking

Werner Ulrich
University of Fribourg, Fribourg, Switzerland
The Open University, Milton Keynes, UK

Introduction

Critical systems thinking (CST) is a development of systems thinking that aims to support good practice of all forms of applied systems thinking and professional intervention. In its simplest definition, CST is applied systems thinking in the service of good practice. Three essential ideas are as follows:

1. Professional practice in all its stages and activities, from the formulation of problems to the implementation of solutions and the evaluation of outcomes, involves choices that need to be made transparent and require systematic examination and validation.
2. Systems thinking, although it does not protect against the need for such choices, at least offers a methodological basis for examining them systematically.
3. Consequently, applied systems thinking should make it standard practice to employ not only a hard (quantitative, scientific) and/or a soft (qualitative, interpretive) but always also a systematically critical (reflective, questioning) perspective and mode of analysis.

Taking these three elements together, CST not only recognizes that all applied systems thinking involves choices in need of critical reflection but also draws on systems thinking itself as a source of systematic critical reflection and deliberation.

CST and OR

Critical systems thinking has essential roots in operations research and management science (OR/MS), along with some equally important roots in philosophy, social theory, and other disciplines. It has applications in OR/MS as well as in many other professional fields that it is increasingly influencing; among them are environmental planning and management, public policy analysis, information systems design, social planning, evaluation research, technology assessment and risk regulation, and others. Unlike most of these fields, OR/MS was from the outset conceived as applied systems thinking; its systems perspective was to distinguish it from conventional notions of applied science and professional intervention. Critical systems thinking may be understood as an expansion of that original idea. CST's focus is on the fundamental theoretical and normative assumptions that inform the formulation and analysis of problems within their contexts, rather than on the more technical aspects of model building, analysis, and validation, or on procedural aspects of project management and consensus formation.

Two Main Sources of CST Within OR/MS

Critical systems thinking developed from the confluence of two largely independent strands of thought about OR practice. The first strand originated in the 1970s at the University of California at Berkeley and can be regarded as a development of, and response to, Churchman's (1968, 1971, 1979) philosophy of social systems design, which itself was a development of his earlier pioneering work on OR/MS (Churchman et al. 1957). The second strand originated in the 1980s at the University of Hull in England and can be regarded as a response to the development, in British OR, of soft systems methodology (Checkland 1981, 1985; Checkland and Scholes 1990), along with a number of soft OR methods or problem structuring methods (Rosenhead 1989) and

some other approaches to complex and dynamic problem contexts (e.g., management cybernetics and viable systems diagnosis, Beer 1972, 1985), all of which not only led to a growing variety of methods and underlying research paradigms but also to a perception of paradigmatic insecurity or crisis in parts of the OR profession.

Two Key Issues of Critical Systems Thinking

CST responded to these developments in American and British OR/MS by focusing its methodological efforts on two key issues:

- The first issue emerged from recognizing that the way professionals understand and define problem contexts has value implications, in the practical sense that it may do more or less justice to the different views and needs of people. Professional practice cannot avoid, in every specific context of intervention, choices as to what views (observations, data) and what needs (concerns, interests) of people are to be considered relevant and what other views and needs should not or cannot be considered equally relevant. The question is: "What should constitute the basis of knowledge and values for rational practice?" When it comes to this normative core of practice, there is a need to support professionals and everyone else concerned in handling their assumptions in a transparent and self-critical way, as well as to deal adequately with the consequences these assumptions may have for the different parties concerned.
- The second issue emerged from recognizing that different problem situations put different demands on professional competence and accordingly also on the methods professionals use. Professional practice cannot avoid, in defining and employing its methods of analysis and intervention, assumptions about the nature of problem situations, particularly with respect to the kind of complexity that matters; for real-world complexity takes different forms and there is consequently no single best way to understand and handle it. The question is: "What are the assumptions, strengths, and weaknesses of different approaches and methods regarding the nature of problem contexts, that is, different kinds of social reality? When it comes to the variety of methodological options available today in applied systems thinking, there is a need to support professionals in handling these options in a theoretically informed and justifiable way.

Critical systems thinking, then, is the use of systems ideas for probing into these two different (though not entirely independent) sources of contextual selectivity, that is, assumptions that shape the understanding and handling of problem contexts—the selection of relevant facts and values, and the selection of adequate methodologies and methods. Both shape the way problems will be understood within their contexts. However, they place rather different demands on good practice. What assumptions different systems approaches make regarding the nature and complexity of problem contexts depends on their theoretical underpinnings and thus can be identified theoretically once and for all; good practice in this respect means informed methodology choice. By contrast, relevant facts and values need to be identified anew in each specific problem situation and therefore are mainly a responsibility of practice itself; good practice in this respect means reflective practice.

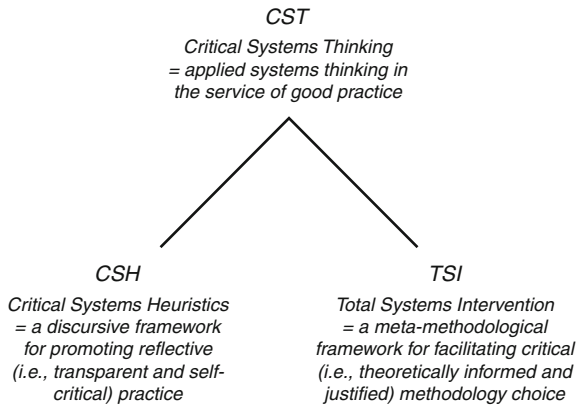
Two different strands of critical systems thinking have accordingly developed: critical systems heuristics (CSH) and total systems intervention (TSI). Their shared core idea is that systems thinking can be a useful source of critical reflection about contextual selectivity. A precise yet comprehensive definition of CST may therefore be formulated as follows.

Definition

Critical systems thinking (CST) is an application of systems thinking that aims to support good practice with regard to (a) the normative core of the knowledge and value basis that informs professional findings and conclusions and (b) the theoretical assumptions that inform the variety of methodologies and methods employed. The common denominator of (a) and (b) is that they both condition the perception of relevant problem contexts.

Terminology: CST, CSH, and TSI

The term "critical systems thinking" was coined in July 1989, when the originators of the two strands met at the 33rd Annual Conference of the International Society for the Systems Sciences (ISSS) in Edinburgh, Scotland, and decided to unite their efforts under the umbrella of critical systems



Critical Systems Thinking, Fig. 1 Critical systems thinking (CST) and its two strands—basic terminology (Source: Adapted from Ulrich 2003, 327)

thinking. Due to differing methodological conceptions and philosophical backgrounds, the cooperation between the two strands of CST remained a brief episode in the late 1980s and early 1990s; but the term CST has survived as a name for their shared interest in handling contextual assumptions critically.

Some confusion was subsequently caused by the circumstance that both strands have continued to refer to their efforts as critical systems thinking. For the sake of terminological clarity, it is advisable to use the term as a higher-level concept under which CSH and TSI may meaningfully be subsumed, rather than identifying it with either strand (Fig. 1).

Due to their separate development and also to different theoretical foundations, the two strands, despite their shared core idea and complementary ends, have brought forth partly incompatible frameworks for CST. They are therefore introduced separately. However, to facilitate comparison and synthesis, the account follows the same structure and uses the same criteria.

Critical Systems Heuristics (CSH): Facing the Normative Core of Professional Practice

CSH was fully worked out in the late 1970s at the University of California at Berkeley but became widely known only in the early 1980s, when the main theoretical work (Ulrich 1983) was published with some delay after the author's return to Switzerland. With a view to submitting his work to the test of practice, Ulrich assumed a position as chief policy analyst and evaluation researcher in the public sector and also returned to teaching at his home university, the University of Fribourg (Philosophical Faculty). This double experience in public policy making and university teaching has helped Ulrich to develop CSH continuously since. CSH has meanwhile found resonance and applications in many applied disciplines and is gradually evolving into a more comprehensive framework for reflective practice in the civil society (Ulrich 2000), critical pragmatism (Ulrich 2006 and 2007), and philosophy for professionals (Ulrich 2007).

Core Idea

Professional practice involves validity claims (e.g., to truth, rightness, sincerity, objectivity, rationality, and relevance) that have practical consequences but which it cannot fully justify. Applied systems thinking makes no exception, for its effort to appreciate the systemic nature of problems, and thus to gain a comprehensive or whole-systems view of problem situations, does not supersede the need for making value judgments as to what exactly is to be considered the problem to be dealt with (i.e., what merits improvement), what constitutes the relevant problem context (i.e., what is the sum total of the relevant facts and concerns), and wherein would consist a good solution (i.e., how to define improvement). No kind of systems methodology or other methodology can fully justify the answers to such inevitable questions as "whose problem is to be solved in the first place?" and "for whom should improvement be achieved and for whom should it not?" What is possible, however, is a conscious and careful handling of this normative core of all professional intervention.

Critical systems thinking as understood in CSH therefore begins with the idea that holistic or whole-systems thinking—the quest for comprehensiveness—is a meaningful effort but not a meaningful claim. Doing full and equal justice to the views and values of all the people concerned is an ideal; but applied systems thinking should not be expected to achieve ideals. To put it differently, holism is not a philosophically and methodologically credible source of justification, it is a problem. Hence, rather than trying to be holistic, CSH tries to support practice—professionals as well as ordinary citizens—in

appreciating the inevitable selectivity of the claims involved (e.g., to putting a problem well and to securing improvement) with regard to the facts (observations) and values (concerns) it takes to be relevant and on which its rationality and consequences depend.

In practical contexts of action, selectivity usually translates into partiality in the sense that different parties will be affected differently. CSH consequently also aims to help professionals and citizens in analyzing these consequences and how they may change if assumptions about relevant observations and concerns are modified. Good practice cannot avoid selectivity and partiality, but it will make it transparent to all those concerned how the selectivity of assumptions and the partiality of consequences depend on one another. It will give all the parties an opportunity to articulate their critique, and will then try to modify assumptions and consequences accordingly. Critical systems thinking, thus understood, is reflective practice—a methodologically disciplined effort to support such processes of critique systematically.

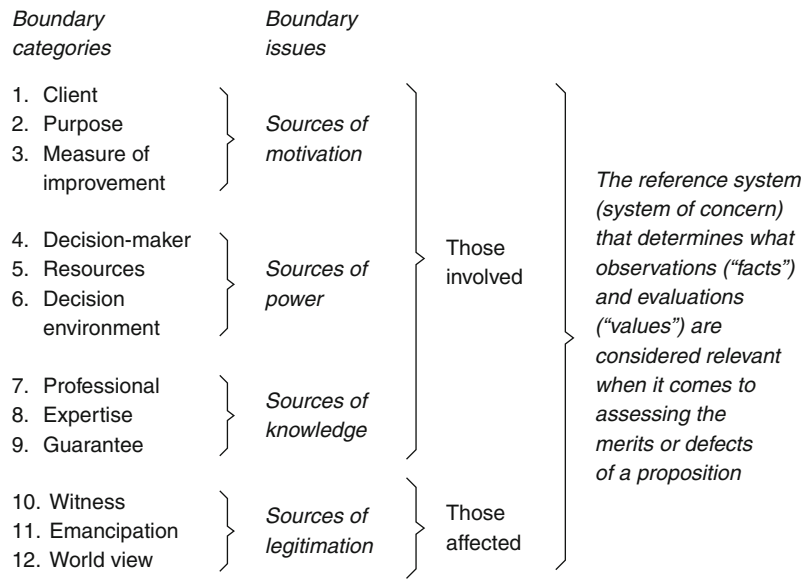
Methodological Approach

CSH is both a new philosophical foundation and a practical implementation of a discursive framework for value clarification and critique. Like the previously used concept of the normative core of rational practice, the term "value clarification" again refers to the selectivity of both considerations of facts (the empirical or knowledge basis of rational action) and of values (the normative or value basis of rational action) in contexts of practical action. The choice of the knowledge basis of professional interventions—of relevant data, judgments of fact, personal views, and other empirical conjectures (e.g., anticipated consequences of action)—has no less normative implications than does the choice of its value basis, that is, of relevant concerns, notions of improvement, and ethical standards. Both sources of selectivity and partiality demand a critical handling.

But applied systems thinking not only implies empirical and normative selectivity, it also holds a key to handling such selectivity critically. Systems thinking compels professionals, as well as everyone else concerned, to pay attention to the systems boundaries that delimit any specific system of interest. Systems thinking can thus be understood as a tool for reflecting about the boundaries of concern that (consciously or not) inform all analysis of problems and related proposals and arguments, regardless of whether systems terms are used in the first place or others. Systems thinking then becomes a source of critique—of questioning boundary assumptions and the ways they condition validity claims—rather than, as it is more usually understood, a source of justification, that is, a way of buttressing validity claims by more comprehensive considerations of fact and value.

In the terms of CSH, critical systems thinking can support professionals and all the parties concerned in identifying and questioning boundary judgments that delimit the reference systems for defining problems and relevant contexts, solution designs, evaluations, proposals for action, and so on. Boundary judgments determine for a number of basic boundary issues and related boundary categories what is to be considered and what is to be left out when it comes to defining relevant observations (judgments of fact) and concerns (judgments of value). A reference system is the set of boundary judgments that together define the context of application which a specific claim or proposal refers to and for which it is valid.

Boundary judgments are the perfect device for questioning the relevance and quality of reference systems; for unlike what one might assume at first glance, they define not just the scope of the context considered (i.e., how narrow or comprehensive it is delimitated) but equally its content, that is, what observations about that context are collected and taken to be relevant; how they are formulated, interpreted, and used; what importance is attached to them; and how well related conjectures are argued. This is so because any aspects of a problem situation that are not properly considered, say, because those involved argue incoherently or anticipate consequences incorrectly, or fail to do justice to the concerns of others, have in fact been excluded from the relevant knowledge and value basis. Even if one recognizes some aspects as relevant and agrees with others they should be considered but then fails to take them properly into account, due to lacking knowledge, to an error of judgment, or some communicative misunderstanding, or because those in control of the situation decide to suppress their discussion, the aspects are in fact (deliberately or not) excluded from the considered reference system. Thus the



Critical Systems Thinking, Fig. 2 Boundary categories of critical systems heuristics (Source: Ulrich 1983, 258)

argumentative quality of a validity claim or related discussion very well reflects itself in boundary judgments.

The main device to promote such argumentative quality is critical systems discourse, a dialogical form of boundary critique. Boundary critique is basically a systematic process of identifying the boundary judgments that are built into any specific validity claims, in an effort to unfold their normative core (selectivity) and what it may mean for the parties concerned (partiality). A second basic aim is to show that there are always options for defining boundary judgments, and to make it visible how different the claims in question may look in the light of such options. In cooperative settings where the parties are prepared to try and see whether they can agree on their boundary judgments, these can be modified accordingly. In controversial settings this may not be possible; boundary critique then gains a new meaning and consists in employing boundary judgments for critical purposes against those who are not prepared to disclose and question them or who even try to impose them on the basis of authority and power rather than argumentation. Critical systems discourse thus becomes a discursive process of challenging validity claims by demonstrating that and how they depend on boundary judgments that have not been declared or are imposed by nonargumentative means.

To be sure, selectivity, not comprehensiveness, is the fate of everyone who tries to solve problems and to do something about the state of the world. The point of boundary critique consists, in terms of CSH, in a critical turn of applied systems thinking and its notion of good professional practice. It recognizes that there is no objective but only a critical solution to the fundamental problem of practical reason, of how claims to rational practice can be justified in the face of their inevitable selectivity and partiality. The problem has remained unresolved in practical philosophy, the philosophical discipline concerned with the normative dimension of rational action, in that no theoretical solutions have been found that would at the same time be practicable. (A more complete account of the concept of a critical solution is given in Ulrich (1983, 2001, 2003).)

Methodological Core Principle

CSH's answer to the unresolved problem of practical reason is the principle of boundary critique. It says that both the meaning and the validity of claims depend on the reference system which these claims refer to and, hence, that one cannot understand and qualify (appreciate and criticize) their adequacy without examining the boundary judgments that define that reference system. The basic idea and aim of CSH, then, is to support systematic processes of boundary critique as a way to secure at least a critical solution of the problem of practical reason. To this end, there are 12 CSH boundary categories (Fig. 2).

These boundary categories stand for four crucial sources of selectivity built into all practice. Each boundary category translates into two boundary questions: one asking what is the case ("is" mapping, i.e., descriptive analysis) and the other what should be the case ("ought" mapping, i.e., normative analysis). This yields an extensive checklist of boundary questions that explicitly define the precise intent of each boundary category (Ulrich 1987, 1996, 2000; Ulrich and Reynolds 2010). They can be used, first, to identify boundary judgments systematically; second, to examine how alternative boundary judgments may change the way one sees problem definitions, findings, and conclusions, and thus what is considered to be adequate and rational; and third, to challenge any claims to knowledge, rationality, or improvement that rely on hidden boundary judgments or take them for granted.

The last-mentioned application leads to an argumentative employment of boundary judgments, known as polemical or emancipatory boundary critique, that creates an improved symmetry of critical competence among all the parties concerned, professionals and citizens alike, regardless of their theoretical knowledge or special expertise with respect to the problem at issue. As a practicable model of cogent critical argumentation (Ulrich 1983, 1993, 2000), it embodies a critical pragmatization of Habermas' (1973, 1979) ideal model of rational practical discourse (a model that underpins his discourse ethics and confines it to being a moral theory rather than a practicable model of moral justification). It constitutes a chief methodological backing of the critical turn of the concept of rational practice proposed above.

In sum, CSH can be defined as a methodological framework for boundary critique, that is, for identifying and debating boundary judgments, with the aim of securing at least a critical solution to the unsolved problem of practical reason—the question of how claims to rational practice can be justified despite the unavoidable selectivity and partiality of all practice. Despite its emancipatory implications (the aspect for which it is best known), CSH should not be misunderstood and used as an emancipatory systems approach only; its principle of systematic boundary critique is vital for sound professional practice in general, whatever importance may be attached to emancipatory issues. For the same reason, CSH does not aim to be a self-contained systems methodology, but is better understood as an approach that should inform all critical professional practice, whatever specific methodology is used.

Practical Implementation (Main Procedure)

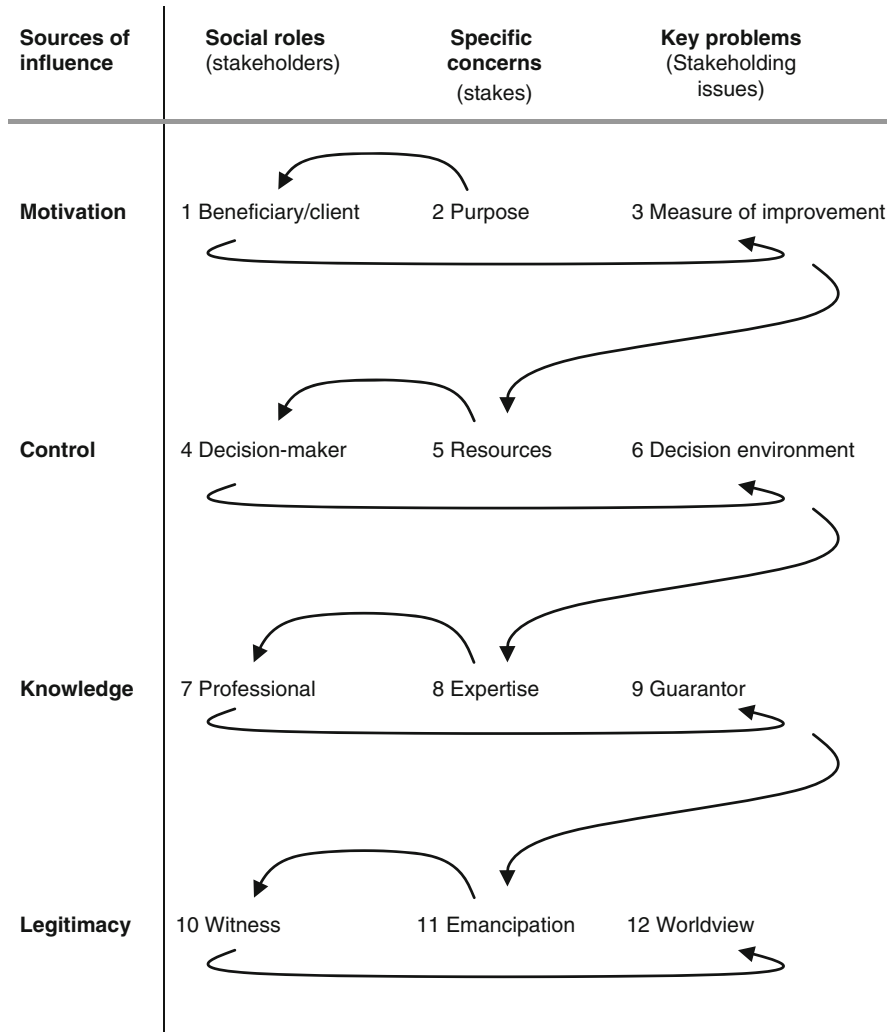
Boundary critique is best implemented as an iterative process of reflecting on, and discussing, the implications of alternative boundary judgments. When some boundary judgment changes, the reference system of which it is constitutive will change too; consequently, all other boundary judgments may need being reconsidered and adapted. However, iterative processes are not particularly easy to learn and to handle; experience with boundary critique suggests that it is useful for beginners to have available, and follow, a standard sequence for unfolding the boundary categories and questions of CSH (Fig. 3).

Total Systems Intervention (TSI) or Creative Holism (CH): Ensuring Informed Methodology Choice

TSI stems from work done at the University of Hull, England, in the mid and late 1980s, about the evolution of OR and systems thinking in terms of changing underlying theoretical assumptions. This work resulted in the early 1990s in the proposal of a meta-methodology for choosing among methodologies according to situational requirements (Flood and Jackson 1991; Jackson 1991). By that time CSH and TSI had joined their efforts under the new name of "critical systems thinking" (CST), after previously using different names such as critical systems approach (CSH) and critical management science (TSI). But due to differing notions of what critical practice was to mean, the two strands of CST ultimately found it difficult to integrate their approaches and consequently returned to developing their frameworks separately. Both have nevertheless continued to use the name critical systems thinking. Meanwhile, Jackson (2003; 2006b) refers to his work on critical systems thinking and practice as creative holism (CH).

Core Idea

Applied systems thinking depends for its choice of systems methodologies and methods on basic assumptions regarding the nature of the problem contexts (typically: organizational contexts) with



Critical Systems Thinking, Fig. 3 CSH's process of unfolding: a standard sequence of boundary critique (Source: Ulrich and Reynolds 2010, 259; Adapted from Reynolds 2007, 106)

which it is dealing. Some of these assumptions can usefully be captured in terms of a number of sociological paradigms for describing the nature of social reality as they have been analyzed, for example, by Burrell and Morgan (1979), as well as by organizational images or systems metaphors as they have been described most systematically by Morgan (1986). Different systems methodologies, because they usually are developed with different problem contexts in mind, can similarly be characterized in terms of underlying metaphors and paradigms. Hence, since the characteristics of both problem contexts and systems methodologies can be captured in terms of adequate paradigms and metaphors, it becomes possible to match contexts and methodologies in a systematic way and thus to support professionals in choosing among the increasing number of available systems methodologies and conforming methods that are best suited to deal with a problem situation at hand.

CST as understood in TSI/CH therefore begins with the idea that systems thinking—the attempt to understand organizational or societal problem contexts in systems terms—is meaningful only to the extent people are aware of the sociological paradigms and organizational metaphors that inform it. Since different systems methodologies rely on different paradigms and metaphors—that is, on different theoretical assumptions about the nature of problem contexts—applied systems thinking depends for its justification and rationality on paradigmatic fit between systems methodologies and problem contexts.

In applied OR/MS, as in other forms of applied research, the requirement of paradigmatic fit translates into a need for informing the selection and use of methodologies and methods by previous paradigm analysis as well as, where relevant, metaphor analysis, as a condition for doing justice to the nature of the problem context at issue. TSI/CH consequently puts its critical focus on the theoretical underpinnings of alternative research paradigms rather than on the normative core of professional practice, as does CSH. CST, thus understood, is about methodology choice—a theoretically informed way to support processes of matching methodologies and methods with problem contexts.

Methodological Approach

The basic strategy of TSI/CH can be described as a contingency approach to methodology choice, based on paradigm analysis and, to a lesser degree, also on metaphor analysis of the three major traditions of systems thinking thus far—hard, soft, and critical systems thinking. The idea is that there is no such thing as a best systems methodology and underpinning tradition of systems thinking; rather, situational aspects of the problem context at hand determine what tradition of systems thinking is best suited as a source of methodological guidance and specific methods or tools of intervention. In OR/MS, such an approach promises to resolve the OR in crisis debate of the 1970s and 1980s, for it allows hard and soft OR approaches to be seen as appropriate for dealing with different problem contexts rather than competing for the same ones.

Contingency frameworks are also called contingency theories, as they involve theoretical generalizations about the crucial aspects of the application domain to which the framework is to be applied. This theoretical device is often used in the social sciences (e.g., in management and organization theories) when a variety of approaches are required to handle a given class of problems, as the proper approach is dependent (contingent) on the situation or, more precisely, on a range of changing situations.

Applied to contexts of professional intervention, using a contingency approach implies that some independent (contextual) variables can be identified empirically which regularly, for reasons that can be explained theoretically, may be expected to condition the outcome of interventions. A contingency approach can then (and only then) make sure that the way one deals with a situation matches situational requirements and, on that basis, can also justify the credibility of the results. To the extent this condition is fulfilled, one can properly speak of a contingency theory. It follows that the crucial question for any contingency approach is whether it can identify and theoretically justify a small number of empirical dimensions (ideally only two) in terms of which the range of situations in question can be usefully classified, so that each type of empirical situation can then be identified in a relevant and reliable way.

Methodological Core Principle

TSI/CH's answer to the problem of ensuring paradigmatic fit between intervention approaches and problem contexts is a classification of problem contexts, and of systems methodologies assigned to them, called the system of systems methodologies (SOSM). It says that systems methodologies and conforming methods are well chosen if their underlying systems metaphor (machine, organism, etc.) and/or paradigm (functionalist, interpretive, etc.) match with the nature of the problem context, or more exactly, with assumptions about the kind of complexity that needs to be handled in the problem context. The basic idea and aim of TSI/CH, then, is to support systematic processes of informed methodology choice, as a way to secure paradigmatic fit between intervention methods and intervention contexts. To this end, TSI/CH proposes the SOSM (see Fig. 4).

There was an earlier, four-celled version of the SOSM (Jackson and Keys 1984) that is now often cited as the origin of the TSI strand of CST. However, it only distinguished hard and soft methodologies, and its discussion in that early paper did not yet introduce the notion of critical systems thinking.

CSH became known to Jackson and Keys shortly after publishing their 1984 paper. First hints at a planned extension of their work appeared in a few articles in the late 1980s (Jackson 1987, 1990); the extended SOSM was presented later in Flood and Jackson (1991) and Jackson (1991).

Due to the underlying logic of the SOSM, the extended scheme could not manage to include CSH except by constricting its notion of critical systems thinking considerably. This logic assumes that any

Systems dimension of contexts (increasing complexity)	Participants dimension of contexts (increasing diversity of values)		
	Unitary (paradigm: functional) **HARD SYSTEMS THINKING**	*Pluralist (paradigm: interpretive)* **SOFT SYSTEMS THINKING**	*Coercive (paradigm: emancipatory)* **EMANCIPATORY SYSTEMS THINKING**
Simple	*Simple-unitary problem contexts (systems metaphor: machine)* • Operations research (OR) • Systems engineering (SE) • Systems analysis (SA)	*Simple-pluralist problem contexts (systems metaphors: culture, coalition)* • Systems approach (Churchman) • Strategic assumption surfacing and testing (SAST)	*Simple-coercive problem contexts (systems metaphor: prison)* • Critical systems heuristics (CSH)
Complex	*Complex-unitary problem contexts (systems metaphors: organism, brain)* • Organizational cybernetics/ viable systems diagnosis (VSD) • Socio-technical systems thinking	*Complex-pluralist problem contexts (systems metaphors: culture, coalition)* • Interactive planning (Ackoff) • Soft systems methodology (SSM)	*Complex-coercive problem contexts (systems metaphor: prison)* • ?

Critical Systems Thinking, Fig. 4 The extended system of systems methodologies (SOSM) (Source: Adapted from Flood and Jackson 1991, 42; Jackson 1991, 29 and 31; 2000, 359)

methodology can be meaningfully assigned to a single type of problem context and to a conforming (dominant) theoretical paradigm. There is no room in such a scheme for an approach that focuses on the genuinely normative core of practice as such, whatever the theoretical paradigm adopted may be and the choice of methodology and conforming methods it may inspire. This makes it understandable why the extended SOSM rather arbitrarily assigned CSH a merely emancipatory purpose, as opposed to the critical purpose of the SOSM. To render this choice more plausible, CSH was associated with a prison metaphor, which then seemed to render CSH adequate for coercive problem contexts only and thus provided a rationale for assigning it to a specific emancipatory paradigm (for critical discussion and alternatives, see Ulrich 2003). In this way, CSH became in the SOSM scheme an apparently self-contained methodology that, quite against its original intentions, was to be chosen (or not) as an alternative to soft and hard systems methodologies. Its concern for the practical-normative side of all practice thus moved out of sight.

In British OR/MS, CSH was henceforth understood mainly through the lens of the SOSM, and critical systems thinking became widely identified with TSI. Consequently, CST was now almost the same as the SOSM—an updated contingency framework for methodology choice, as well as for continuing discussions about the evolution of OR/MS (e.g., Jackson 2006a). Both uses attracted much interest and the mentioned difficulties of the extended SOSM did not hamper its success in helping to raise awareness in the profession that there are options for conceiving of good professional practice. The discussion that the SOSM was able to generate in turn has helped to make CSH more known, so that its core principle of boundary critique is increasingly being recognized as an important, independent source of critical thought on practice. These diverse successes of the SOSM certainly have contributed to the comparatively high level of methodological awareness and discussion by which the OR/MS profession distinguishes itself from other fields, which has allowed it to pioneer soft and critical systems ideas that are now radiating into many other fields.

Practical Implementation (Main Procedure)

To support methodology choice in practice, the SOSM still needed to be embedded in a methodology, properly speaking, that is, a framework that would guide practitioners in asking relevant questions and proceeding systematically. This is what total systems intervention (TSI), a name adopted in 1991, is all about. It stands for the practical procedure of methodology choice and implementation that Flood and Jackson (1991) proposed on the basis of the SOSM. The aim is to provide a meta-methodology for methodology choice and implementation. The procedure may be employed in a linear or iterative way. Originally it consisted of three phases labeled creativity, choice, and implementation, to which Jackson (2003, 2006b) later, in the extended

Critical Systems Thinking, Table 1 The meta-methodology of TSI/CH: standard phases of methodology choice and use

Phase	Activity/Aim
(1) CREATIVITY	
Task	To identify major aims and issues of the problem context
Tools	Use of different metaphors and paradigms to gain different perspectives
Outcome	Appreciation of dominant and dependent metaphors/paradigms and related issues
(2) CHOICE	
Task	To choose appropriate systems methodologies and methods
Tools	Use of SOSM to reveal strengths and weaknesses of methodologies and methods
Outcome	Choice of dominant and dependent systems methodologies and methods
(3) IMPLEMENTATION	
Task	To arrive at and implement specific positive change proposals
Tools	Systems methodologies and methods used properly according to the logic of TSI/CH
Outcome	Relevant change according to the concerns of the different paradigms
(4) REFLECTION	
Task	To evaluate the intervention and ensure methodological learning
Tools	Understanding of the concerns of different paradigms regarding good practice
Outcome	Methodological progress

TSI total systems intervention = phases 1–3, *CH* creative holism = phases 1–4, *SOSM* system of systems methodologies
(Source: Adapted from Flood and Jackson 1991, 54; Jackson 1991, 276; 2000, 372; and 2006b, 654)

version he now prefers to call creative holism, added a fourth phase, Reflection (Table 1).

The creativity phase is intended to encourage consideration of what alternative systems paradigms and root metaphors might mean for thinking about a problem context at hand, so that a dominant metaphor could be identified as most adequate, that is, in effect, preference for a hard (functionalist), soft (interpretive), or critical (emancipatory) orientation. In the choice and implementation phases, a conforming particular systems methodology should then be chosen based on the SOSM and used to implement specific change proposals.

A new element in CH as compared to its predecessor TSI is the reflection phase, which brings in an element of reflective practice as CSH understands it, by looking at the outcomes of methodology choice and implementation rather than at its theoretical justification only. Although the underlying notion of evaluation is still not genuinely practical in the sense of CSH and practical philosophy, this development does promise to open up new chances for reflective practice.

Another new element, following a considerable amount of discussion in the literature about methodological complementarism or pluralism (Jackson 1997, 1999), mixing methods (Midgley 1997), and multi-methodology (Mingers and Gill 1997), is that creative holism, unlike TSI, no longer insists on choosing a single dominant paradigm. Instead, a combination of methodologies, or parts of methodologies and conforming methods, is now encouraged, which makes the framework more flexible and brings it closer to actual practice. As Jackson describes it, CH now is a "meta-methodological" framework that aims to help practitioners to "harness the various systems methodologies, methods and models" by being "multi-paradigm, multi-methodology and multi-method in orientation" (Jackson 2006b, 248 and 253).

A Summary Comparison of CSH and TSI

To provide an overview of the discussed aspects of critical systems thinking, Table 2 summarizes the accounts of CSH and TSI in a way that should facilitate comparison.

Concluding Remarks

The claim of professional practice to relevance, rigor, and rationality depends on many requirements. Among these, two crucial ones are putting the problem well

Critical Systems Thinking, Table 2 CSH and TSI compared

Aspect	CSH	TSI/CH
Core idea	Professional practice involves *validity claims* that cannot be justified theoretically but at least can be handled openly and critically in the process of intervention itself	Professional practice involves *methodological choices* that can be justified theoretically by analyzing underpinning research paradigms and systems metaphors
Critical focus	*Reflective practice:* surfacing the reference systems underpinning all judgments of fact and value and analyzing how they condition practical claims (e.g., problem definitions, relevant contexts, standards of improvement, and proposals for action)	*Paradigm analysis:* surfacing the theoretical underpinnings of alternative research paradigms (e.g., functionalist, interpretive, emancipatory, or post-modern) and analyzing how they condition different perceptions of problem contexts and suitable methodological choices
Approach	*Critical systems discourse:* a discursive framework for value clarification and critique	*Contingency theory:* a contingency framework for methodology choice and use
Methodological core principle	*Boundary critique:* unfolding the selectivity of reference systems	*Informed methodology choice:* matching systems methodologies with problem contexts
Main critical device	*Checklist of boundary questions:* a definition of boundary categories for "is" and "ought" mapping (i.e., descriptive and normative analysis) of reference systems	*System of systems methodologies* (SOSM): a classification of problem contexts and conforming systems methodologies
Implementation	A discursive *process of unfolding selectivity:* a standard sequence of boundary critique	A holistic *meta-methodology of paradigm analysis:* standard phases of methodology choice and reflection

CSH critical systems heuristics, *TSI/CH* total systems intervention/creative holism

and tackling it by means of adequate methods. In different ways, both of these embody crucial requirements of professional competence. Both of them stand for efforts to make sure that relevant issues are properly identified and the implications of related assumptions are made transparent and evaluated.

- *Putting problems well* is an issue that involves empirical (observational) as well as normative (ethical) problem structuring and reflection. The selection of relevant facts and values depends on a proper understanding of the problem, which is hardly achievable without questioning the scope and diversity of the social context that matters. It also depends on the extent to which justice is done in practice to the diversity of views and concerns of the different parties concerned. A problem may be ill defined so long as this normative core of any quest for rational practice is not well understood.
- *Choosing and employing methods properly* involves analysis and reflection about the demands of problem situations on the one hand and about the availability of methods that respond to these demands on the other. The selection of adequate methodologies and methods depends on a proper understanding of the theoretical and paradigmatic assumptions involved, which is hardly achievable without questioning the nature of the complexity that matters. It also depends on the extent to which the matching of such assumptions with specific situations is successful in practice. A methodology and conforming methods may be ill chosen so long as this theoretical core of the quest for rational practice is not well understood.

Neither effort replaces or precludes the other. Critical systems thinking, properly understood, aims to promote good practice with regard to both. To this end, the two strands of CST bring to bear within the field of OR/MS, and in the applied sciences in general, new philosophical and theoretical foundations, along with new practical tools for analyzing contextual complexity and diversity. CSH draws on practical philosophy and consequently conceives of rational practice in terms of discursive tools of value clarification and critique, in particular boundary critique and discourse. TSI/CH draws on organizational sociology and conceives of rational practice in terms of theoretically informed tools of methodology choice, in particular paradigm analysis and metaphor analysis.

Different as the resulting frameworks of CSH and TSI are, their shared concern remains the idea that

good professional practice depends crucially on making sure that problems are well put and methods of intervention are well chosen; and that to meet both requirements, it is essential to properly situate problems in their contexts and make sure one understands those contexts well. Formulated in everyday terms, the essential message of CST to professionals might thus be summarized as follows:

Critical Systems Thinking: Its Operational Imperative

As a professional intervening in a specific context, pay attention to your contextual assumptions and try to identify and examine them systematically, so as to understand them well. Then make sure everyone concerned understands them well too. Work toward mutual understanding about how problem definitions and solutions depend on and change with the facts and values considered relevant. Make sure divergent views and values are properly addressed. Adapt your choice of methodologies and methods to the amount of diversity that you find in the problem context, and to the resulting nature of the complexity that matters. Finally, whatever problem definitions and methods your professional practice ultimately relies on, reflect on the validity claims your professional findings and conclusions imply and how, if taken as a basis for action, they may affect the different parties concerned. Make boundary critique a standard practice to this end, and always remember that no professional intervention can do justice to all views and values, that is, can justify all its implications. But at least it can deal with this inevitable lack of complete justification in a transparent and self-reflecting way. This is what critical professional practice is all about.

See

- ▶ Community OR
- ▶ Cybernetics and Complex Adaptive Systems
- ▶ Practice of Operations Research and Management Science
- ▶ Problem Structuring Methods
- ▶ Soft Systems Methodology
- ▶ System Dynamics
- ▶ Systems Analysis

References

Beer, S. (1972). *Brain of the firm* (2nd ed. Chichester: Wiley, 1981). Harmondsworth: Penguin Press.

Beer, S. (1985). *Diagnosing the system for organizations*. Chichester: Wiley.

Burrell, G., & Morgan, G. (1979). *Sociological paradigms and organizational analysis: Elements of the sociology of corporate life*. London: Heinemann.

Checkland, P. (1981). *Systems thinking, systems practice*. Chichester: Wiley.

Checkland, P. (1985). From optimizing to learning: A development of systems thinking for the 1990s. *Journal of the Operational Research Society, 36*, 757–767.

Checkland, P., & Scholes, J. (1990). *Soft systems methodology in action*. Chichester: Wiley.

Churchman, C. W. (1968). *The systems approach*. New York: Dell Publishing.

Churchman, C. W. (1971). *The design of inquiring systems: Basic concepts of systems and organization*. New York: Basic Books.

Churchman, C. W. (1979). *The systems approach and its enemies*. New York: Basic Books.

Churchman, C. W., Ackoff, R. L., & Arnoff, E. L. (1957). *Introduction to operations research*. New York/London: Wiley/Chapman & Hall.

Flood, R. L., & Jackson, M. C. (1991). *Creative problem solving: Total systems intervention*. Chichester: Wiley.

Habermas, J. (1973). Wahrheitstheorien. In H. Fahrenbach (Ed.), *Wirklichkeit und Reflexion: Walter Schulz zum. 60 Geburtstag* (pp. 211–265). Neske: Pfullingen.

Habermas, J. (1979). What is universal pragmatics? In J. Habermas (Ed.), *Communication and the evolution of society* (pp. 1–68). Boston: Beacon Press.

Jackson, M. C. (1987). New directions in management science. In M. C. Jackson & P. Keys (Eds.), *New directions in management science* (pp. 133–164). Aldershot: Gower.

Jackson, M. C. (1990). Beyond a system of systems methodologies. *Journal of the Operational Research Society, 41*, 657–668.

Jackson, M. C. (1991). *Systems methodology for the management sciences*. New York: Plenum.

Jackson, M. C. (1997). Pluralism in systems thinking and practice. In J. Mingers & A. Gill (Eds.), *Multimethodology: The theory and practice of integrating management science methodologies* (pp. 347–378). Chichester: Wiley.

Jackson, M. C. (1999). Towards coherent pluralism in management science. *Journal of the Operational Research Society, 50*, 12–22.

Jackson, M. C. (2000). *Systems approaches to management*. New York: Kluwer/Plenum.

Jackson, M. C. (2003). *Systems thinking: Creative holism for managers*. Chichester: Wiley.

Jackson, M. C. (2006a). Beyond problem structuring methods: Reinventing the future of OR/MS. *Journal of the Operational Research Society, 57*, 868–878.
Jackson, M. C. (2006b). Creative holism: A critical systems approach to complex problem situations. *Systems Research and Behavioral Science, 23*, 647–657.
Jackson, M. C., & Keys, P. (1984). Towards a system of system methodologies. *Journal of the Operational Research Society, 35*, 473–486.
Midgley, G. (1997). Mixing methods: Developing systemic intervention. In J. Mingers & A. Gill (Eds.), *Multimethodology: The theory and practice of integrating management science methodologies* (pp. 249–290). Chichester: Wiley.
Mingers, J., & Gill, A. (Eds.). (1997). *Multimethodology: The theory and practice of integrating management science methodologies*. Chichester: Wiley.
Morgan, G. (1986). *Images of organization* (3rd ed. 2006). Beverly Hills, CA: Sage.
Reynolds, M. (2007). Evaluation based on critical systems heuristics. In B. Wiliams & I. Imam (Eds.), *Systems concepts in evaluation: An expert anthology* (pp. 101–122). Point Reyes, CA: Edge Press.
Rosenhead, J. (Ed.). (1989). *Rational analysis for a problematic world: problem structuring methods for complexity, uncertainty and conflict*. Chichester: Wiley (Revised edition: Rosenhead, J., & Mingers, J. (Eds.). (2001). *Rational analysis for a problematic world revisited: Problem structuring methods for complexity, uncertainty and conflict*. Chichester: Wiley).
Ulrich, W. (1983). *Critical heuristics of social planning: A new approach to practical philosophy*. Bern: Paul Haupt. Reprinted Chichester: Wiley (1994).
Ulrich, W. (1987). Critical heuristics of social systems design. *European Journal of Operational Research, 31*, 276–283.
Ulrich, W. (1993). Some difficulties of ecological thinking, considered from a critical systems perspective: A plea for critical holism. *Systems Practice, 6*, 583–611.
Ulrich, W. (1996). *A primer to critical systems heuristics for action researchers*. Hull: Centre for Systems Studies, University of Hull.
Ulrich, W. (2000). Reflective practice in the civil society: The contribution of critically systemic thinking. *Reflective Practice, 1*, 247–268.
Ulrich, W. (2001). The quest for competence in systemic research and practice. *Systems Research and Behavioral Science, 18*, 3–28.
Ulrich, W. (2003). Beyond methodology choice: Critical systems thinking as critically systemic discourse. *Journal of the Operational Research Society, 54*, 325–342.
Ulrich, W. (2006). Critical pragmatism: A new approach to professional and business ethics. In L. Zsolnai (Ed.), *Interdisciplinary yearbook of business ethics* (Vol. 1, pp. 53–85). Oxford: Peter Lang.
Ulrich, W. (2007). Philosophy for professionals: Towards critical pragmatism. *Journal of the Operational Research Society, 58*, 1109–1113.
Ulrich, W., & Reynolds, M. (2010). Critical systems heuristics. In M. Reynolds & S. Holwell (Eds.), *Systems approaches to managing change: A practical guide* (pp. 243–292). London: Springer (in association with The Open University, Milton Keynes, UK).

Cross-Entropy Method

Dirk P. Kroese[1], Reuven Y. Rubinstein[2], Izack Cohen[2], Sergey Porotsky[3] and Thomas Taimre[1]
[1]The University of Queensland, Brisbane, Australia
[2]Technion – Israel Institute of Technology, Haifa, Israel
[3]A.L.D. Ltd., Tel-Aviv, Israel

Introduction

The cross-entropy (CE) method is a versatile Monte Carlo technique introduced by Rubinstein (1999, 2001), extending earlier work on variance minimization (Rubinstein 1997). A tutorial on the CE method is given in de Boer et al. (2005). A comprehensive treatment can be found in Rubinstein and Kroese (2004); see also Rubinstein and Kroese (2007, Chap. 8).

The CE method can be applied to two types of problems:

1. *Estimation*: Estimate $\ell = \mathbb{E}[H(\mathbf{X})]$, where $\mathbf{X}$ is a random object taking values in some set $\mathcal{X}$ and H is a function on $\mathcal{X}$. An important special case is the estimation of a probability $\ell = \mathbb{P}(S(X) \geq \gamma)$, where S is another function on $\mathcal{X}$.
2. *Optimization*: Optimize (i.e., maximize or minimize) $S(\mathbf{x})$ over all $\mathbf{x} \in \mathcal{X}$, where S is some objective function on $\mathcal{X}$.

In the estimation setting, the CE method can be viewed as an adaptive importance sampling procedure that uses the cross-entropy or Kullback–Leibler divergence as a measure of closeness between two sampling distributions. In the optimization setting, the optimization problem is first translated into a rare-event estimation problem, and then the CE method for estimation is used as an adaptive algorithm to locate the optimum.

Estimation

Consider the estimation of

$$\ell = \mathbb{E}_f[H(\mathrm{X})] = \int H(\mathrm{x}) f(\mathrm{x})\, \mathrm{dx}, \qquad (1)$$

where H is a real-valued function and f is the probability density function (pdf) of the random vector $\mathbf{X}$. It is assumed, for simplicity, that $\mathbf{X}$ is a continuous random variable. For the discrete case, replace the integral in (1) by a sum. Let g be another pdf—which must be nonzero for every $\mathbf{x}$ for which $H(\mathbf{x})\, f(\mathbf{x}) \neq 0$. Using the pdf g, ℓ can be represented as

$$\ell = \int H(\mathbf{x}) \frac{f(\mathbf{x})}{g(\mathbf{x})} g(\mathbf{x}) \mathrm{d}\mathbf{x} = \mathbb{E}_g \left[H(\mathbf{X}) \frac{f(\mathbf{X})}{g(\mathbf{X})} \right], \quad (2)$$

where the subscript g indicates that the expectation is taken with respect to g rather than f. Consequently, if $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are independent random vectors with pdf g, written as $\mathbf{X}_1, \ldots, \mathbf{X}_N \sim_{\text{iid}} g$, then

$$\hat{\ell} = \frac{1}{N} \sum_{k=1}^{N} H(\mathbf{X}_k) \frac{f(\mathbf{X}_k)}{g(\mathbf{X}_k)} \quad (3)$$

is an unbiased estimator of ℓ—a so-called importance sampling estimator. The optimal importance sampling pdf, that is, the pdf g^* that minimizes the variance of $\hat{\ell}$, is proportional to $|H|\, f$ (see, e.g., Rubinstein and Kroese (2007, 132)), but is in general difficult to evaluate. The idea of the CE method is to choose the importance sampling pdf g in a specified class of pdfs such that the Kullback–Leibler divergence between the optimal importance sampling pdf g^* and g is minimal. The Kullback–Leibler divergence between two pdfs g and h is given by

$$\begin{aligned} \mathcal{D}(g,h) &= \mathbb{E}_g \left[\ln \frac{g(\mathbf{X})}{h(\mathbf{X})} \right] = \int g(\mathbf{x}) \ln \frac{g(\mathbf{x})}{h(\mathbf{x})} \, \mathrm{d}\mathbf{x} \\ &= \int g(\mathbf{x}) \ln g(\mathbf{x}) \, \mathrm{d}\mathbf{x} - \int g(\mathbf{x}) \ln h(\mathbf{x}) \, \mathrm{d}\mathbf{x}. \quad (4) \end{aligned}$$

In most cases of interest the function H is nonnegative, and the nominal pdf f is parameterized by a finite-dimensional vector $\mathbf{u}$; that is, $f(\mathbf{x}) = f(\mathbf{x}; \mathbf{u})$. It is then customary to choose the importance sampling pdf g in the same family of pdfs; thus, $g(\mathbf{x}) = f(\mathbf{x}; \mathbf{v})$ for some reference parameter $\mathbf{v}$. The CE minimization procedure then reduces to finding an optimal reference parameter vector, say $\mathbf{v}^*$, by cross-entropy minimization:

$$\begin{aligned} \mathbf{v}^* &= \operatorname*{argmin}_{\mathbf{v}} \mathcal{D}(g^*, f(\cdot; \mathbf{v})) \\ &= \operatorname*{argmax}_{\mathbf{v}} \int H(\mathbf{x}) f(\mathbf{x}; \mathbf{u}) \ln f(\mathbf{x}; \mathbf{v}) \, \mathrm{d}\mathbf{x} \\ &= \operatorname*{argmax}_{\mathbf{v}} E_{\mathbf{u}} H(\mathbf{X}) \ln f(\mathbf{X}; \mathbf{v}) \\ &= \operatorname*{argmax}_{\mathbf{v}} E_{\mathbf{w}} H(\mathbf{X}) \ln f(\mathbf{X}; \mathbf{v}) \frac{f(\mathbf{X}; \mathbf{u})}{f(\mathbf{X}; \mathbf{w})}, \quad (5) \end{aligned}$$

where $\mathbf{w}$ is any reference parameter. This $\mathbf{v}^*$ can be estimated via the stochastic counterpart of (5):

$$\hat{\mathrm{v}} = \operatorname*{argmax}_{\mathrm{v}} \frac{1}{N} \sum_{k=1}^{N} H(\mathrm{X}_k) \frac{f(\mathrm{X}_k; \mathrm{u})}{f(\mathrm{X}_k; \mathrm{w})} \ln f(\mathrm{X}_k; \mathrm{v}), \quad (6)$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_N \sim_{\text{iid}} f(\cdot; \mathbf{w})$. The optimal parameter $\hat{\mathrm{v}}$ in (6) can often be obtained in explicit form, in particular when the class of sampling distributions forms an exponential family; see, for example, Rubinstein and Kroese (2007, 319–320). Indeed, analytical updating formulas can be found whenever explicit expressions for the maximal likelihood estimators of the parameters can be found, cf. de Boer et al. (2005, 36).

Example: Exponential Random Variables. Consider the case where $\mathbf{X}_1 = (X_1, \ldots, X_n)$ is a vector of independent exponential random variables with expectations $u_1, \ldots, u_n$. Let $\mathbf{u} = (u_1, \ldots, u_n)$ and let $\mathbf{v} = (v_1, \ldots, v_n)$ be the reference parameter of the importance sampling pdf $f(\mathbf{x}; \mathbf{v})$, given by

$$f(\mathbf{x}, \mathbf{v}) = \prod_{i=1}^{n} \frac{\mathrm{e}^{-x_i / v_i}}{v_i}.$$

Hence, under this importance sampling pdf, $X_1, \ldots, X_n$ are again independent and exponentially distributed, but now with expectations $v_1, \ldots, v_n$. Writing $H_k = H(\mathbf{X}_k)$ and the likelihood ratio $W_k = f(\mathbf{X}_k; \mathbf{u}) / f(\mathbf{X}_k; \mathbf{w})$ in (6), the optimal parameter $\hat{\mathrm{v}}$ is found by maximizing

$$\begin{aligned} &\sum_{i=1}^{n} \sum_{k=1}^{N} H_k W_k \ln f(\mathbf{X}_k; \mathbf{u}) \\ &= \sum_{i=1}^{n} \sum_{k=1}^{N} H_k W_k \left(\frac{-X_{ki}}{v_i} - \ln v_i \right), \quad (7) \end{aligned}$$

where X_{ki} is the i-th component of $\mathbf{X}_k$. This maximum can be found by differentiating and equating to zero the right-hand side of (7) for each v_i, resulting in the equations

$$\sum_{k=1}^{N} H_k W_k \left(\frac{X_{ki}}{v_i^2} - \frac{1}{v_i} \right) = 0, \quad i = 1, \ldots, n,$$

from which it follows that

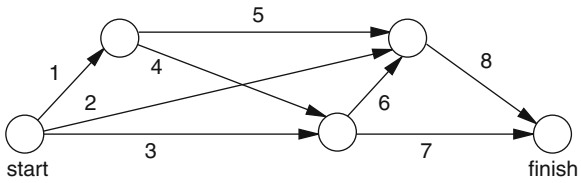
$$\hat{v}_i = \frac{\sum_{k=1}^{N} H_k W_k X_{ki}}{\sum_{k=1}^{N} H_k W_k}, \quad i = 1, \ldots, n. \tag{8}$$

Often $\ell = \mathbb{P}(S(\mathrm{X}) \geqslant \gamma)$ for some function S and level γ, in which case $H(\mathbf{x})$ takes the form of an indicator function: $H(\mathrm{x}) = I_{\{S(\mathrm{x}) \geqslant \gamma\}}$; that is, $H(\mathbf{x}) = 1$ if $S(\mathbf{x}) \geqslant \gamma$, and 0 otherwise. A complication in solving (6) occurs when ℓ is a rare-event probability; that is, a very small probability (say less than 10^{-4}). Then, for moderate sample size N, most or all of the values $H(\mathbf{X}_k)$ in (6) are zero, and the maximization problem becomes useless. In that case a multilevel CE procedure is used, where a sequence of reference parameters and levels is constructed with the goal that the former converges to $\mathbf{v}^*$ and the latter to γ. This leads to the following algorithm; see, for example, Rubinstein and Kroese (2007, 238).

Algorithm 1 (CE Algorithm for Rare-Event Estimation).

1. *Define* $\hat{\mathbf{v}}_0 = \mathbf{u}$. *Let* $N^{\mathrm{e}} = \lceil \varrho N \rceil$. *Set* $t = 1$ (*iteration counter*).
2. *Generate* $\mathbf{X}_1, \ldots, \mathbf{X}_N \sim_{\mathrm{iid}} f(\cdot; \hat{\mathbf{v}}_{t-1})$. *Calculate* $S_i = S(\mathbf{X}_i)$ *for all* i, *and order these from smallest to largest:* $S_{(1)} \leq \ldots \leq S_{(N)}$. *Let* $\hat{\gamma}_t$ *be the sample* $(1 - \varrho)$*-quantile of performances; that is,* $\hat{\gamma}_t = S_{(N - N^{\mathrm{e}} + 1)}$. *If* $\hat{\gamma}_t > \gamma$, *reset* $\hat{\gamma}_t$ *to* γ.
3. *Use the* **same** *sample* $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *to solve the stochastic program* (6), *with* $\mathbf{w} = \hat{\mathbf{v}}_{t-1}$. *Denote the solution by* $\hat{\mathbf{v}}_t$.
4. *If* $\hat{\gamma}_t < \gamma$, *set* $t = t + 1$ *and reiterate from Step 2; otherwise, proceed with Step 5.*
5. *Let* T *be the final iteration counter. Generate* $\mathbf{X}_1, \ldots, \mathbf{X}_{N_1} \sim_{iid} f(\cdot; \hat{\mathbf{v}}_T)$ *and estimate* ℓ *via importance sampling, as in* (3).

Apart from specifying the family of sampling pdfs, the sample sizes N and N_1, and the rarity parameter ϱ (typically between 0.01 and 0.1), the algorithm is completely self-tuning. The sample size N for determining a good reference parameter can usually be chosen much smaller than the sample size N_1 for the final importance sampling estimation, say $N = 1000$ versus $N_1 = 100{,}000$. Under certain technical conditions the deterministic version of Algorithm 1 is guaranteed to terminate (reach level γ) provided that ϱ is chosen small enough; see Sect. 3.5 of Rubinstein and Kroese (2004).



Cross-Entropy Method, Fig. 1 A stochastic activity network

Example: Rare-Event Probability Estimation. A stochastic activity network is a frequently used tool in project management to schedule concurrent activities. Each arc corresponds to an activity and is weighted by the duration of that activity. The maximal project duration corresponds to the length of the longest path in the graph. Figure 1 shows a stochastic activity network with eight activities. Suppose the durations of the activities are independent exponential random variables $X_1, \ldots, X_8$, each with mean 1.

Let $S(\mathbf{X})$ denote length of the longest path in the graph; that is,

$$\begin{aligned} S(\mathbf{X}) = \max\{&X_1 + X_4 + X_6 + X_8,\ X_1 + X_4 + X_7,\ X_1 \\ &+ X_5 + X_8,\ X_2 + X_8,\ X_3 + X_6 + X_8,\ X_3 + X_7\} \end{aligned}$$

Suppose the objective is to estimate the rare-event probability $\mathbb{P}(S(\mathbf{X}) \geq 20)$ using importance sampling where the random vector $\mathbf{X} = (X_1, \ldots, X_8)$ has independent exponentially distributed components with mean vector $\mathbf{v} = (v_1, \ldots, v_8)$. Note that the nominal pdf is obtained by setting $v_i = 1$ for all i. At the t-th iteration of the multilevel CE Algorithm 1, the solution to (6) with $H(\mathbf{X}) = \mathrm{I}_{\{S(\mathbf{X}) \geq \hat{\gamma}_t\}}$, using (8), is given by

$$\hat{v}_{t,i} = \frac{\sum_{k=1}^{N} \mathrm{I}_{\{S(\mathrm{X}_k) \geq \hat{\gamma}_t\}} W_k X_{ki}}{\sum_{k=1}^{N} \mathrm{I}_{\{S(\mathrm{X}_k) \geq \hat{\gamma}_t\}} W_k}, \tag{9}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_N \sim_{\mathrm{iid}} f(\cdot; \hat{\mathbf{v}}_{t-1})$, $W_k = f(\mathbf{X}_k; \mathbf{u}) / f(\mathbf{X}_k; \hat{\mathbf{v}}_{t-1})$, and X_{ki} is the i-th element of $\mathbf{X}_k$.

Cross-Entropy Method, Table 1 Convergence of the sequence $\{(\hat{\gamma}_t, \hat{\mathbf{v}}_t)\}$

t	$\hat{\gamma}_t$	$\hat{v}_t$							
0	–	1	1	1	1	1	1	1	1
1	7.32	1.93	1.12	1.39	1.83	1.32	1.81	1.37	1.96
2	12.01	3.33	1.09	1.58	2.98	1.50	2.95	1.58	3.32
3	20	5.03	1.00	1.88	4.63	1.51	4.73	1.47	5.14

Table 1 lists the successive estimates for the optimal importance sampling parameters obtained from the multilevel CE algorithm, using $N = 10^5$ and $\varrho = 0.1$.

The last step in Algorithm 1 gives an estimate of $4.15 \cdot 10^{-6}$ with an estimated relative error of 1%, using a sample size of $N_1 = 10^6$. A typical crude Monte Carlo estimate (i.e., taking $\mathbf{v} = \mathbf{u} = (1,1,\ldots,1)$) using the same sample size is $3 \cdot 10^{-6}$, with an estimated relative error of 60%, and is therefore of little use.

For large-size activity networks the accurate estimation of the optimal parameters via (9) runs into problems due to the degeneracy behavior of the likelihood ratio; cf. Rubinstein and Kroese (2007, 133). For such systems it is recommended to estimate the optimal CE parameters by drawing samples directly from g^*, for example, via Markov chain Monte Carlo; see Chan (2010).

Optimization

Let $\mathcal{X}$ be an arbitrary set of states and let S be a real-valued performance function on $\mathcal{X}$. Suppose the goal is to find the maximum of S over $\mathcal{X}$, and the corresponding maximizer $\mathbf{x}^*$ (assuming, for simplicity, that there is only one). Denote the maximum by γ^*, so that

$$S(\mathbf{x}^*) = \gamma^* = \max_{\mathbf{x} \in \mathcal{X}} S(\mathrm{X}). \tag{10}$$

Associate with the above problem the estimation of the probability $\ell = \mathbb{P}(S(\mathbf{X}) \geq \gamma)$, where $\mathbf{X}$ has some probability density $f(\mathbf{x}; \mathbf{u})$ on $\mathcal{X}$ (e.g., corresponding to the uniform distribution on $\mathcal{X}$) and γ is some level. Thus, for optimization problems, randomness is purposely introduced in order to make the model stochastic, as in the estimation setting. If γ is chosen close to the unknown γ^*, then ℓ is typically a rare-event probability, and the CE approach of section "Estimation" can be used to find an importance sampling distribution close to the theoretically optimal importance sampling density, which concentrates all its mass on point $\mathbf{x}^*$. Sampling from such a distribution thus produces optimal or near-optimal states. Note that the final level $\gamma = \gamma^*$ is generally not known in advance, in contrast to the rare-event simulation setting. The CE method for optimization produces a sequence of levels $\{\hat{\gamma}_t\}$ and reference parameters $\{\hat{\mathbf{v}}_t\}$ such that the former tends to the optimal γ^* and the latter to the optimal reference vector $\mathbf{v}^*$ corresponding to the point mass at $\mathbf{x}^*$; see, for example, Rubinstein and Kroese (2007) p. 251.

Algorithm 2 (CE Algorithm for Optimization).

1. *Choose an initial parameter vector* $\hat{\mathbf{v}}_0$. *Let* $N^{\mathrm{e}} = \lceil \varrho N \rceil$. *Set* $t = 1$ (*level counter*).
2. *Generate* $\mathbf{X}_1, \ldots, \mathbf{X}_N \sim_{\mathrm{iid}} f(\cdot; \hat{\mathbf{v}}_{t-1})$. *Calculate the performances* $S(\mathbf{X}_i)$ *for all* i, *and order them from smallest to largest:* $S_{(1)} \leq \ldots \leq S_{(N)}$. *Let* $\hat{\gamma}_t$ *be the sample* $(1 - \varrho)$*-quantile of performances; that is,* $\hat{\gamma}_t = S_{(N - N^{\mathrm{e}} + 1)}$.
3. *Use the* **same** *sample* $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *and solve the stochastic program*

$$\max_{\mathbf{v}} \frac{1}{N} \sum_{k=1}^{N} I_{\{S(X_k) \geq \hat{\gamma}_t\}} \ln f(\mathbf{X}_k; \mathbf{v}). \tag{11}$$

Denote the solution by $\hat{\mathbf{v}}_t$.

4. *If some stopping criterion is met, stop; otherwise, set* $t = t + 1$, *and return to Step 2.*

To run the algorithm, one needs to provide the class of sampling pdfs, the initial vector $\hat{\mathbf{v}}_0$, the sample size N, the rarity parameter ϱ, and the stopping criterion. Any CE algorithm for optimization involves thus the following two main iterative phases:

1. *Generate* a random sample of objects in the search space $\mathcal{X}$ (trajectories, vectors, etc.) according to a specified probability distribution.
2. *Update* the parameters of that distribution, based on the N^{e} best performing samples (the so-called elite samples), using CE minimization.

Note that Step 5 of Algorithm 1 is missing in Algorithm 2. Another main difference between the two algorithms is that the likelihood ratio term $f(\mathbf{X}_k; \mathbf{u})/f(\mathbf{X}_k; \hat{\mathbf{v}}_{t-1})$ in (6) is missing in (11).

Often a smoothed updating rule is used, in which the parameter vector $\hat{\mathbf{v}}_t$ is taken as

$$\hat{\mathbf{v}}_t = \alpha\,\tilde{\mathbf{v}}_t + (1-\alpha)\hat{\mathbf{v}}_{t-1}, \tag{12}$$

where $\tilde{\mathbf{v}}_t$ is the solution to (11) and $0 \le \alpha \le 1$ is a smoothing parameter. Many other modifications can be found in Kroese et al. (2006), Rubinstein and Kroese (2004, 2007). When there are two or more optimal solutions, the CE algorithm typically "fluctuates" between the solutions before focusing on one of the solutions. The effect that smoothing has on convergence is discussed in detail in Costa et al. (2007). In particular, it is shown that with appropriate smoothing the CE method converges and finds the optimal solution with probability arbitrarily close to 1. Necessary conditions and sufficient conditions under which the optimal solution is generated eventually with probability 1 are also given. Other convergence results, including a proof of convergence along the lines of the convergence proof for simulated annealing can be found in Margolin (2005). The CE method is also effective for solving noisy optimization problems, for example, when the objective function value is obtained via simulation. Typical examples may be found in Alon et al. (2005) and Cohen et al. (2007).

Combinatorial Optimization

When the state space $\mathcal{X}$ is finite, the optimization problem (10) is often referred to as a discrete or combinatorial optimization problem. For example, $\mathcal{X}$ could be the space of combinatorial objects such as binary vectors, trees, paths through graphs, permutations, etc. To apply the CE method, one needs to first specify a convenient parameterized random mechanism to generate objects $\mathbf{X}$ in $\mathcal{X}$. An important example is where $\mathbf{X} = (X_1, \ldots, X_n)$ has independent components such that $X_i = j$ with probability p_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$. In that case, the CE updating rule (see de Boer et al. 2005, 56) at the t-th iteration is

$$\hat{p}_{t,ij} = \frac{\sum_{k=1}^{N} \mathbf{I}_{\{S(\mathbf{X}_k) \ge \hat{\gamma}t\}}\, \mathbf{I}_{\{X_{ki}=j\}}}{\sum_{k=1}^{N} \mathbf{I}_{\{S(\mathbf{X}_k) \ge \hat{\gamma}t\}}}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, m, \tag{13}$$

where $\mathbf{X}_1, \ldots, \mathbf{X}_N$ are independent copies of $\mathbf{X} \sim \{\hat{p}_{t-1,ij}\}$ and X_{ki} is the *i-th* element of $\mathbf{X}_k$. Thus, the updated probability $\hat{p}_{t,ij}$ is simply the number of elite samples for which the i-th component is equal to j, divided by the total number of elite samples.

A possible stopping rule for combinatorial optimization problems is to stop when the overall best objective value does not change over a number of iterations. Alternatively, one could stop when the sampling distribution has "degenerated" enough. For example, when in (13) the $\{\hat{p}_{t,ij}\}$ differ less than some small $\varepsilon > 0$ from the $\{\hat{p}_{t-1,ij}\}$.

Example: Max-Cut Problem. The max-cut problem in a graph can be formulated as follows. Given a weighted graph $G(V, E)$ with node set $V = \{1, \ldots, n\}$ and edge set E, partition the nodes of the graph into two subsets V_1 and V_2 such that the sum of the (nonnegative) weights of the edges going from one subset to the other is maximized. Let $C = (C(i, j))$ be the matrix of weights. The objective is to maximize

$$\sum_{(i,j)\in V_1 \times V_2} (C(i,j) + C(j,i)) \tag{14}$$

over all cuts $\{V_1, V_2\}$. Such a cut can be conveniently represented by a binary cut vector $\mathbf{x} = (1, x_2, \ldots, x_n)$, where $x_i = 1$ indicates that $i \in V_1$. Let $\mathcal{X}$ be the set of cut vectors and let $S(\mathbf{x})$ be the value of the cut represented by $\mathbf{x}$, as given in (14).

To maximize S via the CE method one can generate the random cut vectors by drawing each component (except the first one, which is set to 1) independently from a Bernoulli distribution, that is, $\mathbf{X} = (1, X_2, \ldots, X_n) \sim \text{Ber}(\mathbf{p})$, where $\mathbf{p} = (1, p_2, \ldots, p_n)$. Given an elite sample set $\mathcal{E}$, with size N^{e}, the updating formula (13) is then:

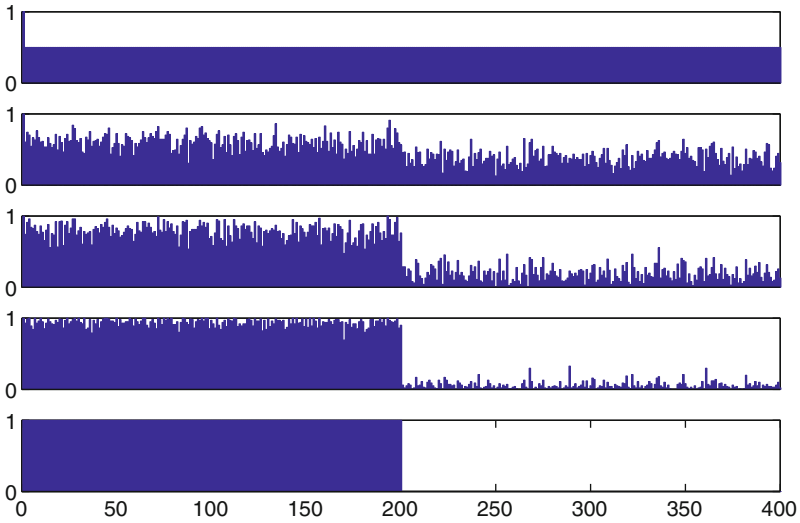
$$\hat{p}_{t,i} = \frac{\sum_{\mathbf{X}\in\mathcal{E}} X_i}{N^{\text{e}}}, \qquad i = 2, \ldots, n. \tag{15}$$

That is, the updated success probability for the *i-th* component is the mean of the i-th components of the vectors in the elite set.

Figure 2 illustrates the evolution of the Bernoulli parameters for a max-cut problem from de Boer et al. (2005) of dimension $n = 400$, for which the optimal solution is given by $\mathbf{x}^* = (1, \ldots, 1, 0, \ldots, 0)$.

Continuous Optimization

When the state space is continuous, in particular when $\mathcal{X} = \mathbb{R}^n$, the optimization problem is often referred to



Cross-Entropy Method, Fig. 2 Sequence of reference vectors for a synthetic max-cut problem with 400 nodes. Iterations 0, 5, 10, 15, and 20 are displayed

as a continuous optimization problem. The sampling distribution on $\mathbb{R}^n$ can be quite arbitrary and does not need to be related to the function that is being optimized. The generation of a random vector $\mathbf{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is most easily performed by drawing the coordinates independently from some 2-parameter distribution. In most applications, a normal (Gaussian) distribution is employed for each component. Thus, the sampling distribution for $\mathbf{X}$ is characterized by a vector of means $\boldsymbol{\mu}$ and a vector of standard deviations $\boldsymbol{\sigma}$. At each iteration of the CE algorithm, these parameter vectors are updated simply as the vectors of sample means and sample standard deviations of the elements in the elite set; see, for example, Kroese et al. (2006).

Algorithm 3 (CE for Continuous Optimization: Normal Updating).

1. ***Initialize***: *Choose* $\hat{\boldsymbol{\mu}}_0$ *and* $\hat{\boldsymbol{\sigma}}_0^2$. *Set* $t = 1$.
2. ***Draw***: *Generate a random sample* $\mathbf{X}_1, \ldots, \mathbf{X}_N$ *from the* $\mathrm{N}(\hat{\boldsymbol{\mu}}_{t-1}, \hat{\boldsymbol{\sigma}}_{t-1}^2)$ *distribution.*
3. ***Select***: *Let* $\mathcal{I}$ *be the indices of the* N^{e} *best performing (= elite) samples.*
   Update: *For all* $j = 1, \ldots, n$ *let*

$$\tilde{\mu}_{t,j} = \sum_{i \in \mathcal{I}} X_{ij} / N^{\mathrm{e}} \tag{16}$$

and

$$\tilde{\sigma}_{t,j}^2 = \sum_{i \in \mathcal{I}} (X_{ij} - \tilde{\mu}_{t,j})^2 / N^{\mathrm{e}}. \tag{17}$$

4. ***Smooth***:

$$\hat{\boldsymbol{\mu}}_t = \alpha \tilde{\boldsymbol{\mu}}_t + (1-\alpha)\hat{\boldsymbol{\mu}}_{t-1}, \quad \hat{\boldsymbol{\sigma}}_t = \alpha \tilde{\boldsymbol{\sigma}}_t + (1-\alpha)\hat{\boldsymbol{\sigma}}_{t-1} \tag{18}$$

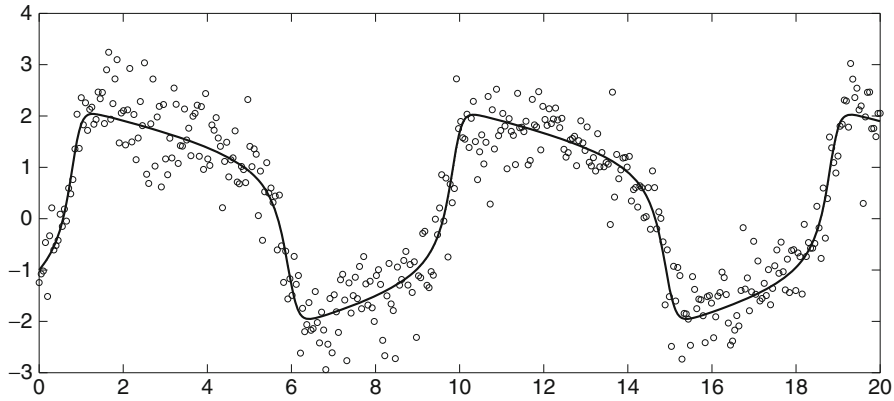
5. *If* $\max_j \{\hat{\sigma}_{t,j}\} < \varepsilon$ stop *and return* $\boldsymbol{\mu}_t$ *as an approximate solution. Otherwise, increase* t *by* 1 *and return to Step 2.*

For constrained continuous optimization problems, where the samples are restricted to a subset $\mathcal{X} \subset \mathbb{R}^n$, it is often possible to replace the normal sampling with sampling from a truncated normal distribution while retaining the updating formulas (16–17). An alternative is to use a beta distribution. Instead of returning $\hat{\boldsymbol{\mu}}_t$ as the final solution, one often returns the overall best solution generated by the algorithm.

Smoothing, as in Step 4, is often crucial to prevent premature shrinking of the sampling distribution. Instead of using a single smoothing factor, it is often useful to use separate smoothing factors for $\hat{\boldsymbol{\mu}}_t$ and $\hat{\boldsymbol{\sigma}}_t$. An alternative is to use dynamic smoothing for $\hat{\boldsymbol{\sigma}}_t$:

$$\alpha_t = \beta - \beta\left(1 - \frac{1}{t}\right)^q, \tag{19}$$

where q is an integer (typically between 5 and 10) and β is a smoothing constant (typically between 0.8 and 0.99). Another approach is to inject extra variance into the sampling distribution, for example, by increasing the components of $\boldsymbol{\sigma}$, once the distribution has degenerated; see Botev and Kroese (2004). Finally, significant speed up can be achieved by



Cross-Entropy Method, Fig. 3 Simulated data for the FitzHugh–Nagumo model and a fitted curve obtained via the CE method

using a parallel implementation of CE; see, for example, Evans et al. (2007).

Example: Parameter Estimation for Differential Equations. Consider the FitzHugh–Nagumo differential equations:

$$\frac{\mathrm{d}V_t}{\mathrm{d}t} = c\left(V_t - \frac{V_t^3}{3} + R_t\right),$$
$$\frac{\mathrm{d}R_t}{\mathrm{d}t} = -\frac{1}{c}(V_t - a + bR_t), \qquad (20)$$

which model the behavior of certain types of neurons (Nagumo et al. 1962). Ramsay et al. (2007) consider estimating the parameters a, b, and c from noisy observations of (V_t) by using a generalized smoothing approach. The simulated data in Fig. 3 correspond to the values of V_t obtained from (20) at times 0, 0.05, ..., 20.0, adding Gaussian noise with standard deviation 0.5. The true parameter values are $a = 0.2$, $b = 0.2$, and $c = 3$. The initial conditions are $V_0 = -1$ and $R_0 = 1$.

Estimation of the parameters via the CE method can be established by minimizing the least-squares performance

$$S(\mathbf{x}) = \sum_{i=0}^{400} (y_i - V_{0.05i}(\mathbf{x}))^2,$$

where $\{y_i\}$ are the simulated data, $\mathbf{x} = (a, b, c, V_0, R_0)$, and $V_t(\mathbf{x})$ is the solution to (20) for parameter vector $\mathbf{x}$. Algorithm 3 was implemented with $\hat{\boldsymbol{\mu}}_0 = (0, 0, 5, 0, 0)$, $\hat{\boldsymbol{\sigma}}_0 = (1,1,1,1,1)$, $N = 100$, $N^e = 10$, and $\varepsilon = 0.001$. Constant smoothing parameters $\alpha_1 = 0.9$ and $\alpha_2 = 0.5$ were used for the $\{\hat{\boldsymbol{\mu}}_t\}$ and the $\{\hat{\boldsymbol{\sigma}}_t\}$, respectively. The following solution was found (note that the initial condition was assumed to be unknown): $\hat{a} = 0.19$, $\hat{b} = 0.21$, $\hat{c} = 3.00$, $\hat{V}_0 = -1.02$, and $\hat{R}_0 = 1.02$. The smooth curve in Fig. 3 gives the corresponding estimated curve, which is practically indistinguishable from the true one.

See

- Monte Carlo Methods
- Monte Carlo Simulation
- Rare Event Simulation
- Simulation of Stochastic Discrete-Event Systems
- Simulation Optimization

References

Alon, G., Kroese, D. P., Raviv, T., & Rubinstein, R. Y. (2005). Application of the cross-entropy method to the buffer allocation problem in a simulation-based environment. *Annals of Operations Research, 134*(1), 137–151.

Botev, Z. I., & Kroese, D. P. (2004). Global likelihood optimization via the cross-entropy method with an application to mixture models. *Proceedings of the 36th Winter Simulation Conference*, Washington, DC, pp. 529–535.

Chan, J. C. C. (2010). *Advanced Monte Carlo methods with applications in finance*. PhD thesis, University of Queensland.

Cohen, I., Golany, B., & Shtub, A. (2007). Resource allocation in stochastic, finite-capacity, multi-project systems through the cross entropy methodology. *Journal of Scheduling, 10*(1), 181–193.

Costa, A., Owen, J., & Kroese, D. P. (2007). Convergence properties of the cross-entropy method for discrete optimization. *Operations Research Letters, 35*(5), 573–580.

de Boer, P. T., Kroese, D. P., Mannor, S., & Rubinstein, R. Y. (2005). A tutorial on the cross-entropy method. *Annals of Operations Research, 134*(1), 19–67.

Evans, G. E., Keith, J. M., & Kroese. D. P. (2007). Parallel cross-entropy optimization. *Proceedings of the 2007 Winter Simulation Conference*, Washington, DC, pp. 2196–2202.
Kroese, D. P., Porotsky, S., & Rubinstein, R. Y. (2006). The cross-entropy method for continuous multi-extremal optimization. *Methodology and Computing in Applied Probability, 8*(3), 383–407.
Margolin, L. (2005). On the convergence of the cross-entropy method. *Annals of Operations Research, 134*(1), 201–214.
Nagumo, J., Arimoto, S., & Yoshizawa, S. (1962). An active pulse transmission line simulating nerve axon. *Proceedings of the IRE, 50*(10), 2061–2070.
Ramsay, J. O., Hooker, G., Campbell, D., & Cao, J. (2007). Parameter estimation for differential equations: A generalized smoothing approach. *Journal of the Royal Statistical Society, Series B, 69*(5), 741–796.
Rubinstein, R. Y. (1997). Optimization of computer simulation models with rare events. *European Journal of Operational Research, 99*(1), 89–112.
Rubinstein, R. Y. (1999). The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability, 1*(2), 127–190.
Rubinstein, R. Y. (2001). Combinatorial optimization, cross-entropy, ants and rare events. In S. Uryasev & P. M. Pardalos (Eds.), *Stochastic optimization: Algorithms and applications* (pp. 304–358). Dordrecht: Kluwer.
Rubinstein, R. Y., & Kroese, D. P. (2004). *The cross-entropy method: A unified approach to combinatorial optimization, Monte Carlo simulation and machine learning*. New York: Springer.
Rubinstein, R. Y., & Kroese, D. P. (2007). *Simulation and the Monte Carlo Method* (2nd ed.). New York: Wiley.

Crossover

A genetic-algorithm operator which exchanges corresponding genetic material from two parent chromosomes (i.e., solutions), allowing genes on different parents to be combined in their offspring.

See

► Genetic Algorithms

CS

Computer science.

See

► Computer Science and Operations Research Interfaces

Curse of Dimensionality

The situation that arises in such areas as dynamic programming, control theory, integer programming, combinatorial problems, and, in general, time-dependent problems in which the number of states and/or data storage requirements increases exponentially with small increases in the problems' parameters or dimensions; sometimes referred to as combinatorial explosion.

See

► Combinatorial Explosion
► Control Theory
► Dynamic Programming
► Integer and Combinatorial Optimization

References

Bellman, R. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press. (Dover Publications reprint 2003).

Customer Distribution

The probability distribution of the state of the process that customers observe upon arrival to a queueing system. In general, it is not the same as the distribution seen by a random outside observer; but the two distributions are the same for queueing systems with Poisson arrivals (PASTA). Since customers entering a queue must also exit, the probability distribution seen by arriving customers who are accepted is the same as that for the number of customers left behind by the departures.

See

► Outside Observer Distribution
► PASTA
► Queueing Theory

Cut

A set of arcs in a graph (network) whose removal eliminates all paths joining a node s (source node) to a node t (sink node).

See

- ▶ Graph Theory
- ▶ Max-Flow Min-Cut Theorem
- ▶ Maximum-Flow Network Problem

Cutset

A minimal set of edges whose removal disconnects a graph.

See

- ▶ Cut
- ▶ Graph Theory

Cutting Stock Problems

Robert W. Haessler
University of Michigan, Ann Arbor, MI, USA

Introduction

Solid materials such as aluminum, steel, glass, wood, leather, paper and plastic film are generally produced in larger sizes than required by the customers for these materials. As a result, the producers or primary converters must determine how to cut the production units of these materials to obtain the sizes required by their customers. This is known as a cutting stock problem. It can occur in one, two or three dimensions depending on the material. The production units may be identical, may consist of a few different sizes, or may be unique. They may be of consistent quality throughout or may contain defects. The production units may be regular (rectangular) or irregular. The ordered sizes may be regular or irregular. They may all have the same quality requirements or some may have different requirements. They may have identical or different timing requirements which impact inventory. The first

Some examples follow:

- cutting rolls of paper from production reels of the same diameter.
- cutting rectangular pieces of glass from rectangular production sheets.
- cutting irregular pieces of steel from rectangular plates.
- cutting rectangular pieces of leather from irregular hides.
- cutting dimensional lumber from logs of various size.

There are two other classes of problems which are closely related to the above cutting problems. The first is the layout problem. An example of this would be the problem of determining the smallest rectangle which will contain a given set of smaller rectangles without overlap. Solving this problem is essentially the same as being able to generate a cutting pattern in the discussion of cutting stock problems which follows. The second type of problem, which in many cases can be solved by the same techniques as cutting stock problems, is the (bin) packing problem. A one-dimensional example of this would be to determine the minimum number of containers required to ship a set of discrete items where weight and not floor space or volume is the determinant of what can be placed in the container. If floor space or volume is the key determinant, then the problem is equivalent to a two or three-dimensional cutting stock problem in which guillotine cuts are not required. Even though the following discussions focuses on cutting stock problems, it is also applicable to solving both packing and layout problems.

Although cutting stock problems are relatively easy to formulate, many of them especially those with irregular shapes, are difficult to solve; there are no efficient solution procedures available. The major difficulty has to do with the generation of feasible low trim loss cutting pattern. As will be discuseed below, this ranges from being simple in one-dimension to complex in two-dimensions, even with regular shapes.

The first known formulation of a cutting stock problem was given in 1939 by the Russian economist Kantorovich (1960). The first and most significant advance in solving cutting problems was the seminal

work of Gilmore and Gomory (1961, 1963) in which they described their delayed pattern generation technique for solving the one-dimensional trim loss minimization problem using linear programming. Since that time, there has been an explosion of interest in this application area. Sweeney and Paternoster (1992) have identified more than 500 papers which deal with cutting stock and related problems and applications. The primary reasons for this activity are that cutting stock problems occur in a wide variety of industries, there is a large economic incentive to find more effective solution procedures, and it is easy to compare alternative solution procedures and to identify the potential benefits of using a proposed procedure.

Cutting stock problems are introduced with a discussion of the one-dimensional problem and the techniques available for solving it. The article concludes with an extension to the regular two dimensional problem.

One-Dimensional Problems

An example of a one-dimensional cutting stock problem is the trim loss minimization problem that occurs in the paper industry. In this problem, known quantities of rolls of various widths and the same diameter are to be slit from stock rolls of some standard width and diameter. The objective is to identify slitting patterns and their associated usage levels that satisfy the requirements for ordered rolls at the least possible total cost for scrap and other controllable factors. The basic cutting pattern feasibility restriction in this problem is that the sum of the roll widths slit from each stock roll must not exceed the usable width of the stock roll.

Let R_i be the nominal order requirements for rolls of width W_i, $i = 1, \ldots, n$, to be cut from stock rolls of usable width UW. We have RL_i and RU_i as lower and upper bounds on the order requirement, for customer order i, reflecting the general industry practice of allowing overruns or underruns within specified limits. Depending on the situation, R_i may be equal to RL_i and/or RU_i. All orders are for rolls of the same diameter. This problem can be formulated as follows, with X_j as the number of stock rolls to be slit using pattern j and T_j as the trim loss incurred by pattern j:

$$\text{minimize} \sum_j T_j X_j \tag{1}$$

$$\text{s.t } RL_i \leq \sum_j A_{ij} X_j \leq RU_i \text{ for all } i \tag{2}$$

$$T_j = UW - \sum_i A_{ij} W_i \tag{3}$$

$$X_j \geq 0, \quad \text{integer.} \tag{4}$$

where A_{ij} is the number of rolls of width W_i to be slit from each stock roll that is processed using pattern j. In order for the elements A_{ij}, $i = 1, \ldots, n$, to constitute a feasible cutting pattern, the following restrictions must be satisfied:

$$\sum_i A_{ij} W_i \leq UW, \tag{5}$$

$$A_{ij} \geq 0, \quad \text{integer} \tag{6}$$

Note that the objective in this example is simply to minimize trim loss. In most industrial applications, it is necessary to consider other factors in addition to trim loss. For example, there may be a cost associated with pattern changes and, therefore, controlling the number of patterns used to satisfy the order requirements would be an important consideration.

Because optimal solutions to integer cutting stock problems can be found only for values of n smaller than typically found in practice, heuristic procedures represent the only feasible approach to solving this type of problem. Two types of heuristic procedures have been widely used to solve one-dimensional cutting stock problems. One approach uses the solution to a linear programming (LP) relaxation of the integer problem above as its starting point. The LP solution is then modified in some way to provide a integer solution to the problem. The second approach is to generate cutting patterns sequentially to satisfy some portion of the remaining requirements. This sequential heuristic procedure (SHP) terminates when all order requirements are satisfied.

Linear Programming Solutions

Almost all LP-based procedures for solving cutting stock problems can be traced back to Gilmore and Gomory (1961, 1963). They described how the next pattern to enter the LP basis could be found by solving

an associated knapsack problem. This made it possible to solve the trim loss minimization problem by linear programming without first enumerating every feasible slitting pattern. This is extremely important because a large number of feasible patterns may exist when narrow widths are to be slit from a wide stock roll. Pierce (1964) showed that in such situations the number of slitting patterns can easily run into the millions. Because only a small fraction of all possible slitting patterns need to be considered in finding the minimum trim loss solution, the delayed pattern generation technique developed by Gilmore and Gomory made it possible to solve trim loss minimization problems in much less time than would be required if all the slitting patterns were input to a general-purpose linear programming algorithm.

A common LP relaxation of the integer programming problem given in (1)–(3) can be stated as follows:

$$\text{minimize} \sum_j X_j \tag{7}$$

$$\text{s.t} \sum_j A_{ij} X_j \geq RU_i \text{ for all } i, \tag{8}$$

$$X_j \geq 0, \quad \text{integer}. \tag{9}$$

Let U_i be the dual variable associated with constraint i. Then the dual of this problem can be stated as:

$$\text{minimize} \sum_i R_i U_i \tag{10}$$

$$\text{s.t.} \sum_i A_{ij} U_i \leq 1 \tag{11}$$

$$U_i \geq 0. \tag{12}$$

The dual constraints in (11) provide the means for determining if the optimal LP solution has been obtained or if there exists a pattern which will improve the LP solution because the dual problem is still infeasible.

The next pattern $\boldsymbol{A} = (A_1, \ldots, A_n)$ to enter the basis, if one exists, can be found by solving the following knapsack problem:

$$Z = \text{maximize} \sum_i U_i A_i \tag{13}$$

$$\text{s.t.} \sum_i W_i A_i \leq UW \tag{14}$$

$$A_i \geq 0, \quad \text{integer} \tag{15}$$

If $Z \leq 1$, the current solution is optimal. If $Z > 1$, then **A** can be used to improve the LP solution.

Once found, the LP solution can be modified in a number of ways to obtain integer values for the X_j which satisfy the order requirements. One common approach is to round the LP solution down to integer values, then increase the values of X_j by unit amounts for any patterns whose usage can be increased without exceeding RU_i. Finally, new patterns can be generated for any rolls still needed using the sequential heuristic described in the next section.

Sequential Heuristic Procedures (SHP)

With an SHP, a solution is constructed one pattern at a time until all the order requirements are satisfied. The first documented SHP capable of finding better solutions than those found manually by schedulers was described by Haessler (1971). The key to success with this type of procedure is to make intelligent choices as to the patterns which are selected early in the SHP. The patterns selected initially should have low trim loss, high usage and leave a set of requirements for future patterns which will combine well without excessive side trim.

The following procedure is capable of making effective pattern choices in a variety of situations:

1. Compute descriptors of the order requirements yet to be scheduled. Typical descriptors would be the number of stock rolls still to be slit and the average number of ordered rolls to be cut from each stock roll.
2. Set goals for the next pattern to be entered into the solution. Goals should be established for trim loss, pattern usage, and number of ordered rolls in the pattern.
3. Search exhaustively for a pattern that meets those goals.
4. If a pattern is found, add this pattern to the solution at the maximum possible level without exceeding R_i, for all i. Reduce the order requirements and return to 1.
5. If no pattern is found, reduce the goal for the usage level of the next pattern and return to 3.

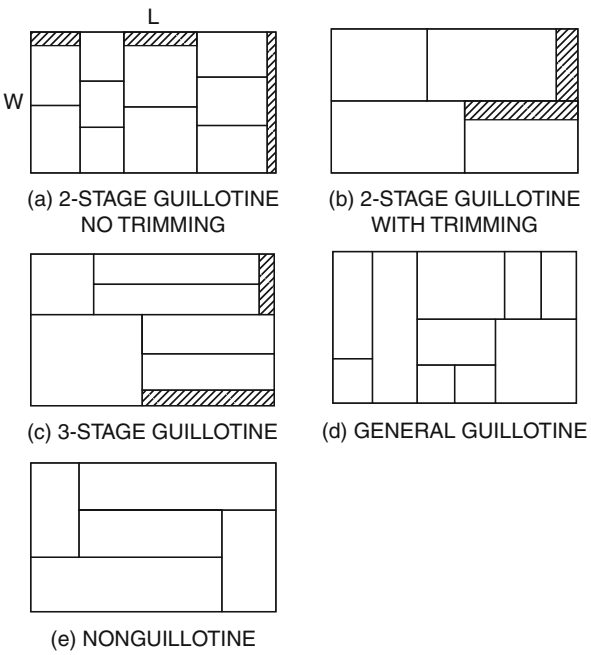
The pattern usage goal provides an upper bound on the number of times a size can appear in a pattern. For example, if some ordered width has an unmet requirement of 10 rolls and the pattern usage goal is 4, that width may not appear more than twice in a pattern. If after exhaustive search no pattern satisfies the goals set, then at least one goal, most commonly pattern usage, must be relaxed. This increases the number of patterns to be considered. If the pattern usage goal is changed to 3 in the above example, then the width can appear in the pattern three times. Termination can be guaranteed by selecting the pattern with the lowest trim loss at the usage level of one.

The primary advantage of this SHP is its ability to control factors other than trim loss and to eliminate rounding problems by working only with integer values. For example, if there is a cost associated with a pattern change, a sequential heuristic procedure which searches for high usage patterns may give a solution which has less than one-half the number of patterns required by an LP solution to the same problem. The major disadvantage of an SHP is that it may generate a solution which has greatly increased trim loss because of what might be called ending conditions. For example, if care is not taken as each pattern is accepted and the requirements reduced, the widths remaining at some point in the process may not have an acceptable trim loss solution. Such would be the case if only 34-inch rolls are left to be slit from 100-inch stock rolls.

Rectangular Two-Dimensional Problems

The formulation of a higher dimensional cutting stock problem is exactly the same as that of the one-dimensional problem given in (1)–(4). The only added complexity comes in trying to define and generate feasible cutting patterns. The simplest two-dimensional case is one in which both the stock and ordered sizes are rectangular. Most of the important issues regarding cutting patterns for rectangular two-dimensional problems can be seen in the examples shown in Fig. 1.

One important issue not covered in Fig. 1 is a limit on the number of times an ordered size can appear in a pattern. This generally is a function of the maximum quantity of pieces, RU_i, required for order i. If R_i is small, it is just as important for the two-dimensional case as the one dimensional case that the number of times size i appears in a pattern should be limited. This becomes less important as R_i becomes larger and as the difference between RU_i and RL_i becomes larger.



Cutting Stock Problems, Fig. 1 Sample cutting patterns

The cutting pattern shown in Fig. 1(a) is an example of two-stage guillotine cuts. The first cut can be in either the horizontal or vertical direction. A second cut perpendicular to the first, yields a finished piece. Figure 1(b) is similar except a third cut can be made to trim the pieces down to the correct dimension. Figure 1(c) shows the situation in which the third cut can create 2 ordered pieces.

For simple staged cutting such as shown in Fig. 1 (a, b, c), Gilmore and Gomory (1965) showed how cutting patterns can be generated by solving two one-dimensional knapsack problems. To simplify the discussion, assume that the orientation of each ordered piece is fixed relative to stock piece and the first guillotine cut on the stock pieces must be along the length (larger dimension) of the stock piece. For each ordered width W_k, find the contents of a strip of width W_k and length L which gives the maximum contribution to dual infeasibility:

$$Z_k = \text{maximize} \sum_{i \varepsilon I_k} U_i A_{ik} \qquad (16)$$

$$\text{s.t.} \sum_{i \varepsilon I_k} L_i A_{ik} \leq L \tag{17}$$

$$A_{ik} \geq 0, \text{ integer.} \tag{18}$$

$$I_k = \{i | W_i \leq W_k\}. \tag{19}$$

Next find the combination of strips which solve the problem

$$Z = \text{maximize} \sum_k Z_k A_k \tag{20}$$

$$\sum_k W_k A_k \leq W \tag{21}$$

$$A_k \geq 0, \quad \text{integer.} \tag{22}$$

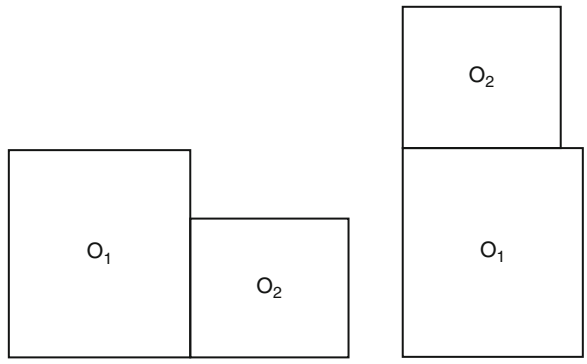
Any pattern for which Z is greater than one will yield an improvement in the LP solution.

The major difficulty with this approach is the inability to limit the number of times an ordered size appears in a pattern. It is easy to restrict the number of times a size appears in a strip and to restrict the number of strips in a pattern. The problem is that small ordered sizes with small quantities may end up as filler in a large number of different strips. This makes the two-stage approach to developing patterns ineffective when the number of times a size appears in a pattern must be limited.

Wang (1983) developed an alternative approach to generating general guillotine cutting patterns which limits the number of times a size appears in a pattern. She combined rectangles in a horizontal and vertical build process as shown in Fig. 2 where O_i is an ordered rectangle of width W_i and length L_i.

She used an acceptable value for trim loss, B, rather than the shadow price of the ordered sizes to drive her procedure which is as follows:

- Step 1
  - (a) Choose a value for B the maximum acceptable trim waste.
  - (b) Define $L^{(O)} = F^{(O)} = \{O_1, O_2, \ldots, O_n\}$, and set $K = 1$.
- Step 2
  - (a) Compute $F^{(K)}$ which is the set of all rectangles T satisfying (i) T is formed by a horizontal or vertical build of two rectangles from $L^{(K-1)}$, (ii) the amount of trim waste in T does not exceed B, and (iii) those rectangles O_i, appearing in T do not violate the constraints on the number of times a size can appear in a pattern.



(a) Horizontal build of O_1 and O_2 (b) Vertical build of O_1 and O_2

Cutting Stock Problems, Fig. 2 Guillotine cutting patterns

  - (b) Set $L^{(K)} = L^{(K-1)} UF^{(K)}$. Remove any equivalent (same component rectangles) rectangle patterns from $L^{(K)}$.
- Step 3. If $F^{(K)}$ is non-empty, set $K = K + 1$ and go to Step 2; otherwise, set $M = K - 1$, and choose the rectangle in $L^{(M)}$ which has the smallest total trim waste when placed in the stock rectangle.

Concluding Remarks

It is clear that moving from one to two-dimensions causes significant difficulty in the pattern generating process. This is all the more alarming in light of the fact that only rectangular shapes were considered.

This clearly suggests that there is much more research needed on procedures for solving two-dimensional cutting stock problems. An alternative worth considering, especially in those cases where there are many different ordered sizes with small order quantities, might be to first select a subset of orders to consider by solving a one-dimensional knapsack problem as in (13)–(15) based on area and then see if the resulting solution can be put together into a feasible two-dimensional pattern. Wang's algorithm seems to be ideal for this purpose inasmuch as the trim loss in the pattern would be known.

A candidate set of items to be included in the next pattern could be found by solving the following problem:

$$Z = \text{maximize} \sum_i U_i A_i \tag{23}$$

$$\sum_i AR_i A_i \leq UAR \text{ for all } i \tag{24}$$

$$A_i \leq b_i \tag{25}$$

$$A_i \geq 0, \quad \text{integer.} \tag{26}$$

where AR_i is the area of ordered rectangle i, UAR is the usable area of the stock rectangle, and b_i is the upper limit on the number of times order i can be included in the pattern.

The candidate pattern $(A_1, \ldots, A_n)$ could then be tested for feasibility using Wang's procedure. If the AR_i are small, the chances are that there will be little trim loss in the candidate patterns generated. This may require that UAR be reduced to force some trim loss to make it more likely that feasible patterns are found.

See

- ▶ Bin-Packing
- ▶ Integer and Combinatorial Optimization
- ▶ Linear Programming

References

Gilmore, P. C., & Gomory, R. E. (1961). A linear programming approach to the cutting stock problem. *Operations Research, 9*, 848–859.

Gilmore, P. C., & Gomory, R. E. (1963). A linear programming approach to the cutting stock problem, part II. *Operations Research, 11*, 863–888.

Gilmore, P. C., & Gomory, R. E. (1965). Multistage cutting stock problems of two and more dimensions. *Operations Research, 13*, 94–120.

Gilmore, P. C., & Gomory, R. E. (1966). The theory and computation of knapsack functions. *Operations Research, 14*, 1045–1074.

Haessler, R. W. (1971). A heuristic programming solution to a nonlinear cutting stock problem. *Management Science, 17*, 793–802.

Haessler, R. W., & Sweeney, P. E. (1991). Cutting stock problems and solution procedures. *European Journal of Operational Research, 54*, 141–150.

Kantorovich, L. V. (1960). Mathematical methods of organizing and planning production. *Management Science, 6*, 366–422. reprinted in.

Paull, A. E. (1956). Linear programming: A Key to optimum newsprint production. *Paper Magazine of Canada, 57*, 85–90.

Pierce, J. F. (1964). *Some large scale production problems in the paper industry*. Englewood Cliffs, NJ: Prentice-Hall Inc.

Sweeney, P. E., & Paternoster, E. R. (1992). Cutting and packing problems: A categorized, application-oriented research bibliography. *Journal of the Operational Research Society, 43*, 691–706.

Wang, P. Y. (1983). Two algorithms for constrained two-dimensional cutting stock problems. *Operations Research, 31*, 573–586.

C

CV

- ▶ Coefficient of Variation

Cybernetics and Complex Adaptive Systems

Andrew P. Sage
George Mason University, Fairfax, VA, USA

Introduction

Cybernetics is a term that is occasionally used in the literature of such areas as systems engineering and OR/MS to denote the study of control and communication in, and, in particular between humans, machines, organizations, and society. The word cybernetics comes from the Greek word *Kybernetes*, which means controller, or governor, or steersman. The first modern use of the term was due to Professor Norbert Wiener, an MIT professor of mathematics, who made many early and seminal contributions to mathematical system theory (Wiener 1949). The first book formally on this subject was titled *Cybernetics* and published in 1948 (Wiener 1948). In this book, Wiener defined the term as "control and communication in the animal and the machine." This emphasized the concept of feedback control as a construct presumably of value in the study of neural and physiological relations in the biological and physical sciences. In the historical evolution of cybernetics, major concern was initially devoted to the study of feedback control and servomechanisms, studies which later evolved into the area of control systems or control engineering (Singh 1990). Cybernetic concerns also have involved analog and digital computer development, especially computer

efforts that were presumed to be models of the human brain and the combination of computer and control systems for purposes of automation and remote control (Ashby 1952, 1956; George 1971; Lerner 1976).

There were a number of other early influences on cybernetics, including artificial intelligence (AI). The two are quite different subjects, however. Artificial Intelligence is generally concerned with endowing computers with machine intelligence such that they can emulate certain forms of human behavior, generally cognitive behavior. Cybernetics is an epistemological subject that is fundamentally concerned with limits on how we know what we know. It seeks to understand systems in a variety of media — technological, biological, social, or organizational — and descriptions of these limits as a most important result. So while AI seeks to endow computers with human cognitive capabilities, a subject associated with much controversy (Dreyfus 1992), cybernetics is much more concerned with using computational capabilities to develop models of systems based on the information, feedback, and control properties of these systems. In a cybernetic system, information and knowledge are attributes of interactions that occur within the system. It was the initial presumed resemblance, at a neural or physiological level, between physical control systems and the central nervous system and human brain that concerned Wiener. He and close associates, Warren McCulloch, Arturo Rosenblueth, and Walter Pitts, were the initial seminal thinkers in this new field of cybernetics. Soon, it became clear that it was fruitless to study control independent of information flow; cybernetics thus took on an identification with the study of communications and control in humans and machines. An influence in the early notions of cybernetics was the thought that physical systems could be made to perform better by, somehow, enabling them to emulate human systems at the physiological or neural level. Thus, early efforts in what is now known as neural networks began as cybernetic studies.

Another early concept explored in cybernetics was that of homeostasis, which has come to be known as the process by which systems maintain their level of organization in the face of disturbances, often occurring over time, and generally of a very large scale (Ashby 1952). Cybernetics soon became concerned with purposive organizational systems, or viable systems, as contrasted with systems that are static over time and purpose (Beer 1979). Further, organizations operate in the face of incomplete and redundant information by establishing useful patterns of communications (Beer 1979). Thus, organizations can potentially be modeled and have been modeled as cybernetic systems (Steinbrunner 1974).

Cybernetics has often been viewed as a way of looking at systems, or as a philosophical perspective concerning inquiry, as contrasted with a very specific method. This is perhaps much more the case now than during the very early history of use of the term. An excellent collection of Norbert Wiener's original papers on cybernetics studies is contained in Volume IV of an edited anthology (Masani 1985). Fundamental to any cybernetic study is the notion of modeling, and, in particular, the interpretation of the results of a modeling effort as theories that have normative or predictive value. Today, there is little explicit or implicit agreement concerning a precise definition for cybernetics. Some users of the term cybernetics infer that the word implies a study of control systems. Some uses refer to modeling only at the neural and physiological level. Some refer to cognitive ergonomic modeling without necessary consideration of, or connection to, neuronal level elements. Other uses of the word are so general that cybernetics might seem to infer either nothing, or everything. Automation, robotics, artificial intelligence, information theory, bionics, automata theory, pattern recognition and image analysis, control theory, communications, human and behavioral factors, and other topics have all, at one time or another, been assumed to be a portion of cybernetics.

Complex adaptive systems (CAS) involve phenomena associated with interactions of many individual agents that self-organize at higher aggregate system levels. This results in emergent and adaptive properties that are not exhibited by the individual agents. These systems are cybernetic like systems that receive data and information from their environments, find regularities in the data and information, and then identify internal models that process this data and information in order to describe and forecast likely futures. These systems are

evolutionary in the sense that these internal models are subject to selection pressures based on particular environmental conditions and this results in changes to the structure and parameters associated with the internal models. These systems function best under conditions between chaos and order, sometimes referred to as at the edge of chaos (Langton 1990) or self-organized criticality (Bak and Chen 1991; Bak 1996). The emergent characteristics of a particular complex system (Holland 1996) are often equivalent to individual agents acting in a higher level complex system. Adaptation occurs when either the functional or structural properties of an agent change in such a manner as to improve survival probabilities in the environment of the agent. Often the only way to study these complex adaptive systems is through computer simulation.

Definition of Cybernetics and Complex Adaptive Systems

The notion of the physiological aspects of the human nervous system as playing a necessarily critical role in modern cybernetics has all but vanished, except in very specialized classic works. This does not suggest that interest in neural type studies has vanished as there is much interest today in neural networks and related subjects (Freeman and Skapura 1991; Zurada 1992). A much more cognitive perspective is now prevalent, at least in many systems engineering views of cybernetics. In this article, cybernetics is defined as the study of the communication and control processes associated with human-machine interaction in systems that are intended to support accomplishment of purposeful tasks. While this is not a universally accepted definition of cybernetics, it is a useful one for many systems engineering studies involving human-system interaction through communications and control (Sage 1992). Complex systems theory is a general approach to understanding the overall behavior of system comprised of many nonlinearly interacting parts. The complex systems approach tries to construct minimal underlying rule sets from which desired behaviors naturally emerge. Complex adaptive system theory also assumes that systems are composed of interacting agents that continually adapt by changing their internal rules as the environment, and their experience in that environment, evolve over time. Systems transition naturally between equilibrium points through environmental adaptation and self-organization. A complex adaptive system behaves and evolves according to three key principles: (1) order is emergent as opposed to predetermined, (2) the system's history is irreversible, and (3) the system's future is generally unpredictable. Complex adaptive systems are complex systems consisting of many nonlinearly interacting parts or agents. These agents can adapt to changing environments where each agent typically exists within a nested hierarchy of agents within agents.

The purpose of this article is to discuss cybernetics and complex adaptive systems, and the design of support systems based on these concepts for such purposes as knowledge support to humans. Especial concern is with the human-system interactions that occur in such an effort. Thus, the discussions here are particularly relevant to knowledge-based system design concerns relative to human-machine cybernetic problem solving tasks, such as fault detection, diagnosis and correction. These are very important concerns for a large number of knowledge-support systems engineering applications that require fundamentally cognitive support to humans in supervisory control tasks (Sage 1991, 1992, 1995; Sheridan 1992; Rasmussen 1986; Rasmussen et al. 1994).

The need for humans to monitor and maintain the conditions necessary for satisfactory operation of systems and to cope with poorly structured and imprecise knowledge is greater than ever. Ultimately, these primarily cognitive efforts, which involve a great variety of human problem solving activities, are often translated into physical control signals for controlling or manipulating some physical process. As a consequence of this, there are a number of human interface issues that naturally occur between the human and the machines over which the human must exercise control. Many advances in information technology result in systems that enable a significant increase in the amount of information that is available for judgment and decision-making tasks at the problem solving level. Even the highest quality information, however, will generally be associated with considerable uncertainty, imprecision, and other forms of imperfections. Above all else, there is

a major need for information to be associated with context such that it becomes knowledge useful for judgment and choice. The contemporary use of information technology has led to and is expected to continue to lead to major organizational transformations in the future (Harrington 1991; Scott Morton 1991; Davenport 1993; Drucker 1995, 1998.

Cybernetics, Complex Adaptive Systems, and Systems Management

A human-machine cybernetic system may be defined as a functional synthesis of a human system and a technological system or machine. The interaction and functional interdependence between these two elements pre-dominantly characterize human-machine systems. The introduction of communication and control concerns results in a cybernetic system. All kinds of technological systems, regardless of their degree of complexity, may be viewed as parts of a human-machine cybernetic system: industrial plants, vehicles, manipulators, prostheses, computers or management information systems. A human-machine system may, of course, be a subsystem that is incorporated within another system. For example, a decision support system may be incorporated as part of a larger enterprise management, process control, or computer-aided design system that also involves human interaction. This use of the term human-machine cybernetic system corresponds, therefore, to a specific way of looking at technological systems through the integration of technological systems and human-enterprise systems, generally through a systems management or systems engineering process.

The overall purpose of any human-machine cybernetic system is to provide a certain function, product, or service with reasonable cost under constraint conditions and disturbances. This concept involves and influences the human, the machine, and the processes through which they function as an integrated whole. The primary inputs to a human-machine cybernetic system are a set of purposeful performance objectives that are typically translated into a set of expected values of performance, costs, reliability, and safety. Also, the design must be such that an acceptable level of workload and job satisfaction is maintained. It is on the basis of these that the human is able to perform the following activities (Sage 1992):

1. Identify task requirements, such as to enable determination of the issues to be examined further and the issues to be not considered;
2. Identify a set of hypotheses or alternative courses of actions which may resolve the identified issues to be resolved;
3. Identify the probable impacts of the alternative courses of action;
4. Interpret these impacts in terms of the objectives or inputs to the task;
5. Select an alternative for implementation and implement the resulting control;
6. Monitor performance such as to enable determination of how well the integrated combination of human and system are performing.

Many researchers have described activities of this sort in a number of frameworks that include behavioral psychology, organizational management, human factors, systems engineering, operations research and management science.

Many questions can be raised concerning the use of information for judgment and choice activities, as well as activities that lead to the physical control of an automated process. Any and all of these questions can arise in different application areas. These questions relate to the control of technological systems. They concern the degree of automation with respect to flexible task allocation. They also concern the design and use of computer-generated displays. Further, they relate to all kinds of human-computer interaction concerns, as well as management tasks at different organizational levels: strategic, tactical, and operational. For example, computer-based support systems to aid human performance continue to invade more and more areas of the engineering of systems: design, operation, maintenance, and management. The importance of augmenting hardware and micro-level programming aspects of system design to architectural and software systems management considerations is great. The integrated consideration of systems engineering and systems management for software productivity is expressed by the term software systems engineering (Sage and Palmer 1990).

Human tasks in human-machine cybernetic systems can be condensed into three primary categories: (1) controlling (physiological); (2) communicating

(cognitive), and (3) problem solving (cognitive) (Johannsen et al. 1983). In addition, there exists a monitoring or feedback portion of the effort that enables learning over time. Ideally, but not always, the humans involved learn well. There needs to be metalevel learning, or learning how to learn if improvements are to truly be lasting, as contrasted with only specific task performance learning. Associated with the rendering of a single judgment and the associated control implementation, the human monitors the result of the effect of these activities. The effect of present and past monitoring is to provide an experiential base for present problem conceptualization. In the categorization above, activities 1 through 4 may be viewed as problem (finding and) solving, activity 5 involves implementation or controlling, and activity 6 involves communications or monitoring and feedback in which responses to the question "How good is the process performance?" enables improvement and learning through iteration. Of course, the notion of information flow and communication is involved in all of these activities.

These three human task categories are fairly general. Controlling should be understood in a much broader sense than in many control theory studies. Controlling in this narrower sense includes open loop vs. closed-loop and continuous vs. intermittent controlling, as well as discrete tasks such as reaching, switching and typing. It is only through these physiological aspects of controlling that outputs of the human-machine cybernetic system can be produced. Controlling, in the sense of the cognitive ergonomic concerns that support human information processing and associated judgment and choice, is also included. Although human functions on a cognitive level can and do play a role in control implementation, their major importance occurs in problem solving activities. Tasks such as fault detection, fault diagnosis, fault compensation or managing, and planning are particularly important in problem solving. Fault detection concerns the identification of a potential difficulty concerning the operation of a system. Fault diagnosis is concerned with identification of a set of hypotheses concerning the likely cause of a system malfunction, and the evaluation and selection of a most likely cause. It is primarily a cognitive activity. Fault compensation or managing is concerned with solving problems in actual failure situations. This may occur through the use of rules that are based on past experience, and the updating of certain rules based on the results of their present application. It is accomplished with the objective of returning the overall system to a good operating state. Fault compensation or managing involves both cognitive and physiological activities. Planning is a cognitive activity concerned with solving possible future problems in the sense of mentally generating a sequence of appropriate alternatives. Appropriate planning involves the use of knowledge perspectives, knowledge principles, and knowledge practices (Sage 1992). They are based on experiential familiarity with analogous situations and are often expressed in the form of and through the use of skills, rules, and formal knowledge based reasoning efforts (Rasmussen 1986; Rasmussen et al. 1994). Human error issues are of particular importance, especially those concerned with the design of systems that cope with human error through avoidance and amelioration efforts (Reason 1990).

Many of these systems can only be described as complex. While some components of them may be naturally adaptive, they often need to be engineered to possess adaptive characteristics. The subject of complex adaptive systems is closely related to that of complexity theory. Complexity theory (Kaufman 1995; Axelrod 1997; Holland 1998) is a field of study that has evolved from five major knowledge areas: mathematics, physics, biology, organizational science, and computational intelligence and engineering. Fundamentally, a system is complex if it cannot be understood through simple cause-and-effect relationships or other standard methods of systems analysis. In a complex system, the interplay of individual elements cannot be reduced to the study of individual elements considered in isolation. Often, several different models of the complete system, each at a different level of abstraction, are needed.

There are several sciences of complexity, and they generally deal with approaches to understanding the dynamic behavior of units that range from individual organisms to the largest technical, economic, social, and political organizations. Often, such studies involve complex adaptive systems and hierarchical systems, are multidisciplinary in nature, and involve or are at the limits of scientific knowledge (Arthur 1994; Coveney and Highfield 1995; Arthur et al. 1997; Epstein 1997).

Complexity studies attempt to pursue knowledge and discover features shared by systems described as complex. These include studies such as complex adaptive systems, complex systems theory, complexity theory, dynamic systems theory, complex nonlinear systems, and computational intelligence. Many scientific studies, prior to the development of simulation models and complexity theory, involved the use of linear models. When a study resulted in anomalous behavior, the failure was often incorrectly blamed on noise or experimental error. It is now recognized that such errors may reflect inherent inappropriateness of linear models — and linear thinking. Meeting the modeling challenge is complicated by the fact that not all critical phenomena cannot be fully understood, or even anticipated, based on analysis of the decomposed elements of the overall system. Complexity not only arises from there being many elements of the system, but also from the possibility of collective behaviors that even the participants in the system could not have anticipated (Casti 1997).

Thus, many critical phenomena can only be studied once they emerge. In other words, the only way to identify such phenomena is to let them happen. The challenge is to create ways to recognize the emergence of unanticipated phenomena and be able to manage their consequences, especially in situations where likely consequences are highly undesirable. One measure of system complexity is the complexity of the simulation model necessary to effectively predict system behavior (Casti 1997). The more the simulation model must embody the actual system to yield the same behavior, the more complex the system. In other words, outputs of complex systems cannot be predicted accurately based on models with typical types of simplifying assumptions. Consequently, creating models that will accurately predict the outcomes of complex systems is very difficult. A model can be created, however, that will accurately simulate the processes the system will use to create a given output.

This awareness has profound impacts for organizational efforts. For example, it raises concerns related to the real value of creating organizational mission statements and plans with expectations that these plans will be inexorably executed and missions thereby realized. It may be more valuable to create a model of an organization's planning processes themselves, subject this model to various input scenarios, and use the results to generate alternative output scenarios. The question then becomes one of how to manage an organization where this range of outputs is possible.

Interestingly, most studies of complex systems often run completely counter to the trend toward increasing fragmentation and specialization in most disciplines. Complexity studies tend to reintegrate the fragmented interests of most disciplines into a common pathway. This needed transdisciplinarity (Wilson 1998) provides the basis for creating a cohesive systems ecology (Sage 1998) to guide the use of information technology for managing complex systems. Whether they be human-made systems, human systems, or organizational systems, the use of systems ecology could more quickly lead to organizing for complexity (McMaster 1996), and associated knowledge and enterprise integration.

An important aspect of complex systems is path dependence (Arthur 1994). The essence of this phenomenon begins with a supposedly minor advantage or inconsequential head start in the marketplace for some technology, product, or standard. This minor advantage can have important and irreversible influences on the ultimate market allocation of resources, even if market participants make voluntary decisions and attempt to maximize their individual benefits. Such a result is not plausible with classical economic models that assume that the maximization of individual gain leads to market optimization unless the market is imperfect due to the existence of such effects as monopolies. Path dependence is a failure of traditional market mechanisms and suggests that users are locked into a sub-optimal product, even though they are aware of the situation and may know that there is a superior alternative.

This type of path lock-in is generally attributed to two underlying drivers: 1) network effects, and 2) increasing returns of scale. Both of these drivers produce the same result, namely that the value of a product increases with the number of users. network effects, or network externalities, occur because the value of a product for an individual consumer may increase with increased adoption of that product by other consumers. This, in turn, raises the potential value for additional users. An example is the telephone, which is only useful if at least one other person has one as well, and becomes increasingly

beneficial as the number of potential users of the telephone increases.

Increasing returns of scale imply that the average cost of a product decreases as higher volumes are manufactured. This effect is a feature of many knowledge-based products where high initial development cost dominates low marginal production and distribution cost. Thus, the average cost per unit decreases as the sales volume increases and the producing company is able to continuously reduce the price of the product. The increasing returns to scale, associated with high initial development costs and the low sales price, create barriers against market entry by new potential competitors, even though they may have a superior product.

The controversy in the late 1990s over the integration of the Microsoft Internet Explorer with the Windows Operating System may be regarded as a potential example of path dependence, and appropriate models of this phenomenon can potentially be developed using complexity theory. These would allow exploration of whether network effects and increasing returns of scale can potentially reinforce the market dominance of an established but inferior product in the face of other superior products, or whether a given product is successful because its engineers have carefully and foresightedly integrated it with associated products such as to provide a seamless interface between several applications.

Information technology enables systems where the interactions of many loosely structured elements can produce unpredictable and uncertain responses that may be difficult to control. The challenge is to understand such systems at a higher level. Control is likely to involve design and manipulation of incentives to participate and rewards for collaborative behaviors. It may be impossible and probably undesirable to control behaviors directly. The needed type of control is similar to policy formulation. Success depends on efficient experimentation much more than possibilities for mathematical optimization due to the inherent complexities that are involved. Thus, insights from complexity theory may be brought to bear on these situations (Merry 1995).

Information access and utilization, as well as management of the knowledge resulting from this process, are complicated in a world with high levels of connectivity and a wealth of data, information, and knowledge. The underlying problem is the usually tacit assumption that more information is inherently good to have. What users should do with this information and how value is provided by this usage are seldom clear. The result can be large investments in information technology with negligible improvements of productivity (Harris 1994). One of the major needs in this regard is to support bilateral transformations between tacit and explicit knowledge (Nonaka 1994; Nonaka and Takeuchi 1995).

Prior to the development of simulation models and complexity theory, most studies involved use of linear models and assumed time-invariant processes (i.e., ergodicity). Most studies also assumed that humans use deductive reasoning and techno-economic rationality to reach conclusions. But, information imperfections and limits on available time often suggest that rationality must be bounded. Other forms of rationality and inductive reasoning are necessary.

There are a number of descriptive models of human problem solving and decision making. Generally, the appropriate model depends upon the contingency task structure, characteristics of the environment, and the experiential familiarity of humans with tasks and environment. Thus, the context surrounding information and the experiential familiarity of users of the information is most important. In fact, it is the use of information within the context of contingency task structures and the environment that results in the transformation from information to knowledge.

It is appropriate to interpret knowledge in terms of context and experience by sensing situations and recognizing patterns. Features similar to previously recognized situations can thus be discerned. The problem can then be simplified by using these to construct internal models, hypotheses, or schemata to use on a temporary basis. Simplified deductions are attempted based on these hypotheses and one acts accordingly. Feedback of results from these interactions enables more to be learned about the environment and the nature of the task at hand. Hypotheses are revised, reinforcing appropriate ones and discarding poor ones. This use of simplified models is a central part of inductive behavior (Holland et al. 1986).

Models of inductive processes can be constructed in the following way. A collection of generally heterogeneous agents is first determined. It is assumed that the agents are able to form hypotheses based on mental models or subjective beliefs. Further, each agent is assumed to monitor performance relative to a personal

set of belief models. These models are based on the results of actions, as well as prior beliefs and hypotheses. Through this iterative procedure, learning takes place as agents discern which hypotheses are most appropriate. Hypotheses, or models, are retained not because they are correct, but because they have worked in the past. Agents differ in their approach to problems and the way in which they subjectively converge to a set of useful hypotheses.

This process may be modeled as a complex adaptive system. As noted, models cannot be created that will accurately predict the outcomes of many complex systems. But, a model can often be created that will accurately simulate the processes the system uses to create outputs. The major constructs associated with such models are: the interactions and feedback relations between the various agents whose choices depend upon the decisions of others, and linearity and return to scale considerations. There are many implications associated with these models. Among them are questions of steady state versus continued evolutionary behavior, the nature and possibility of time-invariant processes (ergodicity), and questions of path dependence.

The Design of Cybernetic and Complex Adaptive Systems

All of this has major implications with respect to the design of systems for the human user and for associated cybernetic and complex adaptive systems as well. It requires, for appropriate system design, an understanding of human performance in problem solving and decision-making tasks. This understanding has to be at a descriptive level, predicting what humans will likely do in particular situations. It has to be at a normative level, understanding what would be best performance under restrictive axiomatic conditions that will generally not exist in practice. Also, this understanding has to be at a prescriptive level such that humans can be aided in various real-world cognitive tasks. This requires much attention to the evolutionary and emergent properties of systems.

Technological advances have changed and will continue to change the specific design requirements for human-machine cybernetic systems needed in any given application area. This is especially true due to the many advances made possible through modern information technologies, for industrial plants with integrated automated manufacturing capabilities, and for aids to cognitive activities in strategic planning, design, or operational activities. Office automation systems and information systems for observation, planning, executive support, management, and command and control tasks in business, defense, and medicine are similarly influenced by efforts in human-machine cybernetic and complex adaptive systems. These involve not only the operation of technological and management oriented information systems by highly skilled and knowledgeable personnel, but also systems that are intended for use by the less skilled. A major use for new generation systems is to provide computer assistance for the maintenance of existing systems and for the design of new systems of all types.

The methods and tools for supporting emergence of a theory of complex systems that will fully satisfy the requirements posed by systems that must intentionally operate satisfactorily at the edge of chaos will always be in a state of continuous evolution. There seems little question that the methods of operations research and management science, especially those associated with modeling and simulation of large systems, have and will play a major role in the theory of design of cybernetic and complex adaptive systems. Addressing the key challenges requires utilizing many of the concepts, principles, methods, and tools of OR/MS. In addition, it will require a new, broader perspective on the nature of information access and utilization, as well as knowledge management. Fortunately, OR/ MS is an inherently dynamic field of study. However, achieving the goal of cybernetic and complex adaptive system understanding and development capability will require much attention to the integration of OR/MS approaches with those in disciplines not often involved in OR/MS studies and the development of knowledge unity and integration perspectives.

See

- ▶ Artificial Intelligence
- ▶ Control Theory
- ▶ Dynamic Programming
- ▶ Neural Networks
- ▶ Simulation Metamodeling

- ▶ Simulation of Stochastic Discrete-Event Systems
- ▶ Simulation Optimization
- ▶ System Dynamics
- ▶ Systems Analysis

References

Arthur, W. B. (1994). *Increasing returns and path dependence in the economy*. Ann Arbor, MI: University of Michigan Press.

Arthur, W. B., Durlauf, S. N., & Lane, D. A. (Eds.). (1997). *The economy as an evolving complex system, II*. Reading, MA: Addison Wesley.

Ashby, W. R. (1952). *Design for a brain*. London: Chapman and Hall.

Ashby, W. R. (1956). *An introduction to cybernetics*. London: Chapman and Hall.

Axelrod, R. (1997). *The complexity of cooperation: Agent based models of competition and collaboration*. Princeton, NJ: Princeton University Press.

Bak, P., & Chen, K. (1991). Self organized criticality. *Scientific American, 271*(1), 46–53.

Bak, P. (1996). *How nature works: The science of self-organized criticality*. New York: Springer.

Barr, A., Cohen, P. R., & Feigenbaum, E. A. (Eds.). (1981/1982). *Handbook of artificial intelligence*, *Vols. I*, *II*, *and III*. Los Altos, CA: William Kaufman.

Beer, S. (1979). *The heart of enterprise*. Chichester, UK: Wiley.

Casti, J. L. (1997). *Would-be worlds how simulation is changing the frontiers of science*. New York: Wiley.

Coveney, P., & Highfield, R. (1995). *Frontiers of complexity: The search for order in a chaotic world*. Columbine, NY: Fawcett.

Davenport, T. H. (1993). *Process innovation: Reengineering work through information technology*. Boston: Harvard Business School Press.

Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. Cambridge, MA: MIT Press.

Drucker, P. (1995). *Managing in a time of great change*. New York: Dutton.

Drucker, P. (1998). *On the profession of management*. Boston: Harvard Business School Press.

Epstein, J. M. (1997). *Nonlinear dynamics, mathematical biology, and social science*. Reading, MA: Addison-Wesley.

Freeman, J. A., & Skapura, D. (1991). *Neural networks: Algorithms, applications and programming techniques*. Reading, MA: Addison-Wesley.

George, F. H. (1971). *Cybernetics*. Middlegreen, Slough, UK: St. Paul's House.

Harrington, H. J. (1991). *Business process improvement: The breakthrough strategy for total quality, productivity, and competitiveness*. New York: McGraw-Hill.

Harris, D. H. (Ed.). (1994). *Organizational linkages: Understanding the productivity paradox*. Washington, DC: National Academy Press.

Holland, J. H. (1996). *Hidden order: How adaptation builds complexity*. Reading, MA: Addison-Wesley.

Holland, J. H. (1998). *Emergence: From chaos to order*. Reading, MA: Addison-Wesley.

Holland, J. H., Holyoak, K. J., Nisbet, R. E., & Thagard, P. R. (1986). *Induction: Processes of inference, learning, and discovery*. Cambridge, MA: MIT Press.

Johannsen, G., Rijnsdorp, J. E., & Sage, A. P. (1983). Human interface concerns in support system design. *Automatica, 19*(6), 1–9.

Kaufman, S. (1995). *At home in the universe: The search for the laws of self-organization and complexity*. New York: Oxford University Press.

Langton, C. G. (1990). Computation at the edge of chaos: Phase transitions and emergent computation. *Physica D: Nonlinear Phenomena, 42*(1–3), 12–37.

Lerner, A. Y. (1976). *Fundamentals of cybernetics*. New York: Plenum.

Masani, P. (Ed.). (1985). *Norbert Wiener: Collected works volume IV–cybernetics, science and society; ethics, aesthetics, and literary criticism; book reviews and obituaries*. Cambridge, MA: MIT Press.

McMaster, M. D. (1996). *The intelligence advantage: Organizing for complexity*. Boston: Butterworth-Heinnemann.

Merry, U. (1995). *Coping with uncertainty: Insights from the new sciences of chaos, self-organization, and complexity*. Westport, CT: Praeger.

Nonaka, I. (1994). A dynamical theory of organizational knowledge creation. *Organizational Science, 5*(1), 14–37.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge creating company*. New York: Oxford.

Rasmussen, J. (1986). *Information processing and human machine interaction: An approach to cognitive engineering*. Amsterdam: North Holland Elsevier.

Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering*. New York: Wiley.

Reason, J. (1990). *Human error*. Cambridge, UK: Cambridge University Press.

Rockart, J. F., & DeLong, D. W. (1988). *Executive support systems: The emergence of top management computer use*. Homewood, IL: Dow Jones-Irwin.

Sage, A. P. (Ed.). (1987). *System design for human interaction*. New York: IEEE Press.

Sage, A. P. (1991). *Decision support systems engineering*. New York: Wiley.

Sage, A. P. (1992). *Systems engineering*. New York: Wiley.

Sage, A. P. (1995). *Systems management: For information technology and software engineering*. New York: Wiley.

Sage, A. P. (1998). Towards a systems ecology. *IEEE Computer, 31*(2), 107–110.

Sage, A. P., & Palmer, J. D. (1990). *Software systems engineering*. New York: Wiley.

Sage, A. P. (Ed.). (1990). *Concise encyclopedia of information processing in systems and organizations*. Oxford, UK: Pergamon Press.

Scott Morton, M. S. (Ed.). (1991). *The corporation of the 1990s: Information technology and organizational transformation*. New York: Oxford University Press.

Shapiro, S. C. (Ed.). (1987). *Encyclopedia of artificial intelligence*. New York: Wiley.

Sheridan, T. B. (1992). *Telerobotics, automation, and human supervisory control*. Cambridge, MA: MIT Press.

Sheridan, T. B., & Ferrell, W. R. (1974). *Man–machine systems: Information, control, and decision models of human performance*. Cambridge, MA: MIT Press.

Singh, M. G. (Ed.). (1990). *Systems and control encyclopedia*. Oxford, UK: Pergamon Press.
Steinbrunner, J. D. (1974). *The cybernetic theory of decision*. Princeton, NJ: Princeton University Press.
Wiener, N. (1948). *Cybernetics, or control and communication in the animal and the machine*. New York: Wiley.
Wiener, N. (1949). *Extrapolation, interpolation and smoothing of stationary time series with engineering applications*. Cambridge, MA: MIT Press.
Wilson, E. O. (1998). *Consilience: The unity of knowledge*. New York: Alfred A. Knopf.
Zurada, J. (1992). *Introduction to artificial neural systems*. St. Paul, MN: West Publishing.

Cycle

A path in a graph (network) joining a node to itself.

See

- ▶ Chain
- ▶ Path

Cyclic Queueing Network

A closed network of queues in which customer routing is serial.

See

- ▶ Networks of Queues
- ▶ Queueing Theory

Cyclic Service Discipline

When a congestion system with several different locations (service centers) of customers are served by a single service facility. For a given period of time determined by an a priori rule, the service process only works on customers from (at) a given location and then switches to the next group when the period is over.

See

- ▶ Queueing Theory

Cycling

A situation where the simplex algorithm cycles (circles) repeatedly through some sequence of bases and corresponding basic feasible solutions. This can occur at a degenerate extreme point solution where several bases correspond to the same extreme point.

See

- ▶ Anticycling Rules
- ▶ Degeneracy
- ▶ Linear Programming
- ▶ Simplex Method (Algorithm)