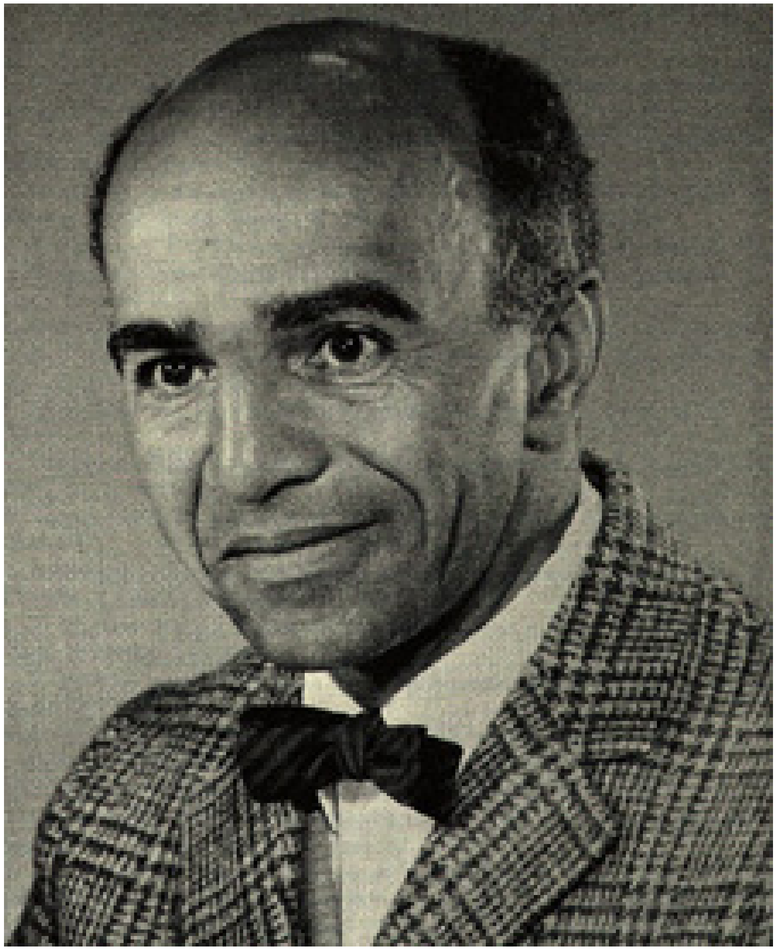
David Blackwell and Dynamic Programming

Bill Sudderth
University of Minnesota

Statistics 169, Dynamic Programming

Blackwell taught a course on dynamic programming at U.C. Berkeley in the 1960's. It was taken by engineers, operations researchers, statisticians, and mathematicians among others. I took the course in 1965. **It was a great course!**

The course met once a week for about two hours.



An example: Spend-or-Save

You begin with s_1 dollars, choose $a_1 \in [0, s_1]$ to spend on consumption, and save $s_1 - a_1$.

You receive $u(a_1)$ in utility, and begin the next stage with cash

$$s_2 = s_1 - a_1 + Y_1.$$

Here Y_1 is your random income and has a given distribution. You then choose $a_2 \in [0, s_2]$, and so on.

Future stages are discounted at rate $\beta \in (0, 1)$, and you want to maximize the expectation of

$$\sum_{n=1}^{\infty} \beta^{n-1} u(a_n).$$

Dynamic Programming (Markov Decision Theory)

Five ingredients: S, A, r, q, β.

Begin at state $s_1 \in S$, select an action $a_1 \in A$, receive a reward $r(s_1, a_1)$.
Move to a new state s_2 with distribution $q(\cdot|s_1, a_1)$. Select $a_2 \in A$, receive $\beta \cdot r(s_2, a_2)$.
Move to s_3 with distribution $q(\cdot|s_2, a_2)$, select $a_3 \in A$, receive $\beta^2 \cdot r(s_3, a_3)$. And so on.

Your total reward is the expected value of

$$\sum_{n=1}^{\infty} \beta^{n-1} r(s_n, a_n).$$

Three Conditions to make $\sum_{n=1}^{\infty} \beta^{n-1} r(s_n, a_n)$ well-defined

Discounted Problems: r bounded, $0 \leq \beta < 1$. Blackwell (1962,1965)

Positive Problems: $r \geq 0$, $\beta = 1$. Blackwell (1967)

Negative Problems: $r \leq 0$, $\beta = 1$. Strauch (1966)

Plans and Rewards

A **plan** π selects each action a_n as a function of the history $(s_1, a_1, \ldots, a_{n-1}, s_n)$. The **reward** from π at the initial state $s_1 = s$ is

$$V(\pi)(s) = E_{\pi,s}[\sum_{n=1}^{\infty} \beta^{n-1} r(s_n, a_n)].$$

The **optimal reward** at s is

$$V^*(s) = \sup_{\pi} V(\pi)(s).$$

Basic problems: Calculate the optimal reward function $V^*(\cdot)$ and find optimal or nearly optimal plans.

Stationary Plans

A stationary plan is one that ignores the past when selecting an action.

Formally, a plan π is **stationary** if there is a function $f : S \mapsto A$ such that $\pi(s_1, a_1, \ldots, a_{n-1}, s_n) = f(s_n)$ for all $(s_1, a_1, \ldots, a_{n-1}, s_n)$.

Notation: $\pi = f^\infty$.

Fundamental Question: Do optimal or nearly optimal stationary plans exist?

Discrete Discounted Dynamic Programming

Theorem 1 *(Blackwell, 1962) If S and A are finite and $0 \leq \beta < 1$, then there is an optimal stationary plan. Indeed, there is a stationary plan that is optimal for all β sufficiently close to 1.*

A plan satisfying the final phrase is now called **Blackwell optimal**. (Hordijk and Yushkevich (2002))

Blackwell Operators for Discounted Problems

Assume: S, A countable, r bounded, $0 \leq \beta < 1$.

Let $\mathbb{B}$ be the Banach space of bounded functions $x : S \mapsto \mathbb{R}$ equipped with the supremum norm.

Let $f : S \mapsto A$ and $\pi = f^\infty$. Define operators T_f and U for $x \in \mathbb{B}$:

$$(T_f x)(s) = r(s, f(s)) + \beta \int x(s')\, q(ds'|s, f(s)),$$

$$(Ux)(s) = \sup_a [r(s, a) + \beta \int x(s')\, q(ds'|s, a)].$$

Theorem 2 *The operators T_f and U are β-contractions on $\mathbb{B}$. The fixed point of T_f is the reward function $V(\pi)(\cdot)$ for $\pi = f^\infty$, the fixed point of U is the optimal reward function $V^*(\cdot)$.*

The Bellman Equation

For $s \in S$, $V^*(s) = UV^*(s)$, or

$$V^*(s) = \sup_a [r(s,a) + \beta \int V^*(s')\, q(ds'|s,a)].$$

This equality is known as the Bellman equation or the **optimality equation**.

Let $\epsilon > 0$. For each $s \in S$, we can select $f(s) \in A$ so that

$$(T_f V^*)(s) \geq V^*(s) - \epsilon(1 - \beta).$$

Blackwell showed that the reward function $V(\pi)(\cdot)$ for the stationary plan $\pi = f^\infty$ satisfies:

$$V(\pi)(s) \geq V^*(s) - \epsilon, \qquad s \in S.$$

So good stationary plans exist.

Measurable Dynamic Programming

The first formulation of dynamic programming in a general measure theoretic setting was given by Blackwell (1965). He assumed:

1. S and A are Borel subsets of some nice measurable space (say, a Euclidean space).

2. The reward function $r(s, a)$ is Borel measurable.

3. The law of motion $q(\cdot|s, a)$ is a regular conditional distribution.

Plans are required to select actions in a Borel measurable way.

Measurability Problems

In his 1965 paper, Blackwell showed by example that for a Borel measurable dynamic programming problem:

The optimal reward function $V^*(\cdot)$ need not be Borel measurable and good Borel measurable plans need not exist.

This led to work by a number of mathematicians including R. Strauch, D. Freedman, M. Orkin, D. Bertsekas, S. Shreve, and Blackwell himself. It follows from their work that for a Borel problem:

The optimal reward function $V^*(\cdot)$ is universally measurable and that there do exist good universally measurable plans.

Blackwell's (1965) Example

Let $S = A = [0,1]$. The state of the system remains fixed: $q(s|s,a) = 1$ for all s, a. The reward function is

$$r(s,a) = 1_B(s,a)$$

where B is a Borel subset of $S \times A$ such that the projection

$$E = \{s : (\exists a)(s,a) \in B\}$$

is not Borel. The optimal reward at s is

$$V^*(s) = \begin{cases} 1 + \beta + \beta^2 + \cdots = 1/(1-\beta), \ s \in E \\ 0, \ s \notin E \end{cases}$$

The optimal reward is not Borel measurable and there are no good Borel measurable plans.

Positive Dynamic Programming

Assume: $\beta = 1$, $r(s,a) \geq 0$ for all (s,a), and the optimal reward function $V^*(s) < \infty$ for all s.

Theorem 3 *(Blackwell 1967). For* $0 < \epsilon < 1$ *and* P *a probability measure on* S *such that* $\int V^* \, dP < \infty$*, there exists a a stationary plan* π *such that*

$$P\{s : V(\pi)(s) \geq V^*(s) - \epsilon\} > 1 - \epsilon.$$

Blackwell showed by example that there need not exist a stationary π such that $V(\pi)(s) \geq V^*(s) - \epsilon$ for all s.

Theorem 4 *(Ornstein 1969, Frid 1972) Given $0 < \epsilon < 1$ and a probability measure P on S, there exists a stationary plan π with payoff $V(\pi)(s)$ at s such that*

$$P\{s \,|\, V(\pi)(s) \geq (1-\epsilon)V^*(s)\} = 1.$$

Question: Is there a stationary plan π such that

$$V(\pi)(s) \geq (1-\epsilon)V^*(s) \text{ for } \textbf{all } s?$$

Answer: Not in general. (Blackwell and Ramakrishnan (1988))

Negative Dynamic Programming

Assume: $\beta = 1, r(s, a) \leq 0$ for all (s, a).

A simple example of Dubins and Savage (1965) shows there need not exist good stationary plans even when S has only three elements and A is countable.

The fundamental paper is by Strauch (1966), based on his PhD thesis under Blackwell. There do exist optimal stationary plans if A is finite.

Question: Optimal Plan $\Rightarrow$ Stationary Optimal Plan?

Yes for discounted or negative problems. If π is optimal, then so is f^∞ where $f(s)$ is the first action for π when the initial state is s.

Theorem 5 *(Ornstein 1969, Blackwell 1970, Orkin 1974) If there is an optimal plan for a positive problem, then, for each probability P on S, there exists a stationary plan π which is optimal with P - probability one.*

Open question: Can the set of probability zero be eliminated?

Convergent Dynamic Programming

Assume: $\beta = 1$ and that

$$\sup_\pi E_{\pi,s}[\sum_{n=1}^{\infty} r^+(s_n, a_n)] < \infty$$

for all $s \in S$.

Many results, such as the Bellman equation, still hold in this general setting (Feinberg, 2002). For A compact, Schal (1983) proved that good stationary strategies exist.

Applications

Blackwell's fundamental work on dynamic programming led to applications in many areas including statistics,finance, economics, communication networks, water resources management, and even mathematics itself.

For information about applications and recent developments, see the **Handbook of Markov Decision Processes** (2002) edited by E. Feinberg and A. Shwartz.

Blackwell's papers on dynamic programming

On the functional equation of dynamic programming (1961). *J. Math. Anal. & Appl.* 2 273-276.

Discrete dynamic programming (1962). *Ann. Math. Statist.* 33 719-726.

Probability bounds via dynamic programming (1964). *AMS Proc. Symp. Appl. Math.* v. XVI 277-280.

Memoryless strategies in finite-stage dynamic programming (1964). *Ann. Math. Statist.* 35 863-865.

Discounted dynamic programming (1965). *Ann. Math. Statist.* 36 226-235.

Positive dynamic programming (1967). *Proc. 5th Berkeley Symp.* 415-418.

On stationary policies (1970). *J. Royal Stat. Soc, A* 133 33-37.

The optimal reward operator in dynamic programming (1974). *Ann. Prob.* 2 926-941 (with D. Freedman and M. Orkin).

The stochastic processes of Borel gambling and dynamic programming (1976). *Ann. Statist.* 4 370-374.

Stationary plans need not be uniformly adequate for leavable, Borel gambling problems (1988). *Proc. AMS* 102 1024-1027 (with S. Ramakrishnan).