Z. Wahrscheinlichkeitstheorie verw. Gebiete
48, 339–346 (1979)

Zeitschrift für
Wahrscheinlichkeitstheorie
und verwandte Gebiete

Solving a General Discounted Dynamic Program by Linear Programming

Wolf-Rüdiger Heilmann

University of Hamburg, Institut für Mathematische Stochastik,
Bundesstr. 55, D-2000 Hamburg 13, Federal Republic of Germany

Summary. As is well known (cf. Derman (1970) and references cited there), dynamic programming problems with finite state and action spaces can be solved by linear programming techniques. In the present paper it will be shown that this statement can be generalized to the case of general state and action spaces.

By this approach, the underlying sequential structure is completely neglected. Instead, vectore space structures and linear programming results (such as existence and duality theorems, complementary slackness) are used to obtain optimality statements.

1. The Dynamic Programming Model

Let $((S, \mathfrak{S}), (A, \mathfrak{A}), D, p, q, r, \beta)$ be a (*discounted stationary Markovian*) *decision model* in the sense of Hinderer [4], i.e. $(S, \mathfrak{S})$ is the *state space*, $(A, \mathfrak{A})$ is the *action space*, for all $s \in S$ the s-section D_s of $D \in \mathfrak{S} \otimes \mathfrak{A}$ denotes the set of admissible actions if the state s has occurred, p is the *initial distribution* on $\mathfrak{S}$, q a transition probability from D to S, the so-called *transition law*, $r\colon D \to \mathbb{R}$ the bounded, measurable *reward function*, $\beta \in (0, 1)$ the *discount factor*.

Throughout, we shall assume that $(S, \mathfrak{S})$ and $(A, \mathfrak{A})$ are SB-spaces, i.e. S and A are elements of σ-algebras generated by complete, separable and metrizable topologies and $\mathfrak{S}$ and $\mathfrak{A}$ are the traces of those σ-algebras in S and A. Furthermore, let D contain the graph of a measurable map from S to A.

A (*randomized stationary Markov*) *policy* π is a transition probability from S to A with $\pi(s, D_s) = 1$ for all $s \in S$. By a theorem of Ionescu-Tulcea, a policy π together with p and q defines a probability measure $P_\pi = p \otimes \pi \otimes q \otimes \pi \otimes \ldots \text{on} \bigotimes_1^\infty (\mathfrak{S} \otimes \mathfrak{A})$ and, by this, a stochastic process $((\zeta_n, \alpha_n), n \in \mathbb{N})$, ζ_n^1 (or α_n) being the projection into the n-th state (or action) space. A policy π^* is *$\bar{p}$-optimal* if π^* maximizes the functional

$$\pi \to V_\pi = E_\pi \left[\sum_{n=1}^{\infty} \beta^{n-1} r \circ (\zeta_n, \alpha_n) \right].$$

0044-3719/79/0048/0339/$01.60

Notation is as in [4], e.g. $\mathbb{N}$ is the set of positive integers, $\mathbb{R}$ is the set of real numbers, B_x is the x-section of the set B, $P_{Y|X}$ is the conditional distribution of Y under X, pr_n is the projection into the n-th coordinate of a cartesian product space.

2. The Linear Programming Model

General linear programs in a symmetric form have been studied by e.g. Krabs [6]. A slight modification of his model leads to the following pair of dual programs: Let E be a normed vector space, F a partially ordered normed vector space with positive cone F_+, $L: E \to F$ a continuous linear mapping with conjugate L^*: $F^* \to E^*$, $c \in E^*$, $b \in F$.

Primal Program (P):

For $x \in E$, $c(x) = \text{Min}!$, $Lx \geqq b$.

Dual Program (D):

For $y^* \in F^*$, $y^*(b) = \text{Max}!$, $L^* y^* = c$, $y^* \geqq 0$.

The results derived by Krabs can easily be transferred to this case. Let $M = \{x \in E: Lx \geqq b\}$, $N = \{y^* \in F^*: L^* y^* = c, y^* \geqq 0\}$, $\sigma = \inf_{x \in M} c(x)$, $\tau = \sup_{y^* \in N} y^*(b)$. σ (or τ) is the *value* of (P) (or (D)). Elements $x \in M$, $y^* \in N$ are *feasible*. If $x \in M$ and $c(x) = \sigma$ (or $y^* \in N$ and $y^*(b) = \tau$), x (or y^*) are *optimal*. (P) and (D) are called *feasible* (or *solvable*), if there are feasible (or optimal) elements x and y^*, respectively.

Theorem 1 ("*weak duality theorem*", cf. [6], *Satz* 3.2).

(i) $\tau \leqq \sigma$.
(ii) *If x and y^* are feasible, and $c(x) = y^*(b)$, then x and y^* are optimal, and $\tau = \sigma$.*

Theorem 2 ("*complementary slackness*", cf. [6], *Satz* 3.3). *Let $x \in M$, $y^* \in N$. Then the following affirmations are equivalent:*

(i) *x and y^* are optimal, and $\tau = \sigma$.*
(ii) $y^*(Lx - b) = 0$.

Theorem 3 ("*existence theorem*" cf. [6], p. 45 *and Satz* 4.14).

a) *If the convex cone*

$$K(L, c) = \{(c(x) + \gamma, Lx - y): x \in E, \gamma \geqq 0, y \in F_+\}$$

is closed (in the topology induced by the usual norm on $\mathbb{R} \times F$), then (P) *is feasible and its value σ is finite if and only if* (D) *is feasible and its value τ is finite. In both cases,* (P) *is solvable and*

$$-\infty < \tau = \sigma < \infty.$$

b) *If there is an $x \in E$ such that $Lx - b$ belongs to the interior of F_+, and if the value σ of* (P) *is finite, then* (D) *is solvable, and $\tau = \sigma$.*

3. The Linear Programming Formulation of the Dynamic Program

Now (and for the rest of the paper) let E (or F) be the space of measurable bounded mappings $v: S \to \mathbb{R}$ (or $w: D \to \mathbb{R}$). Then there is an isometric isomorphism between F^* and the space of bounded additive set functions on $\mathfrak{D} = D \cap \mathfrak{S} \otimes \mathfrak{A}$ (cf. [3], p. 258), and the following pair of linear programs in the above sense can be established:

(P) $\int v\, dp = \text{Min}!$

$$v(s) - \beta \int q(s, a, dt)\, v(t) \geqq r(s, a), \qquad (s, a) \in D.$$

(D) $\int r\, dv = \text{Max}!$

$$\int (u(s) - \beta \int q(s, a, dt)\, u(t))\, v(d(s, a)) = \int u\, dp, \qquad u \in E,$$

$$\int w\, dv \geqq 0, \qquad w \in F_+.$$

Immediately, we get the following results:

Lemma 4. *If v is feasible for* (D), then the following holds:

(i) $v \geqq 0$,
(ii) $v(\mathrm{D}) = 1/(1-\beta)$.

Proof. (i) Putting $w = 1_B$ for $B \in \mathfrak{D}$, we get

$$v(B) = \int 1_B\, dv \geqq 0.$$

(ii) For $u = 1_S$ we have

$$1 = \int u\, dp = \int (u(s) - \beta \int q(s, a, dt)\, u(t))\, v(d(s, a))$$
$$= \int (1-\beta)\, v(d(s, a)) = (1-\beta)\, v(D).$$

Lemma 5. (P) *is feasible, and*

$$\sigma \leqq \sigma' = \frac{1}{1-\beta} \sup_{(s,a) \in D} r(s, a) < \infty.$$

Proof. Let $v(s) = \sigma'$ for all $s \in S$. Then v is feasible for (P), and $\sigma \leqq \int v\, dp = \sigma'$.

A first connection between the dynamic programming problem and the dual program (D) is given by the following

Theorem 6. *For any policy π, v_π, defined by*

$$v_\pi(B) = \sum_{n=1}^{\infty} \beta^{n-1} (P_\pi)_{(\zeta_n, \alpha_n)}(B), \qquad B \in \mathfrak{D},$$

is feasible for (D).

Proof. v_π is a bounded measure on $\mathfrak{D}$, and for $u \in E$ we have

$$\int (u(s) - \beta \int q(s, a, dt)\, u(t))\, v_\pi(d(s, a))$$

$$= \sum_{n=1}^{\infty} \beta^{n-1} \int (u(s) - \beta \int q(s,a,dt)\, u(t))(P_\pi)_{(\zeta_n,\alpha_n)}(d(s,a))$$

$$= \sum_{n=1}^{\infty} \beta^{n-1} (\int u(s)\,(P_\pi)_{(\zeta_n,\alpha_n)}(d(s,a))$$

$$- \beta \int (P_\pi)_{(\zeta_n,\alpha_n)}(d(s,a)) \int q(s,a,dt)\; u(t))$$

$$= \sum_{n=1}^{\infty} \beta^{n-1} (\int u\, d(P_\pi)_{\zeta_n} - \beta \int u\, d((P_\pi)_{(\zeta_n,\alpha_n)} \otimes q)_{\mathrm{pr}_3})$$

$$= \sum_{n=1}^{\infty} \beta^{n-1} (\int u\, d(P_\pi)_{\zeta_n} - \beta \int u\, d(P_\pi)_{\zeta_{n+1}})$$

$$= \sum_{n=1}^{\infty} \beta^{n-1} \int u\, d(P_\pi)_{\zeta_n} - \sum_{n=1}^{\infty} \beta^n \int u\, d(P_\pi)_{\zeta_{n+1}}$$

$$= \int u\, dp.$$

The following result will be needed:

Lemma 7 ([4], Corollary 12.7). *Let $(X, \mathfrak{F})$ and $(Y, \mathfrak{G})$ be SB-spaces, $B \in \mathfrak{F} \otimes \mathfrak{G}$, and let ν be a probability measure on $\mathfrak{F} \otimes \mathfrak{G}$ concentrated on B. If there is a transition probability ψ from $(X, \mathfrak{F})$ to $(Y, \mathfrak{G})$ with $\psi(x, B_x) > 0$ for all $x \in \mathrm{pr}_1(B)$, then there exists a conditional distribution $Q \in \nu_{\mathrm{pr}_2|\mathrm{pr}_1}$ with*

$$Q(x, B_x) = 1 \quad \text{for all } x \in \mathrm{pr}_1(B).$$

From this, we immediately get

Corollary 8. *For any probability measure ν on $\mathfrak{D}$, (a version of) $\nu_{\mathrm{pr}_2|\mathrm{pr}_1}$ is a policy.*

Now let the measure ν be feasible for (D). Then $\tilde{\nu} = (1-\beta)\,\nu$ is a probability measure on $\mathfrak{D}$ (Lemma 4), there is a version $\pi \in \tilde{\nu}_{\mathrm{pr}_2|\mathrm{pr}_1}$ which is a policy (Corollary 8), and ν_π is feasible for (D). Furthermore, we have

Lemma 9. *Let the measure ν be feasible for* (D). *Then $\nu = \nu_\pi$.*

Proof. For any $B \in \mathfrak{D}$,

$$\nu_\pi(B) = \sum_{n=1}^{\infty} \beta^{n-1} (P_\pi)_{(\zeta_n,\alpha_n)}(B)$$

$$= \sum_{n=1}^{\infty} \beta^{n-1} \int p(ds_1) \int \pi(s_1, da_1) \int q(s_1, a_1, ds_2)$$

$$\ldots \int q(s_{n-1}, a_{n-1}, ds_n)\, \pi(s_n, B_{s_n}).$$

For $u_n(s) = \int \pi(s, da_1) \ldots \int q(s_{n-1}, a_{n-1}, ds_n)\, \pi(s_n, B_{s_n})$, $n \in \mathbb{N}$, $s \in S$, we have $u_n \in E$ and

$$\nu_\pi(B) = \sum_{n=1}^{\infty} \beta^{n-1} \int u_n\, dp$$

$$= \sum_{n=1}^{\infty} \beta^{n-1} \int (u_n(s) - \beta \int q(s,a,dt)\, u_n(t))\, \nu(d(s,a))$$

$$= \frac{1}{1-\beta} \sum_{n=1}^{\infty} \beta^{n-1} \left(\int u_n \, d\tilde{\nu}_{\mathrm{pr}_1} - \beta \int \tilde{\nu}(d(s,a)) \int q(s,a,dt)\, u_n(t) \right)$$

$$= \frac{1}{1-\beta} \sum_{n=1}^{\infty} \beta^{n-1} \Big(\int \tilde{\nu}_{\mathrm{pr}_1}(ds_1) \int \pi(s_1, da_1)$$

$$\ldots \int q(s_{n-1}, a_{n-1}, ds_n)\, \pi(s_n, B_{s_n})$$

$$- \beta \int \tilde{\nu}_{\mathrm{pr}_1}(ds) \int \tilde{\nu}_{\mathrm{pr}_2|\mathrm{pr}_1}(s, da) \int q(s,a,dt) \int \pi(t, da_1)$$

$$\ldots \int q(s_{n-1}, a_{n-1}, ds_n)\, \pi(s_n, B_{s_n}) \Big)$$

$$= \frac{1}{1-\beta} \int \tilde{\nu}_{\mathrm{pr}_1}(ds_1)\, \tilde{\nu}_{\mathrm{pr}_2|\mathrm{pr}_1}(s_1, B_{s_1}) = \frac{1}{1-\beta} \tilde{\nu}(B) = \nu(B).$$

Theorem 10. (D) *is feasible, and*

$$-\infty < \frac{1}{1-\beta} \inf_{(s,a)\in D} r(s,a) \leqq \tau \left(\leqq \sigma \leqq \frac{1}{1-\beta} \sup_{(s,a)\in D} r(s,a) < \infty \right).$$

Proof. According to Lemma 6, (D) is feasible, and because of Lemma 4, for any feasible ν we have

$$\tau \geqq \int r \, d\nu \geqq \int \inf_{(s,a)\in D} r(s,a)\, d\nu = \frac{1}{1-\beta} \inf_{(s,a)\in D} r(s,a) > -\infty.$$

The rest follows directly from Lemma 5 and Theorem 1.

A direct consequence of complementary slackness in linear programming is

Theorem 11. *Let v be feasible for* (P), *and let ν be feasible for* (D) *and σ-additive. Then the following statements are equivalent:*

(i) *v and ν are optimal, and $\tau = \sigma$.*

(ii) *$v(s) = r(s,a) + \beta \int q(s,a,dt)\, v(t)$ for ν-almost all $(s,a) \in D$.*

Proof. Because of Theorem 2, (i) and

(ii′) $\int (v(s) - \beta \int q(s,a,dt)\, v(t) - r(s,a))\, \nu(d(s,a)) = 0$

are equivalent. But the integrand is nonnegative, and thus (ii) and (ii′) are equivalent, too.

4. Existence Theorems

For the application of part a) of Theorem 3, we require the following

Lemma 12. *$K(L,c)$ is closed.*

Proof. The proof is merely technical, hence we just give an indication: For

$$\tilde{F} = \{ w = Lu - \rho \in F : u \in E, \rho \in F_+ \},$$

put

$$f(w) = \inf \{ \int u \, dp : Lu - w \in F_+ \}, \qquad w \in \tilde{F}, \tag{4.1}$$

and show that $\tilde{F} = F$, the inf in (4.1) is assumed, and f is lower semicontinuous.

Now from Theorem 3, Theorem 10, and Lemma 12 we get

Theorem 13. (P) *is solvable,* (D) *is feasible, and* $\tau=\sigma$.

Trivially, for the interior $\mathring{F}_+$ of F_+ we have

$$\mathring{F}_+=\{w\in F\colon \inf w>0\}.$$

Now let $\varepsilon>0$, $\delta=\sup_{(s,a)\in D} r(s,a)+\varepsilon$, and $v(s)=\delta/(1-\beta)$ for all $s\in S$. Then $v\in E$, and

$$\begin{aligned}(Lv-r)(s,a)&=v(s)-\beta\int q(s,a,dt)\,v(t)-r(s,a)\\&=\delta/(1-\beta)-\beta\,\delta/(1-\beta)-r(s,a)\\&=\sup_{(s,a)\in D} r(s,a)-r(s,a)+\varepsilon\geqq\varepsilon>0\qquad\text{for all } (s,a)\in D,\end{aligned}$$

i.e. $Lv-r\in\mathring{F}_+$. Hence from part b) of Theorem 3 and Theorem 10, we easily derive

Theorem 14. (D) *is solvable.*

Remark. Our way of proof for the solvability of (D) rests heavily on the special structure of F, which was suggested by the characteristics of the underlying dynamic programming model. When generalized linear programming is used to construct optimal statistical procedures, it is reasonable to define F to be some L_1 space (cf. Krafft and Witting [7]), however. Indeed, the positive cone of L_1 generally has a void interior, and thus a separate proof for an analogue to Theorem 14 is required in this case.

5. The Modified Dual Program

By

$$\pi\sim\pi'\Leftrightarrow v_\pi=v_{\pi'},$$

an equivalence relation in the set of policies is defined, and according to Theorem 6 and Lemma 9, there is a one-to-one correspondence between the set of measures which are feasible for (D), and the set of equivalence classes (with respect to "$\sim$") of policies.

Now the following optimization problem (which is no longer a linear program in the above sense) suggests itself:

(D*) (For measures ν on $\mathfrak{D}$:)

$$\int r\,d\nu=\text{Max}!$$

$$\int\left(u(s)-\beta\int q(s,a,dt)\,u(t)\right)\nu(d(s,a))=\int u\,dp,\qquad u\in E.$$

Clearly, any ν feasible for (D*) is also feasible for (D). So for the value τ^* of (D*) we obtain (cf. Theorem 10)

$$-\infty<\frac{1}{1-\beta}\inf_{(s,a)\in D} r(s,a)\leqq\tau^*\leqq\tau.$$

The following results justify the introduction of (D*) in connection with the dynamic programming problem.

Theorem 15. (D*) *and the problem of finding a $\bar{p}$-optimal policy are equivalent in the following sense:*

a) If π is a policy, then v_π is feasible for (D*), *and* $V_\pi = \int r\, dv_\pi$.

b) *If v is feasible for* (D*), $\tilde{v} = (1-\beta)\, v$, *then there is a policy* $\pi_v \in \tilde{v}_{\mathrm{pr}_2|\mathrm{pr}_1}$, *and* $V_{\pi_v} = \int r\, dv$.

c) $V = \sup_\pi V_\pi = \tau^*$.

Proof. The assertion follows immediately from Theorem 6 and Lemma 9.

Especially we have

Corollary 16. *A $\bar{p}$-optimal policy π^* is obtained by deriving a solution v^* of* (D*) *and choosing a policy* $\pi^* \in ((1-\beta)\, v^*)_{\mathrm{pr}_2|\mathrm{pr}_1}$.

Proof. For any policy π,

$$V_\pi = \int r\, dv_\pi \leqq \int r\, dv^* = V_{\pi^*}.$$

Another relation between the dynamic programming problem and the dual pair (P), (D) of linear programs is provided by the following

Theorem 17. *Let π be any policy, and let*

$$V_\pi(\cdot) = E_\pi\left[\sum_{n=1}^{\infty} \beta^{n-1}\, r\cdot(\zeta_n, \alpha_n) \,|\, \zeta_1 = \cdot\right]$$

be feasible for (P) (*note that* $V_\pi = \int V_\pi(\cdot)\, dp$).

Then $V_\pi(\cdot)$ is optimal for (P), *π is $\bar{p}$-optimal, and v_π is optimal for* (D).

Proof. For all $v \in N$,

$$\int r\, dv \leqq \int V_\pi(\cdot)\, dp = \int r\, dv_\pi \leqq \inf_{v\in M} \int v\, dp.$$

In view of Theorem 14 and Corollary 16, it is worth while to provide conditions which guarantee the "completeness" of the class of σ-additive set functions on $\mathfrak{D}$ with respect to the optimization of the dual program (D). The following well known example shows that even in very simple cases we have to consider set functions which are not σ-additive:

Example ([1], p. 229). Put $S = \{0\}$, $A = \mathbb{N}$, $D = S \times A$, and

$$r(0, n) = (n-1)/n, \qquad n \in \mathbb{N}.$$

Then we have

(P) (For $v \in \mathbb{R}$:)

$$v = \mathrm{Min}!, \qquad v - \beta v \geqq (n-1)/n, \qquad n \in \mathbb{N},$$

and trivially $v = 1/(1-\beta)$ is optimal.

Furthermore,

(D) (For contents v on $\mathbb{N}$:)

$$\int((n-1)/n)\,v(dn)=\mathrm{Max}\,!,\qquad v(\mathbb{N})\cdot(1-\beta)\cdot u=u,\qquad u\in\mathbb{R}.$$

Clearly, an optimal measure v on $\mathbb{N}$ does not exist. From Theorem 14 we know that (D) is solvable, so a content $\tilde{v}$ on $\mathbb{N}$ with $\tilde{v}(B)=0$ for all finite $B\subset\mathbb{N}$ and $\tilde{v}(\mathbb{N})=1$ must exist – a result which has been found by Horn and Tarski (1948), e.g., in quite a different connection.

Acknowledgment. The present paper is an excerpt from the author's doctoral dissertation which was written at the University of Hamburg under the guidance of K. Hinderer, to whom the author wishes to express his gratitude for his advice and encouragement.

References

1. Blackwell, D.: Discounted dynamic programming. Ann. Math. Statist. **36**, 226–235 (1965)
2. Derman, C.: Finite State Markovian Decision Processes. New York: Academic Press 1970
3. Dunford, N., Schwartz, J.T.: Linear Operators, Part I. New York: Interscience Publishers 1957
4. Hinderer, K.: Foundations of Non-stationary Dynamic Programming with Discrete Time Parameter. Berlin-Heidelberg-New York: Springer 1970
5. Horn, A., Tarski, A.: Measures in Boolean Algebras. Trans. Amer. Math. Soc. **64**, 467–497 (1948)
6. Krabs, W.: Optimierung und Approximation. Stuttgart: Teubner 1975
7. Krafft, O., Witting, H.: Optimale Tests und ungünstigste Verteilungen. Z. Wahrscheinlichkeitstheorie und verw. Gebiete **7**, 289–302 (1967)

Received April 19, 1977