ELSEVIER

European Journal of Operational Research 80 (1995) 147–155

EUROPEAN
JOURNAL
OF OPERATIONAL
RESEARCH

Theory and Methodology

A submodular approach to discrete dynamic programming

C.M. Klein *

Department of Industrial Engineering, University of Missouri - Columbia, Columbia, MO 65211, USA

Received April 1993; revised August 1993

Abstract

Submodular functions are playing an increasing role in analyzing many discrete optimization problems. The purpose of this paper is to continue the trend by using submodular functions and their properties to develop a duality for discrete dynamic programming.

Keywords: Submodular functions; Polymatroids; Dynamic programming; Duality

1. Introduction

Duality is one of the most elegant concepts and one of the important computational tools in mathematics. Duality's contributions to analysis and optimization are many. However, as a result of the recursive nature of dynamic programming (DP), duality results for DP are quite limited.

Bellman [1,2] was the first to introduce a duality concept in dynamic programming. Bellman [1] used Lagrange multipliers to reduce the state space. This in itself does not result in a general duality theory for DP. However, this concept, which was later generalized by Everett [4], has played an increasingly important role in discrete optimization vis-à-vis Lagrangian relaxation.

Dinkle and Peterson [3] developed a duality theory for the class of dynamic programming problems that have convex return functions and linear transition functions through the use of generalized geometric programming duality. In their approach, Dinkle and Peterson viewed the primal and dual problems in terms of orthogonal spaces. This approach is not without possible drawbacks. As Rockafellar [9,10] noted, it may not be a very effective scheme since it doesn't always lead to unambiguous Lagrangians. It also does not provide the most natural setting for the study of the optimal value functions and it could create a conceptual stumbling block in application if there is not an obvious subspace at hand. Klein and Morin [6,7] overcame this problem with a more general approach using conjugate functions. Aside from the previously mentioned results and the fact that numerous other authors have

* This research was supported in part by a grant from the Office of Naval Research under contract no. N00014-88-K-0478.

SSDI 0377-2217(93)E0329-V

provided tantalizing hints, a general duality theory for dynamic programming does not appear to have been developed to date.

We will look at the shortest route problem on a directed acyclic network which is the prototype problem for Dynamic Programming (DP). In this paper we will show how a duality can be achieved for the DP formulation of the shortest route problem.

2. Preliminaries

Let $G(V, E)$ be a directed acyclic network. Then the vertices of G can be numbered so that 1 is the source and N is the sink, where $N = |V|$, and for any $(i, j) \in E$, $i < j$. Without loss of generality let $V = \{1, 2, 3, \dots, N\}$. For any edge $(i, j) \in E$, there is a corresponding weight c_{ij} with $c_{ij} = +\infty$ if $(i, j) \notin$ E.

To find the path with minimum weight one can use the following DP recursion:

$$f(i) = \min_{i<j} \{c_{ij} + f(j)\}. \tag{1}$$

Here $f(i)$ represents the shortest distance from i to the sink N. The shortest path for the graph G is given by $f(1)$ and $f(N) := 0$.

In order to develop a DP duality for this formulation we will use the concepts of submodularity and submodular polyhedra [8]. Recall that a set function f defined on all the subsets of a finite set S, i.e., $f: 2^S \to \mathbb{R}$ where 2^S represents the set of all subsets of S, is submodular if

$$f(A \cup B) + f(A \cap B) \leq f(A) + f(B), \tag{2}$$

for all A, $B \subseteq S$. Also, we can define a submodular polyhedron associated with a nonnegative submodular function f as follows.

Definition 1. Let $f: 2^S \to \mathbb{R}$, and let f be submodular. The *submodular polyhedron* associated with f is $P_f = \{x \in \mathbb{R}^S \mid x(T) \leq$ for all $T \subseteq S\}$.

Note that if we restrict $x(T) \geq 0$ then the polyhedron is bounded and is referred to as a polymatroid. The property of interest in terms of submodular polyhedra is that the intersection of two submodular polyhedra, P_f and P_g, is described by the system

$$x(T) \leq \min\{f(T), g(T)\} \quad \text{for all } T.$$

Based on this system description a duality for the intersection of two submodular polyhedra can be derived. This duality is given below.

Let f and g be submodular functions defined on S, $f(\emptyset) = g(\emptyset) = 0$. Then the following holds:

$$\max\{1 \cdot x \mid x \in P_f \cap P_g\} = \min\{f(A) + g(S - A) \mid A \subseteq S\}. \tag{3}$$

The right-hand side of (3) is the convolution operator for set functions. The convolution function h is defined as

$$h(A) = \min_{B \subseteq A} \{g(B - A) + f(A)\} = (g \,\square\, f)(A) \tag{4}$$

In general, h will not be a submodular function. However, if f is submodular and g is modular in (4), then h is submodular.

3. DP duality for the shortest route problem

In order to develop a DP duality for the shortest route problem we will use the functional equation given by (1) and the duality of submodular polyhedra intersection given in (3). First we must rewrite (1) in terms of set functions. To do this we define the following sets and set functions. For $G(V, E)$ directed and acyclic, let

$$J_i = \{i, i+1, \dots, N\},\ i \in V = \{1, 2, \dots, N\}.$$

$$J^0 = \emptyset.$$

$$J^i = \{1, \dots, i\},\ i \in V = \{1, 2, \dots, N\}.$$

Define the function $f: 2^V \to \mathbb{R}$ to be the following function:

$$f(A) = \begin{cases} 0 & \text{if } A = \emptyset, \\ \text{shortest distance from } i \text{ to } N & \text{if } A = J_i, \\ +\infty & \text{otherwise.} \end{cases}$$

Claim 1. *f is submodular.*

Proof. Let $A, B \subseteq V$. If A and B are such that $f(A) + f(B) < +\infty$ then $A = J_i$ for some i and $B = J_k$ for some k. By definition either $J_i \subseteq J_k$ or $J_k \subseteq J_i$. Without loss of generality let $J_i \subseteq J_k$. Then $A \cup B = J_k$, $A \cap B = J_i$ and $f(A) + f(B) = f(A \cup B) + f(A \cap B)$. If A and B are such that $A \neq J_i$ or $B \neq J_i$ for some i then $f(A) + f(B) = +\infty$ and $f(A) + f(B) \geq f(A \cup B) + f(A \cap B)$ holds. □

Note that here $f(A)$ is actually a modular function for all $A \subseteq V$ such that $f(A) < +\infty$.

Now, define the set-functions $g_i: 2^V \to R$, $i = 1, \dots, N-1$, as follows:

$$g_i(A) = \begin{cases} 0 & \text{if } A = \emptyset, \\ c_{i,k+1} & \text{if } A \cup J^{i-1} = J^k, \\ +\infty & \text{otherwise,} \end{cases}$$

$$g_N(\cdot) = 0.$$

Claim 2. *g_i, for $i = 1, \dots, N-1$, is submodular.*

Proof. Let $A, B \subseteq V$. If A and B are such that $g_i(A) + g_i(B) < +\infty$ then $A \cup J^{i-1} = J^k$ for some k and $B \cup J^{i-1} = J^p$ for some p. Without loss of generality assume $J^k \subseteq J^p$. Since $(A \cup J^{i-1}) = J^k$ and $(B \cup J^{i-1}) = J^p$ and $J^k \subseteq J^p$ it follows that $(A \cup B) \cup J^{i-1} = J^p$. Hence, $g_i(A \cup B) = c_{i,p+1}$. Likewise, $(A \cap B) \cup J^{i-1} = J^k$ which implies $g_i(A \cap B) = c_{i,k+1}$. Therefore, $g_i(A) + g_i(B) = g_i(A \cup B) + g_i(A \cap B)$. If A and B are such that $g_i(A) + g_i(B) = +\infty$ then $g_i(A) + g_i(B) \geq g_i(A \cap B) + g_i(A \cup B)$. □

By definition, $f(J_i)$ is the shortest path from i to N. We can write $f(J_i)$ in DP recursion form as follows:

$$f(J_i) = \min_{A \subset J_i} \{g_i(J_i - A) + f(A)\}. \tag{5}$$

Note that the minimum is over $A \subset J_i$ since $A = J_i$ gives the identity $f(J_i) = f(J_i)$. Also note that the right-hand side of (5) is just the convolution of two submodular functions. However, in this case, the convolution is again a submodular function since it is the function f.

Claim 3. *$f(J_i) = \min_{A\subset J_i}\{g_i(J_i - A) + f(A)\}$ yields the shortest distance from i to N in the directed acyclic network $G(V, E)$.*

Proof. In order for $f(J_i) \neq +\infty$ the minimum will have to occur at some set J_{i+k}, $k = 1,\dots, N-i$, otherwise $f(A) = +\infty$. If the minimum occurs at $A = J_{i+k}$ for some $k \leq N-i$ then

$$J_i - J_{i+k} = \{i, i+1, \dots, i+k-1\} \quad \text{and} \quad (J_i - J_{i+k}) \cup J^{i-1} = \{1, 2, \dots, i+k-1\}.$$

Hence $g_i(J_i - A) = c_{i,i+k}$. If $(i, i+k)$ does not exist, $c_{i,i+k} = +\infty$ and the minimum will not be obtained. Therefore, $g_i(J_i - A)$ will yield only those values of the edges leaving vertex i. Hence, in the recursion given by (5) we have $A = J_{i+k}$, for some k, $g_i(J_i - A) = c_{i,i+k}$, $(i, i+k) \in E$ and $f(J_{i+k}) < +\infty$. Since $f(J_{i+k})$ is the shortest distance from $i+k$ to N and $g_i(J_i - J_{i+k}) = c_{i,i+k}$ it follows by the principle of optimality that the minimum given by

$$\min_{A\subset J_i}\{g_i(J_k - A) + f(A)\} = c_{i,i+k} + f(J_{i+k}),$$

for some k, will be the shortest distance from i to N. □

We now can state a DP duality based on this formulation. The shortest distance from the source, node 1, to the sink, node N, is given in set-function form by

$$f(V) = \min_{A\subset V}\{g_1(V - A) + f(A)\}$$

which is just a convolution operator between two submodular functions. Hence,

$$\min_{A\subset V}\{g_1(V-A) + f(A)\} = \max\{1\cdot x \mid x(T) \leq \min\{g_1(T), f(T)\}, T \subseteq V\}. \tag{6}$$

However, since the intersection of g_i and f, for $i = 1,\dots, N$, is a submodular function, the right-hand side, i.e., the dual, can be written as

$$\max\{1\cdot x \mid x(T) \leq \min\{g_1(T), g_2(T), \dots, g_N(T), f(T)\}, T \subseteq V\}, \tag{7}$$

as was shown by Lovasz [8].

The purpose in developing (7) is that the $g_i(T)$-values are based on cost coefficients that are already known. In order to use (6), it is necessary to know a priori the shortest lengths from each node to the sink. It can be shown that the constraints produced by the shortest lengths, i.e., $f(T)$, are redundant in (7).

4. Example

For the graph in Fig. 1 we have the following function evaluations. For those sets not given the function values are $+\infty$. $g_1(\{1\}) = 4$, $g_1(\{1, 2\}) = 3$; $g_2(\{2\}) = 4$, $g_2(\{2, 3\}) = 8$, $g_2(\{2, 3, 4\}) = 7$; $g_3(\{3\}) = 9$, $g_3(\{3, 4\}) = 10$; $g_4(\{4, 5\}) = 6$; $g_5(\{5\}) = 4$; $g_6(\cdot) = 0$; $f(\{6\}) = 0$, $f(\{5, 6\}) = 4$, $f(\{4, 5, 6\}) = 6$, $f(\{3, 4, 5, 6\}) = 14$, $f(\{2, 3, 4, 5, 6\}) = 11$, $f(\{1, 2, 3, 4, 5\}) = 15$.

We need to determine an $x(T) \leq \min\{g_1(T),\dots, g_6(T), f(T)\}$ where $x \in \mathbb{R}^V$, $x(T) = \sum_{e\in T} x(e)$. Here e represents a single element of T, $x(e)$ represents the component. That is, if $T = \{1, 4, 6\}$ then $\sum x(e) = x_1 + x_4 + x_6$. Also recall that V is the set of vertices. Therefore, the maximum possible values for $x = (x_1, x_2, x_3, x_4, x_5, x_6)$ would be $x_{\max} = (4, 4, 9, \infty, 4, 0)$, based on the sets $\{1\}$, $\{2\},\dots,\{N\}$. However, we must also take into account the other sets and the restrictions they place. Taking these into account, we obtain $x = (4, -1, 8, 0, 4, 0)$ as the maximum of $1\cdot x$.

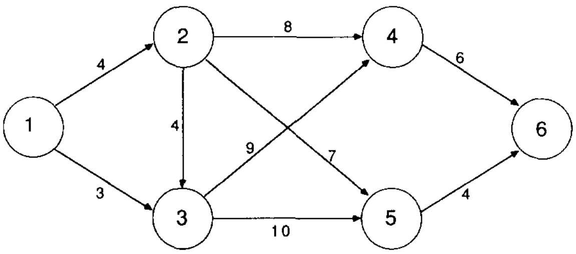

Fig. 1. A directed acyclic graph.

This value can be obtained by solving the corresponding LP of the problem. In this case the LP is given by

(DLP)

$$\max\ x_1+x_2+x_3+x_4+x_5+x_6$$

$$\begin{aligned}\text{subject to}\quad & x_1 \le 4,\\ & x_1+x_2 \le 3,\\ & x_2 \le 4,\\ & x_2+x_3 \le 8,\\ & x_2+x_3+x_4 \le 7,\\ & x_3 \le 9,\\ & x_3+x_4 \le 10,\\ & x_4+x_5 \le 6,\\ & x_5 \le 4,\\ & x_6 \le 0,\\ & x_i \text{ unrestricted.}\end{aligned}$$

Note that this LP does not make use of the restrictions imposed by the function evaluations of f since they are redundant.

5. Analysis of the dual problem

Notice that DLP has a unimodular structure and does not need to be restricted to integer values. This is anticipated since the theory of submodular polyhedron tells us that the intersection of two integral submodular polyhedra will have integral vertices. However, due to the construction of the convolution, what we have is the intersection of two submodular polyhedron being guaranteed to be a submodular polyhedron again. Therefore, we can maximize $1 \cdot x$ over the intersection of all the corresponding submodular polyhedron, i.e.,

$$\left(P_{g_1} \cap P_{g_2} \cap \cdots \cap P_{g_N} \cap P_f\right)$$

and still be guaranteed integral vertices.

To obtain a true DP duality we must take this DLP a step further. The general form of the DLP is

(DLP1)

$$\begin{array}{ll} \max & 1 \cdot x \\ \text{s.t.} & Ax \le c \\ & x \text{ unrestricted.} \end{array}$$

If we restrict x, i.e., set $x = u - w$, $u \ge 0$, $w \ge 0$, we obtain

(DLP2)

$$\begin{array}{ll} \max & 1 \cdot (u - w) \\ \text{s.t.} & Au - Aw \le c \\ & u \ge 0, \quad w \ge 0. \end{array}$$

For ease of presentation let $z = (u_1, \ldots, u_n, w_1, \ldots, w_n)$, $B = (A, -A)$, and $e = (1, \ldots, 1, -1, \ldots, -1)$. Then the problem is

(DLP3)

$$\begin{array}{ll} \max & e \cdot z \\ \text{s.t.} & Bz \le c \\ & z \ge 0 \end{array}$$

The integrality of z is implicitly taken care of through the unimodularity of A. But if we explicitly state that z must be integral, then DLP3 is a multi-dimensional knapsack type problem and we have the DP formulation for DLP3 given below.

$$\max f_i(y) = \max_{z_i \in Z_i(y)} \{e_i z_i + f_{i-1}(y - b_i z_i)\}$$

where b_i is the i-th column of B, B is $m \times 2n$, y is an m-vector where $y \le c$ and $Z_i(y) = \{0, 1, \ldots, \min_j([y_j / b_{ji}])\}$. There is also the boundary condition $f_0(y) = 0$.

Therefore, in terms of DP duality we have the dual of the DP recursion of the shortest path problem being the DP recursion of a knapsack problem. This is intuitively appealing since we already know that a 1-dimensional knapsack problem can be formulated and solved as a longest path problem. This duality is further emphasized by these results.

Another interesting relationship of the dual problems can be found by looking at the LP's involved. As Wagner [11] noted, the DP recursion for the shortest path problem actually solves the dual of the LP formulation of the shortest path problem.

The shortest path problem for a directed acyclic graph can be formulated as

(SRP)

$$\begin{array}{ll} \min & \sum \sum c_{ij} x_{ij} \\ \text{s.t.} & \sum_j x_{kj} - \sum_i x_{ik} = \begin{cases} 1 & \text{if } k \text{ is source,} \\ 0 & \text{other,} \\ -1 & \text{if } k \text{ is sink,} \end{cases} \\ & x_{ij} \ge 0. \end{array}$$

If we look at the example problem and DLP we see that this formulation is in the same basic form as the dual to SRP. The right hand side of DLP is given by the c_{ij}'s and there is one constraint for each edge in the graph. This is similar to the maximum packing of (s, t) cuts given by Fulkerson [5]. That is, there is one constraint for each edge, but this formulation has only one variable for each node and the variables are unrestricted. If we formulate the max packing of (s, t) cuts problem for this example, we obtain the formulation MPLP given below. Note that there is one variable for each cut.

(MPLP)

$$\begin{array}{ll}
\max & \sum_{i=1}^{|K|} x_i \\
\text{s.t.} & x_1+x_2+x_3+x_4 \le 4, \\
& x_1+x_9+x_{10}+x_{11} \le 3, \\
& x_9+x_{10}+x_{11} \le 4, \\
& x_2+x_3+x_6+x_8 \le 9, \\
& x_2+x_4+x_6+x_7 \le 10, \\
& x_3+x_5+x_8+x_{11} \le 4, \\
& x_4+x_5+x_7+x_{10} \le 6, \\
& x_6+x_7+x_9+x_{11} \le 7, \\
& x_6+x_8+x_9+x_{11} \le 8, \\
& x_i \ge 0,
\end{array}$$

where K is the set of all cuts separating s from t, for which there are 11 in this problem.

As stated, this is similar to DLP in terms of the right hand side, but differs in terms of the number of variables and their restriction. If we take the dual of DLP we obtain

(SRP1)

$$\begin{array}{ll}
\min & \sum \sum c_{ij} x_{ij} \\
\text{s.t.} & x_{12}+x_{13} = 1, \\
& x_{13}+x_{23}+x_{24}+x_{25} = 1, \\
& x_{24}+x_{25}+x_{34}+x_{35} = 1, \\
& x_{25}+x_{35}+x_{46} = 1, \\
& x_{46}+x_{56} = 1, \\
& x_{ij} \ge 0,
\end{array}$$

which is a different formulation of the shortest path problem. In this formulation, each constraint represents a particular cut in the network. Constraint i represents the cut that separates the sets of nodes $(1,\dots,i)$ and $(i+1,\dots,N)$. Note, however, that this particularly nice representation of the subset of cuts occurred only due to the manner in which the nodes were numbered. In general, this will not be the subset of cuts that constitutes the linear program.

In general, for the DP duality stated, we are solving the duals to two different formulations of the shortest path problem. This can be seen below. Assume that 1 is the source node and N is the sink. Then for the primal DP we have

(SRP)

$$\min \sum \sum c_{ij} x_{ij}$$

subject to

$$\sum_{(k,j)} x_{kj} - \sum_{(i,k)} x_{ik} = \begin{cases} 1 & \text{if } k \text{ is source,} \\ 0, & \text{other,} \\ -1 & \text{if } k \text{ is sink,} \end{cases}$$

$$x_{ij} \geq 0,$$

(DSRP)

$$\max \ -y_n + y_1$$

subject to

$$y_i - y_j \leq c_{ij},$$

y_k unrestricted,

and the DP recursion for shortest paths solves (DSRP).

For the Dual DP we have

(SRP1)

$$\min \sum \sum c_{ij} x_{ij}$$

subject to

$$Ax = 1,$$

$$x_{ij} \geq 0,$$

(DLP)

$$\max \sum_i y_i$$

subject to

$$A^{\mathrm{T}} y \leq C,$$

y unrestricted,

and the dual DP formulation solves DLP after y is restricted.

These LP's and their duals are interesting in that we see two different formulations that are close to each other, but yield different recursions to solve the duals. In addition, these duals, DSRP and DLP, are dual to each other in terms of the DP recursion.

6. Conclusions

It has been shown that it is possible to develop a DP duality through the shortest route problem. The dual then becomes a multidimensional knapsack problem. This has some intuitive feel since the one-dimensional knapsack problem can be reformulated as a longest path problem. Computationally, this dual does not have any advantage since shortest path problems are well solved. However, the primary result is the development of a duality that has not existed before for general DP. It may be possible though to gain a computational advantage when looking at DP's other than the shortest path problem and their duals. These problems and the relationships and implications of these results warrant further investigation.

References

[1] Bellman, R.E., "Dynamic programming and Lagrange multipliers", *Proceedings of the National Academy Sciences* 42 (1956) 767–769.

[2] Bellman, R.E., "On the maximum transform", *Journal of Mathematical Analysis and Applications* 6 (1963) 65–74.

[3] Dinkle, J., and Peterson, E.L., "A duality theory for dynamic programming problems via geometric programming", Working Paper, Dept. of I.E. and Management Science, Northwestern University, (1974).

[4] Everett, H.M., "Generalized Lagrange multiplied method for solving problems of optimal allocation of resources", *Operations Research* 11 (1963) 399–417.

[5] Fulkerson, D.R., "Networks, frames, blocking systems", in: G.B. Dantzig and A.F. Veinott (eds.), *Lectures in Applied Mathematics, Vol. 11: Mathematics of the Decision Sciences*, American Mathematical Society, Providence, RI, 1968, 303–335.

[6] Klein, C.M., and Morin, T.L., "Duality for dynamic programming", Working Paper Series no. 1210588, Department of Industrial Engineering, University of Missouri - Columbia, 1988.

[7] Klein, C.M., and Morin, T.L., "Conjugate duality and the curse of dimensionally", *European Journal of Operational Research* 50 (1991) 220–228.

[8] Lovasz, L., "Submodular functions and convexity", in: A. Bachem, M. Grotschel and B. Korte (eds.), *Mathematical Programming: The State of the Art*, Springer Verlag, New York, 1983, 235–257.

[9] Rockafellar, R.T., *Conjugate Duality and Optimization*, SIAM, Philadelphia, PA, 1974.

[10] Rockafellar, R.T., *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[11] Wagner, H.M., *Principles of Operations Research*, Prentice-Hall, Englewood Cliffs, NJ, 1975.