Information Relaxations and Dynamic Zero-Sum Games

Martin Haugh
Department of IE and OR, Columbia University, New York, NY 10027, mh2078@columbia.edu.

Chun Wang
Department of IE and OR, Columbia University, New York, NY 10027, cw2519@columbia.edu.

May 20, 2014

Abstract

We apply duality methods based on information relaxations to dynamic zero-sum games. We show these methods can easily be used to construct dual lower and upper bounds for the optimal value of these games. In particular, these bounds can be used to evaluate sub-optimal policies for zero-sum games when calculating the optimal policies and game value is intractable.

Keywords: Information relaxations, duality, zero-sum games.
1 Introduction

In this paper we apply recently-developed duality methods based on information relaxations for stochastic dynamic problems to dynamic two-person zero-sum games. These methods generalized earlier work on optimal stopping problems and were developed independently by Brown, Smith and Sun [4] (hereafter denoted by BSS) and Rogers [13]. They can be used to evaluate sub-optimal policies for dynamic programming (DP) problems where it is too difficult to solve the problems exactly. In particular, a primal bound can be obtained by simply simulating the sub-optimal policy and computing the average total reward (in the case of a maximization problem) across the simulated paths. A dual bound can be computed by solving a suitably defined optimization problem along each simulated path and again averaging the optimal objective functions of these dual problems across the simulated paths. If the dual gap is small, then we know the sub-optimal policy must be close to optimal. In a relatively short period of time these methods have become quiet standard and are now often applied when sub-optimal policies are proposed for hard-to-solve control problems. Recent applications and developments include Brown and Smith [3], Desai et al. [5], Lai et al. [9], Moallemi and Saglam [10] and Haugh and Wang [8].

Our extension of the information relaxation technology to dynamic zero-sum games is motivated in part by Beveridge and Joshi [1] who generalized the dual optimal stopping work of Haugh and Kogan [7] and Rogers [12] to zero-sum optimal stopping games. The main contribution of this paper is to demonstrate that the more general dual approach of BSS and Rogers [13] can also be easily applied to dynamic zero-sum games. While the results are easy to prove, there are many interesting applications including pursuit-evasion games, heads-up poker and many two-person computer games. Our ultimate goal is to apply these dual techniques to very complex zero-sum games which cannot be solved to optimality.

The remainder of this paper is organized as follows. We define the problem and establish the dual theory for finite horizon zero-sum games in Section 2. In Section 3 we outline how the approach can be extended to discounted infinite horizon problems as well as more general stochastic shortest path games using the very recent work of Brown and Haugh [2]. We conclude in Section 4.

2 Dual Bounds for Finite Horizon Zero-Sum Games

We begin with a finite horizon zero-sum game played by two players, A and B. We index time according to \( t = 0, 1, \ldots, T \), and the evolution of information is described by a filtration \( \mathcal{F} = \{ \mathcal{F}_0, \ldots, \mathcal{F}_T \} \) with \( \mathcal{F}_T = \mathcal{F} \). We make the simplifying assumption that both A and B have access to the same information \( \mathcal{F}_t \) at time \( t \) and by insisting that \( \mathcal{F}_t \subseteq \mathcal{F}_{t+1} \subseteq \mathcal{F} \) for all \( t < T \) we model the assumption that the two players do not forget previously known information. We assume \( \mathcal{F}_0 = \{ \emptyset, \Omega \} \) so player A and B know nothing initially. The state vector at time \( t \) is denoted by \( x_t \). We let \( A_t(x_t) \) and \( B_t(x_t) \) be the sets of time-\( t \) actions that are feasible for players A and B, respectively, given the current state is \( x_t \). We assume \( A_t(x_t) \) and \( B_t(x_t) \) are finite sets and that \( x_t \) evolves according to

\[
x_{t+1} = f_t(x_t, a_t, b_t, \omega_{t+1}), \quad t = 0, \ldots, T - 1
\]

where \( a_t \in A_t(x_t) \) and \( b_t \in B_t(x_t) \) are the actions of A and B, respectively, at time \( t \), and \( \omega_{t+1} \) is an \( \mathcal{F}_{t+1} \)-measurable random disturbance. After A and B implement their actions at time \( t \), B makes
a payment of $\tilde{g}_t(x_t, a_t, b_t)$ to $A$. The game ends at time $T$ when $B$ pays $g_T(x_T)$ to $A$. We assume that each $\tilde{g}_t$ is $\mathcal{F}_t$-measurable.

It is useful to make the distinction between pure and mixed policies. A pure policy chooses an action deterministically whereas a mixed policy chooses an action randomly. In particular at time $t$ a mixed policy for player $A$ is a probability distribution $u_t = \{u_t^{(a_t)} | a_t \in \mathcal{A}(x_t)\}$ over $A$’s feasible action set $\mathcal{A}(x_t)$ such that action $a_t$ is selected with probability $u_t^{(a_t)}$. We note that a pure policy is of course also a mixed policy. We use $u = (u_0, ..., u_{T-1})$ to represent player $A$’s policy, and similarly define the probability distribution $v_t = \{v_t^{(b_t)} | b_t \in \mathcal{B}(x_t)\}$ and the policy $v = (v_0, ..., v_{T-1})$ for player $B$. If $u$ or $v$ is a pure policy, we can simply write them as the action sequences $a = (a_0, ..., a_{T-1})$ or $b = (b_0, ..., b_{T-1})$. A joint policy $\{u, v\}$ is $\mathcal{F}_t$-adapted if each $u_t$ and $v_t$ is $\mathcal{F}_t$-measurable. We denote by $\mathcal{A}^F$ and $\mathcal{B}^F$ the set of all $\mathcal{F}_t$-adapted policies for $A$ and $B$, respectively.

We let $g_t(x_t, u_t, v_t)$ denote the time-$t$ expected payment from $B$ to $A$ when they use the mixed policies $u_t$ and $v_t$. Then

$$
g_t(x_t, u_t, v_t) = \sum_{a_t \in \mathcal{A}(x_t)} \sum_{b_t \in \mathcal{B}(x_t)} u_t^{(a_t)} v_t^{(b_t)} \tilde{g}_t(x_t, a_t, b_t), \quad t = 0, ..., T - 1
$$

and $g_t(x_t, u_t, v_t)$ is $\mathcal{F}_t$-measurable because each $\tilde{g}_t(x_t, a_t, b_t)$ is $\mathcal{F}_t$-measurable. The total expected payment from $B$ to $A$ under these policies is then

$$
g(u, v) = g_T(x_T) + \sum_{t=0}^{T-1} g_t(x_t, u_t, v_t).
$$

From $A$’s perspective, $g(u, v)$ is a reward and so $A$ seeks an optimal $\mathcal{F}_t$-adapted policy to maximize this reward. From $B$’s perspective, $g(u, v)$ is a cost and so $B$ seeks an optimal $\mathcal{F}_t$-adapted policy to minimize this cost. Let $J_0(x_0)$ denote the optimal (equilibrium) value of this zero-sum game. Then a classic result due to Shapley [14] states that $J_0(x_0)$ exists and satisfies

$$
J_0(x_0) = \sup_{u \in \mathcal{A}^F} \inf_{v \in \mathcal{B}^F} \mathbb{E}_0[g(u, v)] = \inf_{v \in \mathcal{B}^F} \sup_{u \in \mathcal{A}^F} \mathbb{E}_0[g(u, v)]
$$

where we have used $\mathbb{E}_t[\cdot]$ to denote $\mathbb{E}[\cdot | \mathcal{F}_t]$. We assume the optimal game value $J_0(x_0)$ is bounded in the following discussion. Let $J_t(x_t)$ denote the time-$t$ optimal value function of the game and note that we can compute it recursively according to

$$
J_T(x_T) = g_T(x_T)
$$

$$
J_t(x_t) = \inf_{v_t} \sup_{u_t} \left\{ g_t(x_t, u_t, v_t) + \mathbb{E}_t[J_{t+1}(f(x_t, u_t, v_t, \omega_{t+1}))] \right\}, \quad t = 0, ..., T - 1
$$

with dynamics in $[\Pi]$.}

---

1. This assumption is not too restrictive. Suppose for example the true amount $B$ pays to $A$ is $g(x_t, a_t, b_t, \omega_{t+1})$ so that it depends on the as yet unobserved disturbance, $\omega_{t+1}$. Then we can replace it with $\tilde{g}_t(x_t, a_t, b_t) := \mathbb{E}[g(x_t, a_t, b_t, \omega_{t+1}) | \mathcal{F}_t]$ where the expectation is taken over $\omega_{t+1}$, and thus $\tilde{g}_t(x_t, a_t, b_t)$ is $\mathcal{F}_t$-measurable.

2. If $A$ uses a pure policy, $a_t$, and $B$ uses a mixed policy, $v_t$, the expected payment reduces to $g_t(x_t, u_t, v_t) = \sum_{b_t \in \mathcal{B}(x_t)} v_t^{(b_t)} \tilde{g}_t(x_t, a_t, b_t)$. We note that $g_t(x_t, u_t, b_t)$ and $g_t(x_t, a_t, b_t)$ can be calculated in a similar manner and that these expressions are all $\mathcal{F}_t$-measurable.
2.1 Obtaining Dual Bounds for the Zero-Sum Game

If we fix $B$’s policy to be $\hat{v} \in B^F$ (which may be sub-optimal from $B$’s perspective), $A$’s problem reduces to

$$J_0(x_0; \hat{v}) = \sup_{a \in A^F} E_0[g(a, \hat{v})]. \quad (7)$$

where we have restricted $A$ to choosing a pure policy rather than a mixed policy. This presents no problem since it is well known that if $B$’s policy is fixed then $A$ will have an optimal pure policy. Problem (7) can be solved via the following DP:

$$J_T(x_T; \hat{v}) = g_T(x_T)$$

$$J_t(x_t; \hat{v}) = \sup_{a_t} \left\{ g_t(x_t, a_t, \hat{v}_t) + E_t[J_{t+1}(f(x_t, a_t, \hat{v}_t, \omega_{t+1}))] \right\}, \quad t = 0, ..., T - 1 \quad (9)$$

with dynamics given by (11) and the understanding that $b_t$ is the (generally random) action obtained when $B$ follows $\hat{v}$. Note that the optimal solution, $J_0(x_0; \hat{v})$, obtained from (8) and (9) is an upper bound on the value of the game, $J_0(x_0)$. Moreover we have equality, i.e. $J_0(x_0; \hat{v}) = J_0(x_0)$, if $\hat{v} = v^*$ where $v^*$ is the optimal policy of $B$ obtained from solving (5) and (6). In general, however, it is hard to even solve for $J_0(x_0; \hat{v})$ but we can use the duality methods of BSS (and outlined in Section 2.2 below) to find an upper bound, say $\bar{J}_0(x_0; \hat{v}, z(\hat{v}))$, on $J_0(x_0; \hat{v})$ where $z(\hat{v})$ is a dual-feasible penalty in the language of BSS.

By a similar argument we can find a lower bound, $\underline{J}_0(x_0; \hat{u}, z(\hat{u}))$, on $J_0(x_0; \hat{u})$, the optimal solution to $B$’s problem when $A$’s policy is fixed at $\hat{u}$. Of course we also have $\underline{J}_0(x_0; \hat{u}) \leq J_0(x_0)$ since in general $A$ is free to choose any feasible policy in $A^F$. Combining these observations we have

$$\underline{J}_0(x_0; \hat{u}, z(\hat{u})) \leq J_0(x_0; \hat{u}) \leq J_{0}(x_0; \hat{v}) \leq \bar{J}_0(x_0; \hat{v}, z(\hat{v})). \quad (10)$$

The outer inequalities in (10) follow from the weak duality of BSS. Moreover there exist dual-feasible penalties, $z^*(\hat{u})$ and $z^*(\hat{v})$ say, for which these outer inequalities become equalities. This is strong duality:

$$\underline{J}_0(x_0; \hat{u}, z^*(\hat{u})) = J_0(x_0; \hat{u}) \leq J_{0}(x_0; \hat{v}) = \bar{J}_0(x_0; \hat{v}, z^*(\hat{v})). \quad (11)$$

Finally, and as stated earlier if we choose $\hat{u} = u^*$ and $\hat{v} = v^*$ then the inequalities in (11) become equalities. In the next subsection we outline the weak and strong duality results of BSS that can be used to justify these statements.

2.2 Constructing Dual Bounds for Player $A$’s Dynamic Program

We will focus on the problem of constructing an upper bound for the value function, $J_0(x_0; \hat{v})$, that is the solution of (8) and (9), which we recall is the DP that $A$ must solve when $B$’s policy is fixed at $\hat{v}$. We can obtain lower bounds on $J_0(x_0; \hat{u})$ in a similar manner.

---

3This is related to the so-called principle of indifference; see Part II, Chapter 2 of Ferguson [6].

4BSS can be consulted for all statements in this section regarding dual bounds, weak and strong duality, and the construction of dual feasible penalties.
Proposition 1. (Weak Duality) For any dual feasible penalty, \( z \in \mathcal{Z}_F \), we have

\[
J_0(x_0; \hat{\mathbf{v}}) \leq \sup_{a \in \mathcal{A}_0^F} \mathbb{E}_0[g(a, \hat{\mathbf{v}}) - z(a, \hat{\mathbf{v}}, \omega)] =: \mathcal{J}_0(x_0; \hat{\mathbf{v}}, z(\hat{\mathbf{v}})).
\] (13)

Therefore any dual feasible penalty and information relaxation provides an upper bound on the optimal value of \( A \)'s dynamic program \( J_0(x_0; \hat{\mathbf{v}}) \). For a given information relaxation, we can optimize the upper bound by minimizing it over the set of dual feasible penalties. This yields the dual problem of player \( A \)'s dynamic program:

\[
\inf_{z \in \mathcal{Z}_F} \left\{ \sup_{a \in \mathcal{A}_0^F} \mathbb{E}_0[g(a, \hat{\mathbf{v}}) - z(a, \hat{\mathbf{v}}, \omega)] \right\}.
\] (14)

We then have the following strong duality result:

Theorem 1. (Strong Duality) If the primal problem \((7)\) for player \( A \) is bounded, then the dual problem \((14)\) has an optimal solution, \( z^* \in \mathcal{Z}_F \), and there is no duality gap so that

\[
J_0(x_0; \hat{\mathbf{v}}) = \inf_{z \in \mathcal{Z}_F} \left\{ \sup_{a \in \mathcal{A}_0^F} \mathbb{E}_0[g(a, \hat{\mathbf{v}}) - z(a, \hat{\mathbf{v}}, \omega)] \right\}.
\] (15)

An optimal penalty is therefore a dual feasible penalty, \( z^* \in \mathcal{Z}_F \), for which

\[
J_0(x_0; \hat{\mathbf{v}}) = \sup_{a \in \mathcal{A}_0^F} \mathbb{E}_0[g(a, \hat{\mathbf{v}}) - z^*(a, \hat{\mathbf{v}}, \omega)].
\]

We can construct such a \( z^* \) according to

\[
z^*(a, \hat{\mathbf{v}}, \omega) = \sum_{t=0}^{T-1} \left( \mathbb{E}[J_{t+1}(x_{t+1}; \hat{\mathbf{v}})|\mathcal{G}_t] - \mathbb{E}[J_{t+1}(x_{t+1}, \hat{\mathbf{v}})|\mathcal{F}_t] \right).
\] (16)

If we use a dual optimal penalty \( z^* \) together with the perfect information relaxation \( \mathcal{G} = \mathcal{I} \), then \((15)\) reduces to

\[
J_0(x_0; \hat{\mathbf{v}}) = \sup_{a \in \mathcal{A}} \mathbb{E}_0[g(a, \hat{\mathbf{v}}) - z^*(a, \hat{\mathbf{v}}, \omega)]
\]

\[
= \mathbb{E}_0 \left[ \sup_{a \in \mathcal{A}} \left\{ g(a, \hat{\mathbf{v}}) - z^*(a, \hat{\mathbf{v}}, \omega) \right\} \right]
\] (17)
and in fact the expectation in (17) is unnecessary since it can be shown that
\[
J_0(x_0; \hat{v}) = \sup_{a \in A} \left\{ g(a, \hat{v}) - z^*(a, \hat{v}, \omega) \right\}
\]
almost surely. In general, however, the \(J_t(x_t; \hat{v})\)'s are hard to compute and so it might be not possible to find an optimal penalty, \(z^*\). Instead we can use an approximation, \(\hat{J}_t(x_t; \hat{v})\), to compute a dual feasible\(^5\) penalty, \(\hat{z}(a, \hat{v}, \omega)\), using (16) but with \(J_t\) replaced by \(\hat{J}_t\). We can then obtain a valid and hopefully good upper bound as follows. Assuming again that \(G = I\), we know from weak duality that
\[
J_0(x_0; \hat{v}) \leq \overline{J}_0(x_0; \hat{v}, \hat{z}(\hat{v})) = \mathbb{E}_0 \left[ \sup_{a \in A} \left\{ g(a, \hat{v}) - \hat{z}(a, \hat{v}, \omega) \right\} \right].
\]
The expectation in (18) therefore yields an upper bound on \(J_0(x_0; \hat{v})\) and it's easily estimated via Monte-Carlo simulation: we simply generate \(N\) sample paths \(\omega^{(i)} = (\omega_0^{(i)}, ..., \omega_{T-1}^{(i)})\) for \(i = 1, ..., N\), and solve the deterministic “inner problem"
\[
\overline{J}_0^{(i)}(x_0; \hat{v}, \hat{z}(\hat{v})) := \sup_{a \in A} \left\{ g(a, \hat{v}) - \hat{z}(a, \hat{v}, \omega^{(i)}) \right\}.
\]
along each path. The average, \(\frac{1}{N} \sum_{i=1}^{N} \overline{J}_0^{(i)}(x_0; \hat{v}, \hat{z}(\hat{v}))\), is then an unbiased estimate of the upper bound, \(\overline{J}_0(x_0; \hat{v}, \hat{z}(\hat{v}))\).

**A Simple Example**

We first consider a single-period two-person zero-sum game where the payoff is defined by an \(m \times n\) matrix, \(R\). In this game players \(A\) and \(B\) simultaneously select a row, \(a\), and a column, \(b\), respectively, after which \(B\) pays \(A\) the value \(R_{a,b}\). A mixed policy for player \(A\) is an \(m \times 1\) vector \(u = [u^{(1)} ... u^{(m)}]'\) of probabilities that sum to 1 such that \(A\) chooses the \(a\)-th row with probability \(u^{(a)}\). Similarly, we let \(v = [v^{(1)} ... v^{(n)}]'\) represent player \(B\)'s mixed policy. The expected payoff to \(A\) is then \(u' R v\).

Consider now a 2-period dynamic game defined by the following three matrix games:

\[
R^{(1)} = \begin{bmatrix} 2 & 1 \\ 6 & 8 \end{bmatrix}, \quad R^{(2)} = \begin{bmatrix} 8 & 15 \\ 10 & 12 \end{bmatrix} \quad \text{and} \quad R^{(3)} = \begin{bmatrix} -8 & -10 \\ 3 & -11 \end{bmatrix}.
\]

The state variable \(x_t \in \{i \mid i = 1, 2, 3\}\) determines the game \(R^{(x_t)}\) that is played at time \(t\), and \(\hat{g}(x_t, a_t, b_t)\) is then the payoff of that game when \(A\) and \(B\) choose the \(a_t\)-th row and \(b_t\)-th column, respectively at time \(t\). The game begins at \(t = 0\) with \(A\) and \(B\) playing \(R^{(1)}\) so the initial state is \(x_0 = 1\). The next matrix game to be played at time \(t = 1\) is determined by \(A\) and \(B\)'s actions and a transition probability, \(p(x_t, x_{t+1}, a_t, b_t)\). These transition probabilities are determined by the \(a_t\)-th row and \(b_t\)-th column element in the matrices \(P_{x_t,x_{t+1}}\) which are defined as:

\[
P_{1,2} = \begin{bmatrix} 0.7 & 0.55 \\ 0.4 & 0.5 \end{bmatrix} \quad \text{and} \quad P_{1,3} = \begin{bmatrix} 0.3 & 0.45 \\ 0.6 & 0.5 \end{bmatrix}.
\]

\(^5\)It can be shown that a penalty, \(\hat{z}(a, \hat{v}, \omega)\), constructed from (16) using \(\hat{J}_t(x_t; \hat{v})\) instead of \(J_t\) is indeed a dual feasible penalty.
We estimate the dual bound by simulating 10,000 dual problem instances, i.e. by simulating
bound taking the average of the optimal objective functions. Our simulation yielded an estimated upper
where \( \hat{J} \) uses a perfect information relaxation, and construct the penalty \( \hat{z} \)
the information relaxation approach to compute an upper bound on \( J(2) \) at time \( t \).

For more complex games we would not be able to compute \( J(x_0; \hat{v}) \) but we can instead use
the information relaxation approach to compute an upper bound on \( J_0(x_0; \hat{v}) \). Suppose then we
use a perfect information relaxation, and construct the penalty \( \hat{z} = \hat{J}_1(x_1; \hat{v}) - \mathbb{E} [\hat{J}_1(x_1; \hat{v}) | F_0] \)
where \( \hat{J}_1(x_1; \hat{v}) \) is an approximate value for player A’s DP. For illustrative purposes, we take
\( \hat{J}_1(x_1; \hat{v}) = 8 \cdot 1_{\{x_1=2\}} - 8 \cdot 1_{\{x_1=3\}} \).

We estimate the dual bound by simulating 10,000 dual problem instances, i.e. by simulating
10,000 values of \( \omega \), solving the deterministic dual inner problem \( (19) \) for each instance, and then
taking the average of the optimal objective functions. Our simulation yielded an estimated upper
bound \( \bar{J}_0(x_0; \hat{v}, \hat{z}(\hat{v})) = 5.82 \) with a standard error of 0.02. Note that our numerical results, i.e.
\( J_0(x_0) < J_0(x_0; \hat{v}) < \bar{J}_0(x_0; \hat{v}, \hat{z}(\hat{v})) \), are consistent with weak duality in \( (10) \).

\[ \hat{J}_1(x_1; \hat{v}) = \hat{J}_1(x_1) \]

\[ J_0(x_0; \hat{v}) = \max \left\{ R\hat{v}_0 = \begin{bmatrix} 4.4 \\ 5.6 \end{bmatrix} \right\} = 5.6 \text{ where } R = R^{(1)} + J_1(2)P_{1,2} + J_1(3)P_{1,3} = \begin{bmatrix} 6 & 2 \\ 4 & 8 \end{bmatrix}. \]

\begin{table}
\centering
\caption{Optimal Policies and Game Values}
\begin{tabular}{cccc}
\hline
\( t \) & \( x_t \) & \( u_t^* \) & \( v_t^* \) & \( J_t(x_t) \) \\
\hline
0 & 1 & \begin{bmatrix} 0.5 \\ 0.75 \end{bmatrix}^T & \begin{bmatrix} 0.25 \end{bmatrix}^T & 5 \\
1 & 2 & \begin{bmatrix} 0 & 1 \end{bmatrix}^T & \begin{bmatrix} 0 \\ 1 \end{bmatrix}^T & 10 \\
1 & 3 & \begin{bmatrix} 1 & 0 \end{bmatrix}^T & \begin{bmatrix} 0 & 1 \end{bmatrix}^T & -10 \\
\hline
\end{tabular}
\end{table}

Note of course that \( P_{1,2} + P_{1,3} \) is equal to the matrix with 1 in every entry. For example, if \( A \)
chooses the second row (\( a_0 = 2 \)) and \( B \) chooses the first column (\( b_0 = 1 \) when playing \( R^{(1)} \) at
time \( t = 0 \), \( B \) pays \( A \) amount of \( \hat{g}(x_0 = 1, a_0 = 2, b_0 = 1) = 6 \). At time \( t = 1 \) \( A \) and \( B \) will then
play game \( R^{(2)} \) with probability 0.4, or game \( R^{(3)} \) with probability 0.6. We assume the random
variable, \( \omega_1 \), which drives the state transition is uniformly distributed so that the state evolution
can be written as

\[ x_1 = f_0(x_0, a_0, b_0, \omega_1) = 2 \cdot 1_{\{\omega_1 \leq p(1,2,a_0,b_0)\}} + 3 \cdot (1 - 1_{\{\omega_1 \leq p(1,2,a_0,b_0)\}}) \]

with probability 0, \( x_1 \) at time 1, \( J_1(x_1; \hat{v}) \) is then equal to \( J_1(x_1) \) given in Table 1. In
at time \( t = 0 \), \( R \) is the expected total payoff matrix, and thus \( A \)’s optimal game value under
\( B \)’s fixed policy \( \hat{v} \) is \( J_0(x_0; \hat{v}) = 5.6 \). We note that \( J_0(x_0; \hat{v}) \) is an upper bound on the value of
game, \( J_0(x_0) = 5 \).

Weak Duality

For more complex games we would not be able to compute \( J_0(x_0; \hat{v}) \) but we can instead use
the information relaxation approach to compute an upper bound on \( J_0(x_0; \hat{v}) \). Suppose then we
use a perfect information relaxation, and construct the penalty \( \hat{z} = \hat{J}_1(x_1; \hat{v}) - \mathbb{E} [\hat{J}_1(x_1; \hat{v}) | F_0] \)
where \( \hat{J}_1(x_1; \hat{v}) \) is an approximate value for player A’s DP. For illustrative purposes, we take
\( \hat{J}_1(x_1; \hat{v}) = 8 \cdot 1_{\{x_1=2\}} - 8 \cdot 1_{\{x_1=3\}} \).

This is the game value at time \( t = 1 \) if \( A \) always chooses first row and \( B \) always chooses first column when playing
\( R^{(1)} \).

\[ \bar{J}_0(x_0; \hat{v}, \hat{z}(\hat{v})) = 5.82 \]
**Strong Duality**

If instead we constructed the dual optimal penalty \( z^* = J_1(x_1; \hat{v}) - \mathbb{E}[J_1(x_1; \hat{v}) | \mathcal{F}_0] \) according to (16) using \( J_1(x_1; \hat{v}) \) in (21), each dual inner problem (19) yields an upper bound \( \mathcal{J}_0(x_0; \hat{v}, z^*(\hat{v})) = 5.6 = J_0(x_0; \hat{v}) \). We have therefore demonstrated strong duality (for \( B \)'s fixed policy, \( \hat{v} \)) as given in (11).

Note also that if we fixed \( B \)'s policy at \( v^* \) and repeated the numerical calculations above, then we would obtain \( J_0(x_0) = J_0(x_0; v^*) = J_0(x_0; v^*, z^*(v^*)) \). Corresponding lower bounds for \( J_0(x_0) \) can be obtained by fixing \( A \)'s policy and solving or lower bounding player \( B \)'s resulting DP.

### 3 Dual Bounds for Stochastic Shortest Path Games

In this section we extend the information relaxation approach to more general stochastic shortest-path (SSP) zero-sum games. Note that infinite horizon discounted games can be modeled as SSP zero-sum games so the formulation is quite general and indeed general conditions exist to guarantee that these games have an optimal value. Assuming these games (i) have an optimal value and (ii) have optimal policies under which the game terminates finitely almost surely, we can replicate the information relaxation methods of Section 2 to obtain valid dual bounds. Our discussion below assumes that both (i) and (ii) hold.

One potential difficulty that arises in constructing dual bounds in this case is in simulating a dual sample path. In particular, how many time periods are there in such a dual path? In general the number of periods (before absorption into a terminal state) will depend on the policies employed by the two players but we don’t know these policies in advance so how can we simulate such a path? Very recent work of Brown and Haugh [2] describes several approaches for addressing this problem. One such approach is to use the Rogers [13] approach to finite horizon problems with the perfect information relaxation. Under his approach dual sample paths are simulated under a reference transition probability measure, \( P^* \), that is action-independent. In the context of our (generally infinite-horizon) SSP games, we can use such a \( P^* \) to generate dual sample paths as long as all such paths terminate after a finite number of periods \( P^*- \)almost surely. We also need to choose \( P^* \) so that there exists an optimal policies whose induced probability measure (over sample paths) is absolutely continuous with respect to the measure induced by \( P^* \).

More specifically dual bounds are constructed as follows, where for ease of exposition we use the perfect information relaxation, \( \mathbb{I} \). We generate the state path \( x = (x_0, \ldots, x_T) \) using the action-independent transition matrix \( P^* \) where \( p^*(x_t, x_{t+1}) \) denotes the transition probability from state \( x_t \) to \( x_{t+1} \). Note that while \( T \) is random, this does not present a problem in simulating a dual sample path because \( P^* \) is action-independent and so the realized value of \( T \) does not depend on the policies of the two players.

---

7 We note that (i) and (ii) hold for the two examples that we consider in this section. In fact the game in our first example terminates after a finite number of periods almost surely regardless of what policies the players employ. The game in our second example does have policies under which the game never terminates. However, there does exist an optimal policy for each player under which the game terminates almost surely after a finite number of periods. This is sufficient to guarantee (i) and (ii); see Patek [11].

8 There are better approaches; see Brown and Haugh [2] for details.
Because we used $P^*$ to generate dual paths and not the natural transition matrix induced by the players’ actions, it is necessary to multiply the various rewards and dual value functions by appropriate action-dependent Radon-Nikodym derivative terms; see Rogers [13] or Brown and Haugh [2]. In particular, suppose we fix $B$’s policy at $\hat{\nu}$ and use a dual feasible penalty, $\hat{z}$, that has been constructed from some approximated value function, $\hat{J}_t(x_t; \hat{\nu})$. Then the dual inner problem on a $P^*$-simulated path that player $A$ must solve has the form

$$
sup_{a \in A_t} \left\{ \Lambda_t(a) g_T(x_T) + \sum_{t=0}^{T-1} \Lambda_t(a) \left[ g_t(x_t, a_t, \hat{v}_t) + \mathbb{E} \left[ \hat{J}_{t+1}(x_{t+1}; \hat{\nu}) | F_t \right] - \phi(x_t, x_{t+1}, a_t, \hat{v}_t) \hat{J}_{t+1}(x_{t+1}; \hat{\nu}) \right] \right\} (23)
$$

where

$$
\Lambda_t(a) := \prod_{\tau=0}^{t-1} \phi(x_{\tau}, x_{\tau+1}, a_{\tau}, \hat{v}_{\tau}) (24)
$$

and where the numerator in the Radon-Nikodym derivative term (25) is the probability that the game transitions from $x_t$ to $x_{t+1}$ under $A$’s action $a_t$ and $B$’s mixed policy $\hat{v}_t$. Note also that the expectations in (23) are taken with respect to the action-dependent transitions $p(\cdot)$. The objective in (24) is fully deterministic and can be solved recursively according to

$$
\mathcal{J}_T(x_T; \hat{\nu}, \hat{z}(\hat{\nu})) = g_T(x_T)
$$

$$
\mathcal{J}_t(x_t; \hat{\nu}, \hat{z}(\hat{\nu})) = \sup_{a_t} \left\{ g_t(x_t, a_t, \hat{v}_t) + \mathbb{E} \left[ \hat{J}_{t+1}(x_{t+1}; \hat{\nu}) | F_t \right] + \phi(x_t, x_{t+1}, a_t, \hat{v}_t) \left( \mathcal{J}_{t+1}(x_{t+1}; \hat{\nu}, \hat{z}(\hat{\nu})) - \hat{J}_{t+1}(x_{t+1}; \hat{\nu}) \right) \right\}, \quad t = 0, ..., T - 1.
$$

A dual bound is then estimated by simulating $N$ sample paths under $P^*$, solving (23) along each of these paths, and then taking the average of the optimal objective functions.

**Example: A Stochastic Shortest-Path Matrix Game**

We construct an SSP game using the two matrix games

$$
R^{(1)} = \begin{bmatrix} 4 & 0 \\ 1 & 3 \end{bmatrix} \quad \text{and} \quad R^{(2)} = \begin{bmatrix} 0 & -5 \\ -4 & 1 \end{bmatrix}.
$$

The state variable is $x_t$ and a value of $x_t \in \{1, 2\}$ implies that $R^{(x_t)}$ is played at time $t$. The state $x_t = 3$ denotes a cost-free absorbing termination state and the transition probability matrices are

$$
P_{1,1} = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.5 \end{bmatrix} \quad P_{1,2} = \begin{bmatrix} 0 & 0.4 \\ 0.4 & 0 \end{bmatrix} \quad P_{1,3} = \begin{bmatrix} 0.7 & 0.6 \\ 0.6 & 0.5 \end{bmatrix}
$$

$$
P_{2,1} = \begin{bmatrix} 0.5 & 0 \\ 0 & 0 \end{bmatrix} \quad P_{2,2} = \begin{bmatrix} 0 & 0 \\ 0 & 0.5 \end{bmatrix} \quad P_{2,3} = \begin{bmatrix} 0.5 & 1 \\ 1 & 0.5 \end{bmatrix}.
$$
We note \( \sum_{j=1}^{3} P_{i,j} \) for \( i = 1, 2 \) is a matrix with every entry equal to 1. This of course must be the case. Suppose now, for example, that \( x_0 = 1 \) so that \( A \) and \( B \) initially play \( R(1) \). If \( A \) chooses the second row \( (a_0 = 2) \) and \( B \) chooses the first column \( (b_0 = 1) \), then \( B \) pays \( A \) 1 unit and at time \( t = 1 \) they will play \( R(2) \) with probability 0.4, and enter the absorbing state with probability 0.6.

In order to compute dual bounds we begin with the two sub-optimal policies, \( \hat{u}^{(1)} \) and \( \hat{v}^{(1)} \), for players \( A \) and \( B \), respectively. These policies are defined in the first row of Table 2. We can simulate these policies to obtain estimates of the game value, \( J_0(x_0, \hat{u}^{(1)}, \hat{v}^{(1)}) \) as a function of \( x_0 \). We use these estimates as our approximate value function, \( \hat{J}_t(x_t) \), that we use to construct a dual feasible penalty. Finally, we can then estimate dual lower and upper bounds on the optimal game value by simulating sample paths and solving the corresponding dual problem instances as in (23).

The sample paths (for estimating the dual bounds) were simulated using the action-independent transition matrix

\[
P^* = \begin{bmatrix}
\frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
\frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
0 & 0 & 1
\end{bmatrix}.
\]

and the corresponding dual bounds (mean and standard errors) are displayed in the first two rows of Table 3.

The dual lower and upper bounds are quite far apart and we expect this is because \( \hat{u}^{(1)} \) and \( \hat{v}^{(1)} \) are far from optimal. We therefore perform a policy iteration to find new and improved suboptimal policies \( (\hat{u}^{(i+1)}, \hat{v}^{(i+1)}) \) based on \( J_0(x_0, \hat{u}^{(i)}, \hat{v}^{(i)}) \) for \( i = 1, 2 \). These new policies are also displayed in Table 2. We can also estimate their game values and use them to construct dual feasible penalties and dual bounds. These game values and dual bounds are displayed in Table 3. As expected, we see that the duality gap, i.e. the difference between the dual upper and lower bounds, decreases with each policy iteration for each value of \( x_0 \). In fact, we can use our dual bounds to see that the estimates \( J_0(1) = 2.025 \) and \( J_0(2) = -1.99 \) are very accurate with high probability. Note also that we did not actually compute the optimal game value for this problem.

Example: An Industrial-Waste Inspection Game

We now consider a more practical game as discussed in Patek [11]. The participants of the game are a manufacturer (player \( A \)) who produces industrial waste which has to be dumped every night, and an inspector (player \( B \)) who wants to catch the manufacturer dumping. There is a finite number of geographically disparate sites where industrial waste can be dumped. \( A \) must dump waste at one
of these sites every night while avoiding detection by B. In order to detect dumping activity on a
given night, B must inspect the same site where A is dumping that night and even then, there is
a nonzero probability of failing to catch A.

Conditional on A and B selecting the same site, the probability of detection depends upon two
consideration:

1. The closer the current dump site is to the preceding dump site, the greater the probability of
detection. (This is due to the environment’s limited ability to absorb the waste.)

2. The closer the current inspection site is to the preceding inspection site, the greater the
probability of detection. (This models the fact that the inspector needs less time to travel
and so he has more time to look for dumping activity.)

If B detects A two nights in a row, then A is put out of business. A’s objective is to maximize
its time in business, while B’s objective is to minimize A’s time in business. In deciding where to
dump / inspect each night, we assume that both players know where the dumping and inspection
occurred on the previous night. They also know whether or not A was caught dumping the previous
night.

We can formulate this as a zero-sum SSP game as follows. Let \( \mathcal{L} = \{ l_1, ..., l_N \} \) represent the N
sites where waste may be dumped. Suppose at time \( t - 1 \) A dumped waste at a site \( a_{t-1} \in \mathcal{L} \) and
that B inspected the site \( b_{t-1} \in \mathcal{L} \). Let \( y_t \) be a Boolean variable which is TRUE if A was caught
at time \( t - 1 \). The state vector \( x_t = (a_{t-1}, b_{t-1}, y_t) \) then describes the state of the system at time
\( t \); there are \( N^2 + N \) possible non-terminal states and one absorbing state. Given the manufacturer
A has not yet been shut down at time \( t \), A chooses to dump at site \( a_t \in \mathcal{A}(x_t) = \mathcal{L} \) and B chooses
to search at site \( b_t \in \mathcal{B}(x_t) = \mathcal{L} \). The probability that A will then be detected on night \( t \) is

\[
p^d(x_t, a_t, b_t) = \begin{cases} p^d + \frac{p^d - p^d}{(k_1 + k_2)d} [k_1 d(a_t, a_{t-1}) + k_2 d(b_t, b_{t-1})], & \text{if } a_t = b_t \\ 0, & \text{otherwise} \end{cases}
\]

where \( d(l_i, l_j) \) denotes the distance between the two sites, \( l_i \) and \( l_j \), \( \bar{d} = \max_{i,j} d(l_i, l_j) \) is the
maximum distance between any two sites, \( 0 < p^d < 1 \) are the worst-case and ideal\(^{10} \) proba-

\(|^{10}p^d \) is possible only if there is some pair of sites, \( l_i \) and \( l_j \), for which \( d(l_i, l_j) = 0 \).
Figure 1:

The main contribution of this work is to generalize the work of Beveridge and Joshi [1] on optimal stopping games to dynamic zero-sum games. We used the general results of BSS and Rogers [13] on possibilities of detection, and $k_1$ and $k_2$ are positive weighting factors. If $y_t = \text{FALSE}$, then the system transitions to state $(a_t, b_t, \text{TRUE})$ with probability $p^d(x_t, a_t, b_t)$, otherwise the system transitions to $(a_t, b_t, \text{FALSE})$. If $y_t = \text{TRUE}$, then the game transitions to the absorbing terminal state and ends with probability $p^d(x_t, a_t, b_t)$, otherwise the system transitions to $(a_t, b_t, \text{FALSE})$. The cost is 1 in each time period regardless of the controls applied.\footnote{Any pure policy for the inspector $B$ is improper in this game because, knowing this policy, the manufacturer $A$ can always avoid being detected. On the other hand, a mixed policy which select each site with a positive probability will eventually result in $A$ getting caught, so there exists a proper policy.}

We consider a case with $N = 10$ sites that are located on a straight line with equal distances between successive sites where $d(l_i, l_j) = |j - i|$. We assume parameter values of $p^d = 0.5, \bar{p}^d = 0.95, k_1 = 2$ and $k_2 = 1$. We begin with suboptimal policies where $A$ and $B$ select sites uniformly each night. We simulate the game with these suboptimal policies and obtain corresponding estimates of the game value. We then use these policies and estimated game value to construct dual feasible penalties and then lower and upper dual bounds as in the matrix game example. We then improve these policies via policy iteration, and obtain a new set of estimated game values as well as lower and upper dual bounds. The results are displayed in Figure 1 for the state $(l_1, l_1, \text{FALSE})$. Note that our dual bounds become very tight after just one policy iteration.

4 Conclusions and Further Research

The main contribution of this work is to generalize the work of Beveridge and Joshi [1] on optimal stopping games to dynamic zero-sum games. We used the general results of BSS and Rogers [13] on...
information relaxations for general dynamic programs to do this. While the results were straightforward to prove, there are many interesting applications including pursuit-evasion games, heads-up poker and many two-person computer games. While we considered relatively simple numerical examples in this paper, the ultimate value of this work will be in applying it successfully to more realistic and interesting applications.

There are several potential problems that arise when we consider these more complex applications. For example, many zero-sum games such as heads-up poker are more naturally studied in so-called extensive form whereas our dual formulation requires the games to be modeled in strategic form. While it is certainly possible to go back and forth between these two forms, it is not clear how much work is required to do so. Nor is it clear that sub-optimal policies that have been designed for the extensive form of the game are easily handled in strategic form. Computing easy-to-use dual penalties from these sub-optimal strategies may also be problematic.

Another potential difficulty relates to the issue of private information. Our analysis has assumed the players do not have private sources of information but in many games, however, private information does exist. For example, in heads-up poker each player knows his two “hole” cards but does not know his opponent’s “hole” cards until the end of the game if ever. From a theoretical perspective, it should be easy to handle this complication but it may well raise practical problems that are difficult to overcome with limited computing resources.

There are many other interesting complications related to learning and bounded rationality that would also be interesting to explore. We hope to address some of these problems in future research.

References

[1] C. Beveridge and M. Joshi. Monte carlo bounds for game options including convertible bonds. *Management Science*, 57(5):960–974, 2011.

[2] D.B. Brown and M.B. Haugh. Information relaxation bounds for infinite horizon stochastic control. Technical report, 2014.

[3] D.B. Brown and J.E. Smith. Dynamic portfolio optimization with transaction costs: heuristics and dual bounds. *Management Science*, 57(10):1752–1770, 2011.

[4] D.B. Brown, J.E. Smith, and P. Sun. Information relaxations and duality in stochastic dynamic programs. *Operations Research*, 58(4):785–801, 2010.

[5] V.V. Desai, V.F. Farias, and C.C. Moallemi. Bounds for markov decision processes. Technical report, 2011.

[6] T.S. Ferguson. Game theory. [http://www.math.ucla.edu/~tom/Game_Theory/mat.pdf](http://www.math.ucla.edu/~tom/Game_Theory/mat.pdf).

[7] M.B. Haugh and L. Kogan. Pricing american options: a duality approach. *Operations Research*, 52(2):258–270, 2004.

---

12 See, for example, Ferguson [6] for a detailed discussion on zero-sum games as well as the strategic and extensive forms of these games.
[8] M.B. Haugh and C. Wang. Dynamic portfolio execution and information relaxations. *SIAM Journal on Financial Mathematics*, Forthcoming, 2013.

[9] G. Lai, F. Margot, and N. Secomandi. An approximate dynamic programming approach to benchmark practice-based heuristics for natural gas storage valuation. *Operations Research*, 58:564–582, 2010.

[10] C. Moallemi and M. Saglam. Dynamic portfolio choice with linear rebalancing rules. Technical report, 2011.

[11] S.D. Patek. Stochastic shortest path games: theory and algorithms. Technical report, 1997.

[12] L.C.G. Rogers. Monte carlo valuation of american options. *Mathematical Finance*, 12(3):271–286, 2002.

[13] L.C.G. Rogers. Pathwise stochastic optimal control. *SIAM Journal on Control and Optimization*, 46:1116–1132, 2007.

[14] L.S. Shapley. Stochastic games. *Proceedings of the national academy of sciences, Mathematics*, 39:1095–1100, 1953.