Linear-Quadratic Control and Information Relaxations

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

Andrew Lim
Department of IE and OR, University of California, Berkeley, CA 94720, lim@ieor.berkeley.edu.

This draft: 22-April-2011

Abstract

We apply the recently developed duality methods based on information relaxations to the classic linear quadratic (LQ) control problem. We derive two dual optimal penalties for the LQ problem when the control space is unconstrained. These two penalties, which are derived using value function and gradient methods, respectively, may be used to evaluate sub-optimal policies for constrained LQ problems when it is not possible to determine the optimal policy exactly. We also compare these dual penalties to the dual penalty of Davis and Zervos (1994). This connection to the earlier work of Davis and Zervos is not widely known and demonstrates that some of these duality ideas have been in circulation for some time. We also emphasize that while the three penalties are dual optimal, they are not identical. Indeed their differences have significant implications when the penalties are used via Monte-Carlo to evaluate sub-optimal policies for constrained LQ problems. Our conclusions should apply more generally to other stochastic control problems.

1 Introduction

In this note we apply recently developed duality techniques for stochastic control problems to the classic linear quadratic (LQ) control problem. These techniques were developed independently by Rogers (2007) and Brown, Smith and Sun (2010) and are based on relaxing the decision-maker's information constraints. This work was motivated in part by the duality techniques developed by Davis and Karatzas (1994), Rogers (2002) and Haugh and Kogan (2004) for optimal stopping problems and the pricing of American options in particular. These duality techniques (of Rogers 2007 and Brown et al. 2010) can be used to evaluate sub-optimal policies for control problems that are too difficult to solve exactly. In particular, the sub-optimal policy may be used to compute both primal and dual bounds on the optimal value function. The primal bound can be computed by simply simulating the sub-optimal policy whereas the aforementioned duality techniques can be used with the sub-optimal policy (or indeed some other policy) to compute the dual bound. If the primal and dual bounds are close to one another then we know that the sub-optimal policy is close to optimal. We believe that these techniques will play an increasingly important role in the area of sub-optimal control and that there are many interesting related research questions to be resolved.

In this note we focus on finite horizon LQ problems and we derive two dual optimal penalties when the control space is unconstrained. The first penalty is derived using knowledge of the optimal value function whereas the second penalty is derived using the gradient methods developed by Brown and Smith (2010) in the context of dynamic portfolio optimization under transaction costs. These penalties and others may then be used to evaluate sub-optimal policies for constrained LQ problems when it is not possible to determine the optimal policy exactly. If the controls are not too constrained then we expect the optimal unconstrained penalties to be close to optimal for the constrained problem and therefore to lead to good dual bounds. We emphasize that the derivation of these penalties is quite straightforward and is only a modest contribution to this growing literature.

We also compare these dual techniques to the work of Davis and Zervos (1994) who used Lagrange multipliers to show that a stochastic LQ problem may be reduced to a deterministic LQ problem. Indeed it is easy to show that their Lagrange multipliers are also optimal dual penalties. This connection to the earlier work of Davis and Zervos is not widely known and it highlights that some of these duality ideas have been in circulation for some time. In fact within the stochastic control literature[1] the idea of relaxing the non-anticipativity constraints goes back at least to Davis (1989, 1991). It is also interesting to note that these developments appear to mirror the development of the duality methods for solving optimal stopping problems as mentioned earlier. In this case, Davis and Karatzas (1994) used a dual formulation to characterize the optimal solution to the optimal stopping problem. Rogers (2002) and Haugh and Kogan (2004) were not aware[2] of Davis and Karatzas when they independently developed dual formulations of the optimal stopping problem. Their focus, however, was on using these dual formulations to construct good dual bounds for optimal stopping problems that were too difficult to solve exactly. This was also the focus of Rogers (2007) and Brown et al. (2010) who developed their dual techniques with a view to using them to evaluate sub-optimal strategies. In contrast, the seminal work of Davis and his co-authors appears to have been only on characterizing optimal solutions.

A further contribution of this note is a comparison of the three optimal dual penalties for the unconstrained LQ problem. We emphasize that while the three penalties are dual optimal, they are not actually identical. Indeed as demonstrated by Brown et al. (2010), the penalty function

[1] It has also been a feature of the stochastic programming literature where it has often been applied to stochastic programs with just a few periods.

[2] Davis and Karatzas (1994) published their paper as a book chapter and as a result, was not widely known until sometime afterwards.

constructed using the optimal value function is almost surely optimal whereas the other penalties are only optimal in expectation[3]. This observation should have significant implications when we use dual penalties to evaluate sub-optimal policies for constrained control problems in general. However, it is also worth mentioning that dual penalties constructed using value function approximations can be quite challenging to work with and so we expect the gradient approach to often be a viable alternative.

The remainder of this paper is organized as follows. We briefly review the duality approach of Brown et al. (2010) in Section 2 and then derive optimal dual penalties for the unconstrained LQ problem in Section 3. We review the results of Davis and Zervos (1994) in Section 4 and compare their dual penalty to the dual penalties derived in Section 3. We conclude in Section 5 and identify several directions for further research.

2 Review of Duality Based on Information Relaxations

We begin with a general finite-horizon discrete-time dynamic program with a probability space, $(\Omega, \mathcal{F}, \mathbb{P})$. Time is indexed by $k = 0, \ldots, N$ and the evolution of information is described by the filtration $\mathbb{F} = \{\mathcal{F}_0, \ldots, \mathcal{F}_N\}$ with $\mathcal{F} = \mathcal{F}_N$. We make the usual assumption that $\mathcal{F}_0 = \{\emptyset, \Omega\}$ so that the decision maker starts out with no information regarding the outcome of uncertainty. There is a state vector, $x_k \in S_k$, where S_k is the time k state space. The dynamics of x_k satisfy

$$x_{k+1} \;=\; f_k(x_k, u_k, w_{k+1}), \qquad k = 0, \ldots, N-1 \tag{1}$$

where $u_k \in U_k(x_k)$ is the control taken at time k and w_{k+1} is an $\mathcal{F}_{k+1}$-measurable random disturbance. A feasible strategy, $u := (u_0, \ldots, u_{N-1})$ is one where each individual[4] action, $u_k \in U_k(x_k)$ is $\mathcal{F}_k$-measurable. In particular, we require the decision-maker's strategy, u, to be be $\mathcal{F}_k$-adapted. We use $\mathcal{U}_{\mathbb{F}}$ to denote the set of all such $\mathcal{F}_k$-adapted strategies. The objective is to select a feasible strategy, u, to minimize the expected total cost,

$$g(u) \;:=\; g_N(x_N) \;+\; \sum_{k=0}^{N-1} g_k(x_k, u_k)$$

where we assume[5] each $g_k(x_k, u_k)$ is $\mathcal{F}_k$-measurable. In particular, the decision maker's problem is then given by

$$J_0(x_0) \;\equiv\; \inf_{u \in \mathcal{U}_{\mathbb{F}}} \mathbb{E}_0 \left[g_N(x_N) \;+\; \sum_{k=0}^{N-1} g_k(x_k, u_k) \right] \tag{2}$$

where the expectation in (2) is taken over the set of possible outcomes, $w = (w_1, \ldots, w_{N-1}) \in \Omega$. To emphasize that the total cost is random, we will often write $g(u, w)$ for $g(u)$. Letting J_k denote the time-k value function for the problem (2), the associated dynamic programming recursion is given by[6]

$$\begin{aligned} J_N(x_N) \;&:=\; g_N(x_N) \\ J_k(x_k) \;&:=\; \inf_{u_k \in U_k(x_k)} \{ g_k(x_k, u_k) \;+\; \mathbb{E}_k \left[J_{k+1} \left(f_k(x_k, u_k, w_{k+1}) \right) \right] \} \qquad k = 0, \ldots, N-1. \end{aligned} \tag{3}$$

[3] We will clarify this statement in Section 2.

[4] Brown et al. (2010) use a slightly more general formulation where they assume that $U_k = U_k(u_0, \ldots, u_{k-1})$ can depend on the entire history of past actions and states.

[5] This assumption is without loss of generality. Suppose for example the true time k cost is $\tilde{g}_k(x_k, u_k, w_{k+1})$ so that it depends on the as yet unobserved disturbance, w_{k+1}. Then we can replace this cost with $g_k(x_k, u_k) := \mathbb{E}_k \left[\tilde{g}_k(x_k, u_k, w_{k+1}) \right]$ which is $\mathcal{F}_k$-measurable.

[6] We write $\mathbb{E}_k \left[\cdot \right]$ for $\mathbb{E} \left[\cdot \mid \mathcal{F}_k \right]$ hereafter.

In practice of course it is often too difficult or time-consuming to perform the iteration in (3). This can occur, for example, if the state vector, x_k, is high-dimensional or if the constraints imposed on the controls are too complex or difficult to handle. In such circumstances, we must be satisfied with sub-optimal solutions or policies.

2.1 The Dual Formulation

We now briefly describe the dual formulation of Brown et al. (2010) which should be consulted for further details and proofs of the results given below. Note, however that Brown et al. (2010) focus on problems where the primal problem is a maximization problem. We have chosen to specify our primal problem as a minimization problem so that we are consistent with the usual formulation of linear-quadratic problems where the goal is to minimize expected total costs.

We say the filtration $\mathbb{G} := \{\mathcal{G}_k\}$ is a relaxation of $\mathbb{F}$ if, for each k, $\mathcal{F}_k \subseteq \mathcal{G}_k$. We write $\mathbb{F} \subseteq \mathbb{G}$ to denote such a relaxation. For example, the perfect information filtration, $\mathbb{I} := \{\mathcal{I}_k\}$, is obtained by taking $\mathcal{I}_k = \mathcal{F}$ for all k. We let $\mathcal{U}_\mathbb{G}$ denote the set of all $\mathcal{G}_k$-adapted strategies. It is clear then that for any relaxation, $\mathbb{G} := \{\mathcal{G}_k\}$, we have $\mathcal{U}_\mathbb{F} \subseteq \mathcal{U}_\mathbb{G} \subseteq \mathcal{U}_\mathbb{I}$ so that as we relax the filtration, we expand the set of feasible policies.

The set of penalties, $\mathcal{Z}$, is the set of all functions $z(u, w)$ that, like the set of costs, depend on the choice of actions, u, and the outcome, w. We define the set, $\mathcal{Z}_\mathbb{F}$, of dual feasible penalties to be those penalties that do not penalize temporally feasible, i.e. $\mathcal{F}_k$-adapted, strategies. In particular, we define

$$\mathcal{Z}_\mathbb{F} \;:=\; \{z \in \;\mathcal{Z}:\; \mathbb{E}_0\left[z(u,w)\right] \leq 0 \;\; \text{for all } u \in \mathcal{U}_\mathcal{F}\}. \tag{4}$$

We then have the following version of weak duality, the proof of which follows immediately from the definition of dual feasibility in (4) and because $\mathbb{G}$ is a relaxation of $\mathbb{F}$.

Lemma 1 (Weak Duality)
If u_F and z are primal and dual feasible respectively, i.e. $u_F \in \mathcal{U}_\mathbb{F}$ and $z \in \mathcal{Z}_\mathbb{F}$, then

$$\mathbb{E}_0\left[g(u_F, w)\right] \;\geq\; \inf_{u_G \in \mathcal{U}_\mathbb{G}} \mathbb{E}_0\left[g(u_G, w) + z(u_G, w)\right]. \tag{5}$$

Therefore any dual feasible penalty and information relaxation provides a lower bound on the optimal value function. Clearly weaker relaxations lead to weaker lower bounds as a weaker relaxation will increase the set of feasible policies over which the infimum is taken in (5). In the case of the perfect information relaxation we have $\mathbb{G} = \mathbb{I}$ and the lower bound takes the form

$$\mathbb{E}_0\left[g(u_F, w)\right] \;\geq\; \inf_{u \in \mathcal{U}_\mathbb{I}} \mathbb{E}_0\left[g(u, w) + z(u, w)\right] \;=\; \mathbb{E}_0\left[\inf_{u \in \mathcal{U}_\mathbb{I}} \{g(u, w) + z(u, w)\}\right].$$

For a given information relaxation, we can optimize the lower, i.e., dual bound, by optimizing over the set of dual-feasible penalties. This leads to the dual of the primal DP:

$$\textbf{Dual Problem:} \qquad \sup_{z \in \mathcal{Z}_\mathbb{F}} \left\{\inf_{u_G \in \mathcal{U}_\mathbb{G}} \mathbb{E}_0\left[g(u_G, w) + z(u_G, w)\right]\right\}. \tag{6}$$

By weak duality, if we identify a policy, u_F, and penalty, z, that are primal and dual feasible, respectively, such that equality in (5) holds, then u_F and z must be optimal for their respective problems. Moreover, if the primal problem (2) has a finite solution, then so too has the dual problem (6), and there is no duality gap. This yields the following result.

Theorem 1 (Strong Duality)
Let $\mathbb{G}$ be a relaxation of $\mathbb{F}$. Then

$$\inf_{u \in \mathcal{U}_{\mathbb{F}}} \mathbb{E}_0 \left[g(u_F, w) \right] \;=\; \sup_{z \in \mathcal{Z}_{\mathbb{F}}} \left\{ \inf_{u_G \in \mathcal{U}_{\mathbb{G}}} \mathbb{E}_0 \left[g(u_G, w) + z(u_G, w) \right] \right\} \tag{7}$$

Furthermore, if the primal problem on the left is bounded, then the dual problem on the right has an optimal solution, $z^* \in \mathcal{Z}_{\mathbb{F}}$, and there is no duality gap.

There is also a version of complementary slackness.

Theorem 2 (Complementary Slackness)
Let u_F^* and z^* be feasible solutions for the primal and dual problems, respectively, with information relaxation $\mathbb{G}$. A necessary and sufficient condition for these to be optimal solutions is that $\mathbb{E}_0 \left[z^*(u_F^*) \right] = 0$

$$\mathbb{E}_0 \left[g(u_F^*, w) + z^*(u_F^*, w) \right] \;=\; \inf_{u_G \in \mathcal{U}_{\mathbb{G}}} \; \mathbb{E}_0 \left[g(u_G, w) + z^*(u_G, w) \right]. \tag{8}$$

Note that Theorem 2 implies that with an optimally chosen penalty, z^*, the decision-maker in the dual DP will be happy to choose a non-anticipative control, despite not being restricted to do so. As shown by Brown et al. (2010), we can also take advantage of any structural information regarding the optimal solution to the primal problem. In particular, if it is known that the optimal solution to the primal problem has a particular structure, then we can restrict ourselves to policies with the same structure when solving the dual optimization problem.

2.2 Using the Dual Formulation to Construct Dual Bounds

In practice it is often the case that we are unable to compute the solution to the primal DP exactly. However, we can compute a lower bound on the optimal value function of the primal DP by starting with a dual feasible penalty function, $z(u, w) := \sum_{k=0}^{N-1} z_k(u, w)$, and then using this penalty function on the right-hand-side of (5). In particular, we do not seek to optimize over the dual penalty but hope that the penalty function we have chosen is sufficiently good so as to result in a small duality gap. We will only consider perfect information relaxations in this paper so that $\mathbb{G} = \mathbb{I}$. This is because the perfect information relaxations result in dual problems that are deterministic optimization problems which are often easy to solve. If we use other information relaxations then the resulting dual problems remain stochastic in which case it is generally difficult to handle constraints on the control vector, u. If we use $J_{db}(x_0; z)$ to denote the resulting dual or lower bound from solving the dual problem then we see that $J_{db}(x_0; z)$ satisfies

$$\begin{aligned} J_{db} &= \inf_{u_G \in \mathcal{U}_{\mathbb{G}}} \mathbb{E}_0 \left[g(u_G, w) + z(u_G, w) \right] & (9) \\ &= \inf_{u_G \in \mathcal{U}_{\mathbb{G}}} \mathbb{E}_0 \left[g_N(x_N) \;+\; \sum_{k=0}^{N-1} \left(g_k(x_k, u_k) + z_k(u_G, w) \right) \right] \\ &= \mathbb{E}_0 \left[\inf_{u_G \in \mathcal{U}_{\mathbb{G}}} \left\{ g_N(x_N) \;+\; \sum_{k=0}^{N-1} \left(g_k(x_k, u_k) + z_k(u_G, w) \right) \right\} \right]. & (10) \end{aligned}$$

The optimization problem inside the expectation in (10) can be solved as a deterministic optimization problem after substituting for the x_k's using (1). An unbiased dual bound on the optimal

value function, $J_0(x_0)$, can therefore be estimated by first simulating M paths of the noise process, w. If we label these paths $w^{(i)} := (w_0^{(i)}, \ldots, w_{N-1}^{(i)})$ for $i = 1, \ldots, M$, and set

$$J_{db}^{(i)}(x_0; z) \;:=\; \inf_{u_G \in \mathcal{U}_{\mathbb{G}}} \left\{ g_N(x_N) + z_N(u, w^{(i)}) \;+\; \sum_{k=0}^{N-1} \left(g_k(x_k, u_k) + z_k(u_G, w^{(i)}) \right) \right\} \tag{11}$$

with the x_k's satisfying (1) then

$$J_{db}(x_0; z) \;:=\; \frac{1}{M} \sum_{i=1}^{M} J_{db}^{(i)}(x_0; z) \tag{12}$$

is an unbiased dual or lower bound for $J_0(x_0)$.

2.3 Constructing Dual Penalties

We outline two methods for constructing dual feasible penalties that we will use in Section 3. We emphasize again that we only consider perfect information relaxations so that $\mathcal{G}_k = \mathcal{I}_k$ for all k.

Using Value Function Approximations to Construct Dual Penalties

Brown et al. (2010) propose taking

$$\begin{aligned} z_k(u, w) \;&:=\; \mathbb{E}_k\left[v_k(u, w)\right] \;-\; \mathbb{E}\left[v_k(u, w) \mid \mathcal{G}_k\right] \\ &=\; \mathbb{E}_k\left[v_k(u, w)\right] \;-\; v_k(u, w) \end{aligned} \tag{13}$$

where $v_k(u, w)$ only depends on $(u_0, \ldots, u_k)$ and where (13) follows since we are using the perfect information relaxation here so that $\mathcal{G}_k = \mathcal{F}$ for all k. It is easy to see that penalties defined in this manner are dual feasible. Indeed we easily obtain that $\mathbb{E}_0\left[z_k(u_F)\right] = 0$ for all $u_F \in \mathcal{U}_{\mathbb{F}}$. Brown et al. (2010) call the $v_k(u)$'s *generating functions* and show that if we take $v_k(u) := J_{k+1}(x_{k+1})$ where J_{k+1} is the optimal value function of the primal DP, then the corresponding penalty,

$$z(u, w) \;=\; \sum_{i=0}^{N-1} \left(\mathbb{E}_k\left[J_{k+1}(x_{k+1})\right] \;-\; J_{k+1}(x_{k+1}) \right), \tag{14}$$

is optimal and results in a zero duality gap. Moreover, they show that $g(u_F^*, w) + z(u_F^*, w) = \mathbb{E}_0\left[g(u_F^*, w)\right]$ *almost surely* with this choice of penalty. This will not be true of the gradient based penalty and the penalty of Davis and Zervos (1994) that we discuss in Section 4.

Note that (14) clearly implies that if we know the optimal value function to within a constant then that is enough to obtain a lower bound with a zero duality gap. More generally, this observation suggests that a good approximation to the shape of the value function should often be sufficient for obtaining a good upper bound. In practice, we do not know J_k and therefore cannot compute the dual penalty of (14). Nonetheless if we have a good approximation, say $\tilde{J}_k$ to J_k then we could use

$$\tilde{z}(u, w) \;:=\; \sum_{i=0}^{N-1} \left(\mathbb{E}_k\left[\tilde{J}_{k+1}(x_{k+1})\right] \;-\; \tilde{J}_{k+1}(x_{k+1}) \right) \tag{15}$$

as a dual feasible penalty and hope to still obtain a good lower bound. This program has been implemented successfully in practice in the context of American options and indeed in the examples of Brown et al. (2010) and Brown and Smith (2010). However, the dual penalty of (15) has a number of weaknesses that can result in (11) being difficult to solve in practice. These weaknesses include:

1. It is not always the case that an approximate value function, $\tilde{J}_k(\cdot)$, is readily available. Even if a good sub-optimal policy is available, it is often the case that the value function corresponding to that sub-optimal policy is unknown. While we could simulate the sub-optimal policy and use the resulting rewards to estimate the value function this would require additional work and we would still have to overcome problems two and three below.

2. Even if we do have $\tilde{J}_{k+1}(\cdot)$ available to us for each k, we may not be able to compute $\mathbb{E}_k\left[\tilde{J}_{k+1}(x_{k+1})\right]$ analytically and so we cannot write $\tilde{z}(u,w)$ in (15) as an analytic function of the u_i's. This in turn makes it very difficult in general to solve the optimization problem in (11).

3. Even when we can compute $\mathbb{E}_k\left[\tilde{J}_{k+1}(x_{k+1})\right]$ analytically, it may be the case that the resulting penalty, $\tilde{z}(u,w)$, causes an otherwise easily-solved deterministic optimization problem to become very difficult. For example, it may be the case that (11) is convex and easy to solve if we assume a zero penalty function but that convexity is lost if we construct $\tilde{z}(u,w)$ according to (14). One possible solution to this problem is to use an approximation to $\tilde{z}(u,w)$ that is linear or otherwise convex in u. This approach has been used successfully by Brown and Smith (2010) for solving portfolio optimization problems with transaction costs but there is no guarantee that it will work in general. In particular, if the desired penalty, $\tilde{z}(u,w)$, is very non-linear in u then the linearization approach may result in poor dual bounds.

These weaknesses are not to suggest that dual bounds based on penalties like (14) cannot work well in practice. Indeed as mentioned earlier, there are several applications where they have been used successfully. However, it would appear that, when they can be applied, dual penalties based on gradients are a promising alternative for several reasons. In particular, they do not require an approximation, $\tilde{J}_k(\cdot)$, to the value function and so the first two problems listed above do not arise. The third problem above also turns out to be a non-issue as gradient-based penalties are linear in the control, u. We now describe these gradient penalty functions.

2.3.1 Constructing Dual Penalties Using Gradients

Brown and Smith (2010) developed a *gradient*-based dual penalty function for perfect information relaxations in the context of dynamic portfolio optimization problems with transaction costs. We will describe their gradient penalty for the more general dynamic program of (3). We define

$$z_g^*(u,w) \quad := \quad \nabla_u g(u^*(w))' \left(u^*(w) - u\right) \tag{16}$$

where $u^* = (u_0^*, \ldots, u_{N-1}^*)$ is the optimal control for the primal dynamic programming problem in (3) and $u = (u_0, \ldots, u_{N-1})$ is an arbitrary control policy. Note that we are therefore implicitly assuming that the total cost, $g(u,w)$, is differentiable in the controls, u. If we view the primal problem in (2) as an optimization problem with the entire *strategy*, u, as the decision variable, then assuming the space of feasible strategies is convex, the first order conditions for optimality are

$$\mathbb{E}_0\left[\nabla_u g(u^*(w))' \left(u^*(w) - u\right)\right] \;\leq\; 0 \tag{17}$$

which implies in particular that $z_g(u,w)$ is dual feasible. Moreover, Brown and Smith (2010) showed[7] that when the cost function is convex the dual feasible penalty in (16) is indeed an optimal

[7]Since Brown and Smith's (2010) primal problem was a maximization problem, they needed to show that their reward function, i.e. utility of terminal wealth, was *concave* in the set of feasible trading strategies.

dual penalty. Note that with this choice of penalty the dual deterministic optimization problem in (11) has the form

$$\inf_{u_G \in \mathcal{U}_{\mathbb{G}}} \left\{ g_N(x_N) \; + \; \sum_{k=0}^{N-1} \left(g_k(x_k, u_k) + \nabla_{u_k} g(u^*)' \left(u_k^*(w) - u_k \right) \right) \right\}. \tag{18}$$

Moreover the gradient penalty is linear in u and this suggests that the dual problem with this penalty should be no harder to solve then the deterministic version of the primal problem.

The difficulty with using (18) of course is that we don't know u^*, the optimal control policy. Indeed if we did know u^* then there would be no problem to solve. Brown and Smith (2010), however, recognized that under certain circumstances they could use $z_g(u, w)$ instead of $z_g^*(u, w)$ as their dual penalty where

$$z_g(u, w) \quad := \quad \nabla_u g(\tilde{u}(w))' \left(\tilde{u}(w) - u \right) \tag{19}$$

and where $\tilde{u}$ is the optimal solution to an alternative *approximate* problem. For example, in their dynamic portfolio optimization problem, Brown and Smith (2010) took $\tilde{u}$ to be the optimal control for the dynamic portfolio optimization problem *without* transactions costs. Because (i) $\tilde{u}$ is optimal for this alternative problem and (ii) the space of feasible controls for the alternative problem contains the space of feasible controls for the original problem they could still conclude

$$\mathbb{E}_0 \left[\nabla_u g(\tilde{u}(w))' \left(\tilde{u}(w) - u \right) \right] \; \leq \; 0 \tag{20}$$

for all $\mathcal{F}_k$-adapted trading strategies so that this alternative gradient penalty is also dual feasible. Moreover, intuition suggests that (19) should be similar to (16) in which case we would expect to obtain good dual bounds using (19). This was indeed the case for the problems and parameter values considered by Brown and Smith (2010).

The advantage of the gradient penalty in (18) over the value-function based penalty in (14) is that it does not require knowledge of the value function and that it is liner in the controls, u. The disadvantage of the gradient approach is that it is not always applicable since it assumes that the cost function is differentiable in u. It also requires the space of feasible strategies to be convex. Finally, even if it is applicable we will see in Section 3 that the optimal gradient penalty, while dual optimal, does not result in a dual objective function that equals the primal objective function almost surely. Equality is only in expectation and this is in contrast to the value-function based penalty in (14) as we discussed earlier.

3 Finite Horizon Linear-Quadratic Control Problems

We now apply the duality ideas of Section 2 to constrained LQ problems. We first review the finite horizon, discrete-time LQ problem as formulated, for example, in Berstekas (2000). We consider the complete information version only as it is well known that that the incomplete information version can be reduced to the complete information case. Let x_k denote the n-dimensional state vector at time k. We assume it has dynamics that satisfy

$$x_{k+1} \; = \; A_k x_k + B_k u_k + w_{k+1}, \qquad k = 0, 1, \ldots, N-1 \tag{21}$$

where u_k is an m-dimensional vector of control variables and the w_k's are n-dimensional independent vectors of zero-mean disturbances with finite second moments. It will be useful later to observe

that (21) implies

$$x_k \;=\; \left(\prod_{i=0}^{k-1} A_i\right) x_0 \;+\; \sum_{i=0}^{k-1} \left(\prod_{j=i+1}^{k-1} A_j\right) (B_i u_i + w_{i+1}) \qquad \text{for} \;\; k = 0, \ldots, N \tag{22}$$

with the understanding that an empty product in (22) equals 1. As before we let $\mathcal{F}_k$ denote the filtration generated by the w_k's. The objective then is to choose $\mathcal{F}_k$-adapted controls, u_k, to minimize

$$\mathrm{E}_{\mathcal{F}_0}\left[x_N' Q_N x_N + \sum_{k=0}^{N-1} \left(x_k' Q_k x_k + u_k' R_k u_k\right)\right]$$

where the Q_k's and R_k's are positive semi-definite and positive definite, respectively. The optimal solution is easily seen[8] to satisfy

$$u_k^*(x_k) = L_k x_k$$

where

$$L_k \;:=\; -\left(B_k' K_{k+1} B_k + R_k\right)^{-1} B_k' K_{k+1} A_k \tag{23}$$

and where the symmetric positive semi-definite matrices, K_k, are given recursively by the algorithm $K_N = Q_N$ and

$$K_k \;:=\; A_k' \left(K_{k+1} - K_{k+1} B_k \left(B_k' K_{k+1} B_k + R_k\right)^{-1} B_k' K_{k+1}\right) A_k + Q_k, \quad k = N-1, \ldots, 0. \tag{24}$$

The optimal value function then satisfies

$$\begin{aligned} J_k(x_k) \;&=\; x_k' K_k x_k \;+\; \sum_{i=k}^{N-1} \mathrm{E}\left[w_{i+1}' K_{i+1} w_{i+1}\right] \\ &=\; x_k' K_k x_k \;+\; \sum_{i=k}^{N-1} \mathrm{trace}\left(K_{i+1} \Sigma_{i+1}\right) \end{aligned} \tag{25}$$

where $\Sigma_i := \mathrm{Cov}\left(w_i\right)$.

3.1 The Dual Penalty Constructed from the Optimal Value Function

We can use the unconstrained optimal value function to construct a dual bound. In particular let $\mathcal{G}_k$ be the perfect information relaxation of $\mathcal{F}_k$. Then we can take $z_k(u_k)$ as our dual penalty where

$$z_k(u_k) \;:=\; \mathrm{E}\left[J_{k+1}(x_{k+1}) \mid \mathcal{F}_k\right] \;-\; \mathrm{E}\left[J_{k+1}(x_{k+1}) \mid \mathcal{G}_k\right] \tag{26}$$

$$\begin{aligned} &=\; \mathrm{E}\left[\left(A_k x_k + B_k u_k + w_{k+1}\right)' K_{k+1} \left(A_k x_k + B_k u_k + w_{k+1}\right) \mid \mathcal{F}_k\right] \;+\; \sum_{i=k+1}^{N-1} \mathrm{trace}\left(K_{i+1} \Sigma_{i+1}\right) \\ &\quad -\; \left(A_k x_k + B_k u_k + w_{k+1}\right)' K_{k+1} \left(A_k x_k + B_k u_k + w_{k+1}\right) \;-\; \sum_{i=k+1}^{N-1} \mathrm{trace}\left(K_{i+1} \Sigma_{i+1}\right) \\ &=\; \left(A_k x_k + B_k u_k\right)' K_{k+1} \left(A_k x_k + B_k u_k\right) \;+\; \mathrm{trace}\left(K_{k+1} \Sigma_{k+1}\right) \\ &\quad -\; \left(A_k x_k + B_k u_k + w_{k+1}\right)' K_{k+1} \left(A_k x_k + B_k u_k + w_{k+1}\right) \\ &=\; -w_{k+1}' K_{k+1} \left(A_k x_k + B_k u_k\right) \;-\; w_{k+1}' K_{k+1} w_{k+1} \;-\; \left(A_k x_k + B_k u_k\right)' K_{k+1} w_{k+1} \\ &\quad +\; \mathrm{trace}\left(K_{k+1} \Sigma_{k+1}\right) \\ &=\; -2 \left(A_k x_k + B_k u_k\right)' K_{k+1} w_{k+1} \;-\; w_{k+1}' K_{k+1} w_{k+1} \;+\; \mathrm{trace}\left(K_{k+1} \Sigma_{k+1}\right). \end{aligned} \tag{27}$$

[8] See, for example, Bertsekas (2000).

We know of course from the results of Brown et al (2008) that (27) is an optimal dual penalty for the unconstrained problem. In fact the dual objective (6) with this choice of penalty is given by

$$J_0(x_0) \;=\; \mathrm{E}_{\mathcal{F}_0}\left[\inf_{u_k}\left(x_N' Q_N x_N \;+\; \sum_{k=0}^{N-1}\left(x_k' Q_k x_k + u_k' R_k u_k + 2\left(A_k x_k + B_k u_k\right)' K_{k+1} w_{k+1}\right)\right)\right] \tag{28}$$

subject to the dynamics of (21). Note that the optimization problem inside the expectation in (28) is a standard *deterministic* LQ problem and is easily solved using the usual techniques. It is easy to confirm by direct computation that the optimal control, u_k^*, in (28) is indeed non-anticipative.

Note also that the last two terms in (27) do not appear in (28) since their sum has expectation zero. While perhaps obvious, this observation emphasizes the non-uniqueness of the optimal dual penalty. Indeed let v_k be any random variable with zero expectation that does not depend on the controls or state variables. Then if $z_k(u_k)$ is an optimal dual penalty so too is $z_k(u_k) + v_k$. Note also that an optimal penalty is as good as any other optimal dual penalty in so far as their dual problems result in equal (and optimal) value functions as well as identifying the optimal non-anticipative control. However, the optimal dual penalty of Brown et al. (2008) is such that any instance of the dual problem is guaranteed to equal the optimal value function *almost surely* and not just in expectation. This is not true in general of other optimal dual penalties and suggests that some (optimal) dual penalties will outperform other optimal dual penalties when Monte-Carlo techniques are required to estimate the outer expectation in (10). Similar observations apply when we cannot compute the optimal dual solution but can only estimate it using sub-optimal penalty functions.

An Alternative Representation for the Value-Function Dual Penalty

Since the dual problem is deterministic, we do not need to explicitly associate $z_k(u)$ in (27) with time period k. In particular, it is the total *sum* of the dual penalties that is relevant and we now determine this sum as a function of the u_k's. This representation of the unconstrained optimal dual penalty will be useful in Section 4. Let P_{vf} denote the total penalty and let $C := \sum_{k=0}^{N-1}(\mathrm{trace}\left(K_{k+1}\Sigma_{k+1}\right) - w_{k+1}' K_{k+1} w_{k+1})$. Note that C has no bearing on the optimal control in any instance of the dual

problem. We see that P_{vf} then satisfies

$$\begin{aligned}
P_{vf} &= C - 2\sum_{k=0}^{N-1} u_k' B_k' K_{k+1} w_{k+1} - 2\sum_{k=0}^{N-1} x_k' A_k' K_{k+1} w_{k+1} \\
&= C - 2\sum_{k=0}^{N-1} u_k' B_k' K_{k+1} w_{k+1} \\
&\quad - 2\sum_{k=0}^{N-1} \left(\left(\prod_{i=0}^{k-1} A_i \right) x_0 + \sum_{i=0}^{k-1} \left(\prod_{j=i+1}^{k-1} A_j \right) (B_i u_i + w_{i+1}) \right)' A_k' K_{k+1} w_{k+1} \\
&= C_{vf} - 2\sum_{k=0}^{N-1} u_k' B_k' K_{k+1} w_{k+1} - 2\sum_{k=0}^{N-1} \left(\sum_{i=0}^{k-1} \left(\prod_{j=i+1}^{k} A_j \right) B_i u_i \right)' K_{k+1} w_{k+1} \\
&= C_{vf} - 2\sum_{k=0}^{N-1} u_k' B_k' K_{k+1} w_{k+1} - 2\sum_{i=0}^{N-2} u_i' B_i' \left(\sum_{k=i+1}^{N-1} \left(\prod_{j=i+1}^{k} A_j' \right) K_{k+1} w_{k+1} \right) \\
&= C_{vf} - 2\sum_{k=0}^{N-1} u_k' B_k' K_{k+1} w_{k+1} - 2\sum_{i=0}^{N-1} u_i' B_i' \left(\sum_{k=i+1}^{N-1} \left(\prod_{j=i+1}^{k} A_j' \right) K_{k+1} w_{k+1} \right) \\
&= C_{vf} - 2\sum_{i=0}^{N-1} u_i' B_i' \left(K_{i+1} w_{i+1} + \left(\sum_{k=i+1}^{N-1} \left(\prod_{j=i+1}^{k} A_j' \right) K_{k+1} w_{k+1} \right) \right) \\
&= C_{vf} - 2\sum_{i=0}^{N-1} u_i' B_i' \left(\sum_{k=i}^{N-1} \left(\prod_{j=i+1}^{k} A_j' \right) K_{k+1} w_{k+1} \right) \qquad (29)
\end{aligned}$$

where[9]

$$C_{vf} := C - 2\sum_{k=0}^{N-1} \left(\left(\prod_{i=0}^{k} A_i \right) x_0 + \sum_{i=0}^{k-1} \left(\prod_{j=i+1}^{k} A_j \right) w_{i+1} \right)' K_{k+1} w_{k+1} \qquad (30)$$

is a mean zero term that does not depend on the u_k's. The salient feature of (29) is that we have an explicit expression for the coefficient of u_i in the optimal dual penalty for the unconstrained LQ problem.

3.2 The Gradient Dual Penalty

The gradient-based optimal dual penalty is also straightforward to calculate. First, we define

$$V_0 := \sum_{i=0}^{N-1} u_i' R_i u_i + \sum_{i=0}^{N} x_i' Q_i x_i$$

which of course is the realized cost for the LQ control problem. We may then define

$$z_g(u) := \nabla_u V_T(u^*)'(u^* - u)$$

where $u^* = (u_0^*, \ldots, u_{N-1}^*)$ is the optimal control for the unconstrained problem and $u = (u_0, \ldots, u_{N-1})$ is an arbitrary control policy. By viewing the LQ problem as a convex optimization problem where

[9] We use the notation C_{vf} to emphasize that this term is the *constant* component of the *value-function* based dual penalty. In particular, C_{vf} does not depend on the u_i's.

the strategy, $u = (u_0, \ldots, u_{N-1})$, is the decision vector[10] we see that the first order optimality conditions are

$$\mathrm{E}_0 \left[\nabla_u V_T(u^*)'(u^* - u) \right] \ \leq \ 0. \tag{31}$$

But (31) then implies that $z_g(u)$ is dual-feasible and indeed it is easy to see that $z_g(u)$ is a dual-optimal penalty for the unconstrained LQ control problem. In this case we know that $u_i^* = L_i x_i^*$ where we use x_i^* to denote the trajectory of the state vector under u^*. We then see that

$$z_g(u) \ = \ \sum_{i=0}^{N-1} \nabla_{u_i} V_T(u^*)'(u_i^* - u_i)$$

where the dynamics in (21) imply

$$\begin{aligned} \nabla_{u_i} V_T(u^*)'(u_i^* - u_i) \ &= \ 2 \left[R_i u_i^* \ + \ B_i' \sum_{k=i+1}^{N} \left(\prod_{j=i+1}^{k-1} A_j' \right) Q_k x_k^* \right]' (u_i^* - u_i) \\ &= \ 2 \left[R_i u_i^* \ + \ B_i' \sum_{k=i+1}^{N} \left(\prod_{j=i+1}^{k-1} A_j' \right) Q_k x_k^* \right]' (L_i x_i^* - u_i). \end{aligned} \tag{32}$$

We can iterate $x_k^* = (A_{k-1} + B_{k-1} L_{k-1}) x_{k-1}^* + w_k$ to obtain

$$x_k^* = \left(\prod_{j=0}^{k-1} (A_j + B_j L_j) \right) x_0 \ + \ \sum_{j=0}^{k-1} \left(\prod_{l=j+1}^{k-1} (A_l + B_l L_l) \right) w_{j+1} \tag{33}$$

and then substitute (33) into (32) to obtain an explicit expression for the gradient penalty that is linear in the u_i's. Before doing this, we have the following lemma which we will use to simplify (32).

Lemma 2 *For $i = 0, \ldots, N$ we have*

$$K_i \ = \ \sum_{j=i}^{N} \left(\prod_{k=i}^{j-1} A_k' \right) \ Q_j \left(\prod_{k=i}^{j-1} (A_k + B_k L_k) \right) \tag{34}$$

where L_k is given by (23).

Proof: First note that when $i = N$ (34) reduces to $K_N = Q_N$ which is clearly true. Suppose now that (34) is true for $i+1$. If we can show that (34) is then true for i we are done. Towards this end note that

$$K_i \ = \ A_i' K_{i+1} \left(A_i + B_i L_i \right) \ + \ Q_i \tag{35}$$

$$= \ A_i' \left[\sum_{j=i+1}^{N} \left(\prod_{k=i+1}^{j-1} A_k' \right) \ Q_j \left(\prod_{k=i+1}^{j-1} (A_k + B_k L_k) \right) \right] \ (A_i + B_i L_i) \ + \ Q_i \tag{36}$$

$$= \ \sum_{j=i}^{N} \left(\prod_{k=i}^{j-1} A_k' \right) \ Q_j \left(\prod_{k=i}^{j-1} (A_k + B_k L_k) \right)$$

[10] To be clear, the decision vector, u, is not an $N \times 1$ vector but is infinite-dimensional as $u_i(x_i)$ is a decision variable for each state x_i and all $i = 0, \ldots, N-1$.

where (35) follows from (23) and (24) and where (36) follows from the assumption that (34) holds for $i+1$. ■

We are now in a position to compare the two penalties. In particular we see that the coefficient of u_i' in each of the two penalties is given by

$$\text{Coeff}_{vf}(u_i) \;=\; -2B_i' \sum_{k=i+1}^{N} \left(\prod_{j=i+1}^{k-1} A_j' \right) K_k w_k \qquad \text{(Value-Function Penalty)} \quad (37)$$

$$\text{Coeff}_{g}(u_i) \;=\; -2R_i L_i x_i^* \;-\; 2B_i' \sum_{k=i+1}^{N} \left(\prod_{j=i+1}^{k-1} A_j' \right) Q_k x_k^* \qquad \text{(Gradient Penalty).} \quad (38)$$

Note that (38) follows from (32) with $L_i x_i^*$ substituted for u_i^* and that (37) follows[11] from (29). The following lemma establishes directly that the two coefficients are identical.

Lemma 3 *$\text{Coeff}_{vf}(u_i) = \text{Coeff}_{g}(u_i)$ for $i = 0, \ldots, N-1$.*

Proof: First note that (33) can be restated more generally as

$$x_k^* = \left(\prod_{j=i}^{k-1} (A_j + B_j L_j) \right) x_i^* \;+\; \sum_{j=i}^{k-1} \left(\prod_{l=j+1}^{k-1} (A_l + B_l L_l) \right) w_{j+1}. \quad (39)$$

We can then use (39) to substitute for x_k^* in (38) to obtain

$$\begin{aligned}
-\frac{1}{2}\,\text{Coeff}_{g}(u_i) \;&=\; \left[R_i L_i \;+\; B_i' \sum_{k=i+1}^{N} \left(\prod_{j=i+1}^{k-1} A_j' \right) Q_k \left(\prod_{j=i}^{k-1} (A_j + B_j L_j) \right) \right] x_i^* \\
&\quad + B_i' \sum_{k=i+1}^{N} \left(\prod_{j=i+1}^{k-1} A_j' \right) Q_k \sum_{j=i}^{k-1} \left(\prod_{l=j+1}^{k-1} (A_l + B_l L_l) \right) w_{j+1} \\
&=\; \left[R_i L_i \;+\; B_i' K_{i+1} (A_i + B_i L_i) \right] x_i^* \qquad \text{by (34)} \\
&\quad + B_i' \sum_{k=i+1}^{N} \sum_{j=i}^{k-1} \left(\prod_{m=i+1}^{k-1} A_m' \right) Q_k \left(\prod_{l=j+1}^{k-1} (A_l + B_l L_l) \right) w_{j+1} \\
&=\; B_i' \sum_{j=i}^{N-1} \left[\sum_{k=j+1}^{N} \left(\prod_{m=i+1}^{k-1} A_m' \right) Q_k \left(\prod_{l=j+1}^{k-1} (A_l + B_l L_l) \right) \right] w_{j+1} \qquad (40) \\
&=\; B_i' \sum_{j=i}^{N-1} \left(\prod_{m=i+1}^{j} A_m' \right) \left[\sum_{k=j+1}^{N} \left(\prod_{m=j+1}^{k-1} A_m' \right) Q_k \left(\prod_{l=j+1}^{k-1} (A_l + B_l L_l) \right) \right] w_{j+1} \\
&=\; B_i' \sum_{j=i}^{N-1} \left(\prod_{m=i+1}^{j} A_m' \right) K_{j+1} w_{j+1} \qquad \text{by (34) again} \qquad (41) \\
&=\; -\frac{1}{2}\,\text{Coeff}_{vf}(u_i)
\end{aligned}$$

[11] We have modified the indexing in (37) so that each summation in (37) and (38) runs from $i+1$ to N.

where (40) follows by changing the order of the double summation and noting that $R_i L_i + B_i' K_{i+1}(A_i + B_i L_i) = 0$ by the definition of L_i in (23). ■

Of course Lemma 3 is not at all surprising since the value-function and gradient penalties are both dual optimal. Indeed the more interesting question is whether or not the constant terms in each of the two penalties are equal. If we use C_g to denote the constant component of the gradient penalty, then using (32) and summing over i we see that is is given by

$$C_g = 2\sum_{i=0}^{N-1}\left[R_i u_i^* + B_i' \sum_{k=i+1}^{N}\left(\prod_{j=i+1}^{k-1} A_j'\right) Q_k x_k^*\right]' L_i x_i^* \tag{42}$$

$$= 2\sum_{i=0}^{N-1}\left[B_i' \sum_{j=i}^{N-1}\left(\prod_{m=i+1}^{j} A_m'\right) K_{j+1} w_{j+1}\right] L_i x_i^*. \tag{43}$$

Note that we have used (41) to substitute for the term inside the square brackets in (42). We recall that the corresponding constant component of the value-function based penalty, C_{vf}, is given by (30). It is clear that C_{vf} and C_g are different. For example, C_{vf} contains the term $\sum_{k=0}^{N-1} \text{trace}\,(K_{k+1}\Sigma_{k+1})$ and no such term appears in C_g. This observation demonstrates in general the the value-function based penalty and the gradient penalty do not coincide when the latter is actually defined.

4 The Davis and Zervos Approach

While we have derived the optimal value-function and gradient penalties in Section 3 using the recent results of Brown et al. (2010) and Brown and Smith (2008), it turns out that these penalties are very closely related to the work of David and Zervos (1995) which we will now describe. Davis and Zervos[12] (1995) consider the three classic cases of discrete-time LQ problems: the deterministic, stochastic full information and stochastic partial information cases. They show that the solution to the deterministic problem can be used to solve the two stochastic versions of the problem after including appropriate Lagrange multiplier terms in the objective function. We will show below that the Lagrange multiplier terms of DZ (1994) in the stochastic full information case[13] also constitute a dual optimal penalty. Indeed, the only difference between the DZ penalty and our two earlier penalties are terms that have zero expectation that do not depend on the u_i's. In particular, DK prove the following[14] theorem.

Theorem 3 *Consider the linear system model*

$$x_{i+1} = Ax_i + Bu_i + w_{i+1} \qquad i = 0, \ldots, N-1 \tag{44}$$

[12] DZ, hereafter.

[13] We will not consider the stochastic partial information case as the ideas are identical and of course, it is well known that the partial information case can be reduced to the full information case by expanding the state space.

[14] This Theorem is a combination of Theorem 2 in DZ together with the analysis they provide in "Case 2" immediately following their Theorem 2. (DZ included a cross-term of the form $x_k' T u_k$ in their objective function but we will omit this term without any loss of generality so that we can compare their penalty with our penalty in (29). They also assume that $A_k = A$, $B_k = B$, $Q_k = Q$ and $R_k = R$ for all k and we will maintain this assumption in this subsection, again without loss of generality.)

where $w = (w_1, \ldots, w_N)$ is a sequence of independent zero mean random vectors and $u = (u_0, \ldots, u_{N-1})$ is the control sequence. Let

$$J(u, \lambda) \;=\; E_{\mathcal{F}_0}\left[x_N' Q x_N + \sum_{i=0}^{N-1}\left(x_i' Q x_i + u_i' R u_i + 2\lambda_i' u_i\right)\right] \tag{45}$$

be the cost associated with the pair (u, λ) and let

$$J_d(u, w, \lambda) \;=\; x_N' Q x_N \;+\; \sum_{i=0}^{N-1}\left(x_i' Q x_i + u_i' R u_i + 2\lambda_i' u_i\right). \tag{46}$$

be the cost associated with (u, w, λ) so that $J(u, \lambda) = E[J_d(u, w, \lambda)]$. Assume the matrices Q and R are symmetric positive semi-definite and symmetric positive definite, respectively, and $\lambda = (\lambda_0, \ldots, \lambda_{N-1})$ is a given sequence of vectors. Suppose λ is chosen so that

$$\lambda_i = -B' K_{i+1} w_{i+1} - B' \beta_{i+1}, \quad \text{for} \;\; i = 0, \ldots, N-1 \tag{47}$$

where K_{i+1} satisfies[15] *(24) and where β_i satisfies*

$$\beta_i = A' \beta_{i+1} + A' K_{i+1} w_{i+1}, \qquad \beta_N = 0. \tag{48}$$

Then: (i) $u_i^*(x_i) = L_i x_i$ where L_i is given by (23) is the optimal non-anticipative control vector that minimizes (45). (ii) $u_i^*(x_i) = L_i x_i$ is also the optimal control that minimizes the deterministic objective function of (46) where w is known in advance. Moreover, this choice of λ is almost surely the unique one for which the minimizer of $J_d(u, w, \lambda)$ is non-anticipative and for which the lagrange multiplier terms in (45) disappear.

In order to compare Theorem 3 with our earlier results, we need to compare the penalty terms in the three approaches. But first note that Davis and Zervos do not[16] include a constant term like C_{vf} or C_g and so we can immediately conclude that the Davis and Zervos penalty is different to the two earlier penalties. The following lemma shows, however, that the coefficient of u_i in the penalty term in (45), i.e. $2\lambda_i$, is equal to $\text{Coeff}_{vf}(u_i)$ as given in (37).

Lemma 4 *For $i = 0, \ldots, N-1$, we have*

$$\lambda_i \;=\; -B' \sum_{k=i}^{N-1}\left(A^{k-i}\right)' K_{k+1} w_{k+1}$$

so that the coefficient of u_i in (29) is equal to the coefficient of u_i in (45). In particular the Lagrangian terms of DZ in (45) are also dual optimal in the framework of BSS.

Proof: First note that we can iterate (48) to obtain

$$\beta_{i+1} \;=\; \sum_{k=i+1}^{N-1}\left(A^{k-i}\right)' K_{k+1} w_{k+1}. \tag{49}$$

[15] With the understanding that $A_i = A$, $B_i = B$, $Q_i = Q$ and $R_i = R$ for all i.

[16] It is possible that they simply omitted such a constant term as it has no bearing on the optimal controls.

We can then substitute (49) into (47) to obtain

$$\begin{aligned}\lambda_i &= -B'\left(K_{i+1}w_{i+1}+\beta_{i+1}\right)\\ &= -B'\left(K_{i+1}w_{i+1}+\sum_{k=i+1}^{N-1}\left(A^{k-i}\right)'K_{k+1}w_{k+1}\right)\\ &= -B'\sum_{k=i}^{N-1}\left(A^{k-i}\right)'K_{k+1}w_{k+1}\end{aligned}$$

as desired. ■

Note that while the Lagrangian terms of DZ are dual optimal in the framework of BSS, they do not result in zero-variance dual bounds. This was also the case with the gradient-based penalty, and as suggested earlier, this observation suggests that penalties based on value-function approximations may be more efficient than other penalties when dual bounds need to be computed using Monte-Carlo methods.

When we consider the results of this section and the earlier developments in the optimal stopping literature that we mentioned in Section 1, it becomes clear then that many of the ideas behind the information relaxation duality theory of Brown et al. (2010) and Rogers (2007) have been around for some time[17] and in particular, since the work of Davis and Karatzas (1994) and Davis and Zervos (1995). This is not to say, however, that Brown et al. (2010) and Rogers (2007) is somehow redundant. On the contrary, they have unified these ideas in a discrete-time framework and demonstrated that the dual problem can be used successfully for evaluating sub-optimal strategies when it is not practically feasible to construct optimal policies. This of course parallels the earlier literature on optimal stopping problems. The results of Brown eta al. also apply to information relaxations that are more general than the perfect-information relaxation. Moreover, their optimal dual penalty in the case of the perfect-information relaxation is a *zero-variance* penalty which should be particularly useful when evaluating sub-optimal strategies via Monte-Carlo.

5 Conclusions and Further Research

There are several directions for future research that are particularly interesting. First, we would like to consider constrained LQ problems and compare the dual bounds corresponding to each of the three penalties. Of course these penalties are only optimal for unconstrained LQ problems and they may not produce good dual bounds when the constraints are frequently binding. When that is the case, it would be necessary to construct other dual-feasible penalties, possibly using good sub-optimal policies for the constrained problem. This has already been done for optimal stopping problems and other problems. See for example, Haugh and Kogan (2004), Rogers (2002) and Brown et al. (2010) among others.

A particularly interesting direction for future research is in comparing the efficiency of value-function based penalties with gradient penalties. We know from Brown et al. (2010) that the former are almost surely optimal when the optimal value function is used. Of course the optimal value function is never available in practice and so *approximate* value functions must be used. The question then arises as to whether penalties constructed using approximate value functions are more efficient or have a lower variance than corresponding gradient penalties. Finally, variance

[17] It should also be mentioned that the idea of relaxing the non-anticipativity constraints has been well known in the stochastic programming literature.

reduction methods should be of considerable use when computing dual bounds. For example, the optimal value function of the unconstrained problem (when it is available analytically) should be a good control variate and indeed such a control variate was used by Brown and Smith (2010). More generally, however, the dual instances of these problems can often be very computationally demanding and constructing good variance reduction methods should be of considerable value.

References

Brown, D.B., J.E. Smith and P. Sun. 2010. Information Relaxations and Duality in Stochastic Dynamic Programs. *Operations Research*, 58(4), p. 785-801.

Brown, D.B., and J.E. Smith. 2010. Dynamic Portfolio Optimization with Transaction Costs: Heuristics and Dual Bounds. Working paper, Duke University.

Bertsekas, D.P. 2000. *Dynamic Programming and Optimal Control: Volume One*. Athena Scientific, Belmont, Massachusetts.

Davis, M.H.A. 1989. Anticipative LQG Control. *IMA J. Math. Contr. Info*, Vol. 6, pp 259-265.

Davis, M.H.A. 1991. Anticipative LQG Control II. *Applied Stochastic Analysis*, pp 205-214.

Davis, M.H.A. and I. Karatzas. 1994. A Deterministic Approach to Optimal Stopping. In *Probability, Statistics and Optimization: A Tribute to Peter Whittle*, F.Kelly ed. pp 455-466. J.Wiley and Sons, New York and Chichester.

Davis, M.H.A., and M. Zervos. 1995. A New Proof of the Discrete-Time LQG Optimal Control Theorems. *IEEE Transactions on Automatic Control*, Vol.40, No.8, pp. 1450-1453.

Haugh, M.B., and L. Kogan. 2004. Pricing American Options: A Duality Approach. *Operations Research*, Vol. 52, No. 2, pp 258-270.

Rogers, L.C.G. 2002. Monte-Carlo Valuation of American Options. *Mathematical Finance*, 12:271-286.

Rogers, L.C.G. 2007. Pathwise Stochastic Optimal Control. *SIAM Journal on Control and Optimization*, Vol. 46, No. 3, pp 1116-1132.