Duality in Robust Dynamic Programming: Pricing Convertibles, Stochastic Games and Control

Shyam S Chandramouli[*]

Abstract

Many decision making problems that arise in Finance, Economics, Inventory etc. can be formulated as Markov Decision Problems (MDPs) and solved using Dynamic Programming techniques. Further, to mitigate the statistical errors in estimating the underlying transition matrix or to exercise optimal control under adverserial setup led to the study of robust formulations of the same problems in Ghaoui and Nilim [11] and Iyengar [8]. In this work, we study the computational methodologies to develop and validate feasible control policies for the Robust Dynamic Programming Problem. In terms of developing control policies, the current work can be seen as generalizing the existing literature on Approximate Dynamic Programming (ADP) to its robust counterpart. The work also generalizes the Information Relaxation and Dual approach of Brown, Smith and Sun [4] to robust multi period problems. While discussing this framework we approach it both from a discrete control perspective and also as a set of conditional continous measures as in Ghaoui and Nilim [11] and Iyengar [8]. We show numerical experiments on applications like ... In a nutshell, we expand the gamut of problems that the dual approach can handle in terms of developing tight bounds on the value function.

1 Introduction

Many Dynamic Optimization problems under uncertainity can be modeled as Markov Decision Process (MDP) and can be solved via Dyanmic Programming methodologies. Usually, in such problems the state space transition probability matrix is known to the decision maker (usually modeled time homogeneous) and the uncertainity is primarily the exact state that the underlying Markov Process would transition to. Hence, decision maker optimizes the expected future reward process w.r.t. to the knowledge of the transition probabilities.

Sometimes the exact characterization of the uncertainity is not possible (in this case the transition probability matrix). This is possible due to the statistical errors in estimation of the transitional probabilities from historical data or presence of an adversary who chooses the worst case alternative at each instant for every control. These problems have been studied in the context of Air traffic control by Nilim and Ghaoui [11], robust optimal stopping in Iyengar [8], repeated

[*] IEOR Department, Columbia University, New York, NY; sc3102@columbia.edu

zero sum games in game theory []. The growing litreature, a few mentioned above characterize settings or applications under which the underlying problem can be modeled as a robust version of a dynamic programming problem.

more on exact DP and how the exact robust DP works

Elghouoi and garud talk about all the stuff they have done. Talk about other litreature.

In this current work, our main focus is not on modeling the given problem as a robust dynamic programming problem. But rather, developing computational methodologies once the Bellman recursion has been set up. Robust Dynamic Programming problem is a more general framework than MDP's but can still be solved by backward induction (with the additional inner problem being solved at each step). Since, many practical problems that can be posit as a robust dp problems are high dimensional in terms of the state space, efficient computations is a bottleneck, which is referred to in the litreature as the 'curse of dimensionality'. Many methods that contend this curse develop methodologies based on an approximation architecture (of basis functions) that obtain near optimal policies or a fair approximation to value function. These methods proposed by carroee, Benjamin van roy and others [] and Longstaff schwartz procedure in american options. later many other such ADP methods continoulsy studied by moallemi, weintraub, bertsekas a lot. The success of all these methods in practice, see bertsekas survey These methods generally obtain a lower bound on the value function as they work with sub optimal policies.

To validate these approximations, the dual bounds which generate upper bounds on the value to go function was initially developed by haugh and kogan, rogers in the context of pricing complex american options. Though, primarily motivated by applications in pricing complex high dimensional derivatives, these dual methods gained prominence in inventory others. swing options paper (Cite) The main ideas in the deriving the dual bounds stem from karatzas and davis. Brown smith sun generalize the ideas of martin haugh to general dynamic programming problems and they give a universal representation of dual bounds and ideal penalty functions for general MDP's. Also, followe the gradient penalties for the same. Their approach gives a systematic approach to deriving penalty martingales and dual bounds for any MDP problem.

We generalize their work to robust dynamic programming problems. Unlike MDP's, a feasible policy to a robust DP may not be a strict lower bound to the original problem. We highlight this fact more in section []. Hence, to validate a feasible policy, we need to construct tight lower and upper bounds (unlike earlier, where dual bounds were needed for upper bound only, as any feasible policy is a lower bound). We construct two dual problems (dual upper and dual lower) for every robust dynamic programming problem. Denoting the given problem as the primal, we derive penalties and conditions similar to that of brown smith and sun [] where these 3 problems are equivalent in the sense of optimal value functions (Strong duality). Also, We construct ideal penalties for both these problems and outline a general methodology validating any given feasible policy.

How does the dual problem work? like penalizing anticipating constraints.

Though, the main goal of our work was a validation technique for the primal problem, the lack of efficient computational methodologies to solve the problem motivated to develop heuristic policies. Garud [] outline value iteration and policy iteration techniques for robust and highlight the study of approximate robust dp techniques. This current work extends the theory of approximate DP to approximate robust DP. The idea of fictious plays are used to construct good heuristic policies to start with. As of now, we hope most of the theory of approximate dp would extend to this setup. Further analysis are on the way.

next on conditional continous measures.

Though this is a theoretically motivated paper, practical applications of these uncertaintites are highlighted in the context of .few papers here........ This paper highlights the applications in this and that as important and numerical simulations are used to establish the validity of bounds as a useful technique.

The paper is organized in the following way: In section 2, we discuss games with discrete controls and the dual approach in modeling the same problem. Section 3, we discuss games with adversarial measures. In section 4, we discuss heuristic policies for the Robust Dynamic Programming problem, Finally, we conclude the paper with some numerical experiments for relevant applications.

2 Model - 1 (Games with discrete controls)

2.1 General Framework

Consider a general finite-horizon discrete-time multi-player (Decision Maker (DM) and the Adversary) control problem with a probability space $(\Omega, \mathcal{F}^{(DM)}, \mathcal{P})$ for the DM and $(\Omega, \mathcal{F}^{(A)}, \mathcal{P})$ for the adversary. Time is indexed by the set $\mathcal{T} := \{0, 1, ...T\}$ and the evolution of DM's information is described by the filtration $\mathcal{F}^{(DM)} = \{\mathcal{F}_0^{(DM)}, ..., \mathcal{F}_T^{(DM)}\}$ with $\mathcal{F}^{(DM)} = \mathcal{F}_T^{(DM)}$. We make the usual assumption that $\mathcal{F}_0^{(DM)} = \{0, \Omega\}$ so that the decision maker starts out with no information regarding the outcome of uncertainity. Defined similarly are the dynamics of the adversary. There is a state vector, $x_t \in R^n$, whose dynamics satisfy

$$x_{t+1} \ = \ f_t(x_t, u_t, v_t), \qquad t = 0, \ldots, T-1 \tag{1}$$

where $u_t \in U_t(x_t) \subseteq \mathcal{R}^m$ is the control taken at time t by the decision maker and $v_t \in V_t(x_t, u_t) \subseteq \mathcal{R}^m$ is the control taken at time t by the adversary after observing the control u_t chosen by the decision maker. A *feasible* strategy for the DM, $u := (u_0, \ldots, u_T)$ is one where each individual action satisfies $u_t \in U_t(x_t)$ for all t. We let $\mathcal{U}$ denote the set of such strategies. A feasible *adapted* strategy, $u := (u_0, \ldots, u_T)$ is one where each individual action satisfies $u_t \in U_t(x_t)$ and where u is $\mathcal{F}_t^{(DM)}$-adapted. We let $\mathcal{U}_{\mathcal{F}^{(DM)}}$ denote the set of all such $\mathcal{F}_t^{(DM)}$-adapted strategies. Similarly, a *feasible* strategy for the adversary, $v := (v_0, \ldots, v_T)$ is one where each individual action satisfies $v_t \in V_t(x_t, u_t)$ for all t. We let $\mathcal{V}$ denote the set of such strategies. A feasible *adapted* strategy, $v := (v_0, \ldots, v_T)$ is one where each individual action satisfies $v_t \in V_t(x_t, u_t)$ and where v is $\mathcal{F}_t^{(A)}$-adapted. We let $\mathcal{V}_{\mathcal{F}^{(A)}}$ denote the set of all such $\mathcal{F}_t^{(A)}$-adapted strategies. Note that, if the set $V_t(.)$ is a singleton, then it becomes the classical dynamic programming problem.

The objective of the DM is to select a feasible adapted strategy, u, to maximize the expected total gain (for every feasible adapted strategy $\tilde{v}$ of the adversary)

$$g(u, \tilde{v}) \ := \ \sum_{t=0}^{T} g_t(x_t, u_t, \tilde{v}_t)$$

where we assume each $g_t(x_t, u_t, \tilde{v}_t)$ is $\mathcal{F}_t^{(DM)}$ measurable. In particular, the decision maker's problem is then given by

$$J_0(x_0, v^*) \quad := \quad \sup_{u \in \mathcal{U}_{\mathcal{F}^{(DM)}}} E\left[\sum_{t=0}^{T} g_t(x_t, u_t, v_t^*)\right] \tag{2}$$

where the expectation in is taken over the set of possible outcomes, $w = (w_1, \ldots, w_T) \in \Omega$.

We model the reward structure as a zero-sum game i.e.$-g_t(.)$ is the reward of the adversary whenever $g_t(.)$ is the reward of the DM. Hence, the objective of the adversary is to select a feasible adapted strategy v, to minimize the expected total gain for every adapted strategy,$\tilde{u}$,of the DM.

$$J_0(x_0, u^*) \quad := \quad \inf_{v \in \mathcal{V}_{\mathcal{F}^{(A)}}} E\left[\sum_{t=0}^{T} g_t(x_t, u_t^*, v_t)\right] \tag{3}$$

where the expectation in is taken over the set of possible outcomes, $w = (w_1, \ldots, w_T) \in \Omega$.

Letting $J^*(t)$ denote the value function of the problem, the described multi-player problem reduces to choosing sequential control in the following Bellman recursion setup,

$$J_t^*(x_t) = \sup_{u_t \in U_t(x_t)} \left\{ \inf_{v_t \in V_t(x_t, u_t)} E_t\left[g_t(x_t, u_t, v_t) + J_{t+1}^*(x_{t+1})\right] \right\} \qquad t = 0, \ldots, T \tag{4}$$

witht the understanding that $J_{T+1}^* = 0$. In practice of course, it is often too difficult or time consuming to perform the iteration. This can occur, for example, if the state vector, x_t, is high-dimensional or if the constraints imposed on the controls are too complex or difficult to handle. We refer to it as the 'curse of dimensionality'. In such circumstances, we must be satisfied with sub-optimal policies.

Hence, the focus of the current analysis is from a point where we have a feasible solution $(\tilde{u}, \tilde{v})$ (which generates a a value-to-go function $\tilde{J}_t$) and we need to validate the quality of the feasible policy i.e. we want to check if the robust value gap between the robust value function, $\tilde{J}_t$ and the optimal robust solution,J_t^*, is minimized in some appropriate sense.

If the adversary's decision problem is solvable at each time step easily, i.e. $\tilde{v}_t = v_t^*(x_t, \tilde{u}_t)$, which essentially means both the adversary and the decision maker can solve the inner problem at each instant efficiently, So, the problem boils down to that of a dynamic programming problem in the following way:

$$J_t^*(x_t) = \sup_{u_t \in U_t(x_t)} E_t\left[g_t(x_t, u_t, v_t^*) + J_{t+1}^*(x_{t+1})\right] \qquad t = 0, \ldots, T \tag{5}$$

where $x_{t+1} = f_t(x_t, u_t, v_t^*)$ and if the chooses decision maker chooses sub-optimal control, $\tilde{u}$, obtains a value $\tilde{J}$. It is clear, $\tilde{J}_t \leq J_t^*$, since, the policy $\tilde{u}_t$ is just an arbitrary feasibly policy for the decision maker's problem and the dual analysis of Brown, smith and sun[] can be used to validate the feasible policy. For the sake of completion, we review their analysis in Appendix A below. But, if the controls $(\tilde{u}_t, \tilde{v}_t)$ are just feasible, we cannot infer immidiately if the value function $\tilde{J}_t$ is smaller or greater than J_t^*. Though the player choosing $\tilde{u}_t$ is playing a suboptimal contol policy, the adversary again is not playing the corresponding worst response control. This is because of his own inability in solving the problem exactly due to curse of dimensionality.

This is the central point of deviation from the dual approaches to the dynamic programming problem. Since, the feasible solution is only sub-optimal does not guarantee a value lower or higher than the optimal robust value function. To validate such policies, we need to have a upper and lower bound on the optimal robust value function. In the event, that these bounds are tight, and the feasible policy generates a value function in the range of the bounds, we can conclude that we are implementing "good" control policies on the system.

2.2 The Dual Approach

For this purpose, we construct *dual_lower* and *dual_upper* which are based on information relaxations. Given a feasible control $(\tilde{u}_t, \tilde{v}_t)$, these duals can be used to construct lower and upper bounds respectively. When these duals are fed with optimal controls (u_t^*, v_t^*), then the value of primal robust DP, *dual_lower* and *dual_upper* are the same. In the usual dynamic programming setting, this *dual_lower* is always equal to the primal DP.

Lemma 1 *(Weak Duality)*
If $(\tilde{u}, \tilde{v})$ are primal feasible and z_{up} and z_{low} are dual upper and dual lower feasible respectively, and $\mathcal{G}^{(DM)}, \mathcal{G}^{(A)}$ are relaxations of $\mathcal{F}^{(DM)}$ and $\mathcal{F}^{(A)}$ respectively, then

$$E[g(\tilde{u}, \tilde{v})] \leq \sup_{u \in \mathcal{U}_{\mathcal{F}^{(DM)}}} E[g(u, \tilde{v})] \leq \sup_{u \in \mathcal{U}_{\mathcal{G}^{(DM)}}} E[g(u, \tilde{v}) - z_{up}(u, \tilde{v})] \tag{6}$$

$$E[g(\tilde{u}, \tilde{v})] \geq \inf_{v \in \mathcal{V}_{\mathcal{F}^{(A)}}} E[g(\tilde{u}, v)] \geq \inf_{v \in \mathcal{V}_{\mathcal{G}^{(A)}}} E[g(\tilde{u}, v) - z_{low}(\tilde{u}, v)] \tag{7}$$

Proof: In Appendix (B)

Theorem 1 *(Strong Duality)*
Let $\mathcal{G}^{(DM)}, \mathcal{G}^{(A)}$ are relaxations of $\mathcal{F}^{(DM)}$ and $\mathcal{F}^{(A)}$ respectively and $(\tilde{u}, \tilde{v})$ defined as above. Then,

$$\begin{aligned} \sup_{u \in \mathcal{U}_{\mathcal{F}}^{(DM)}} E[g(u,\tilde{v})] &= \inf_{z_{up} \in \mathcal{Z}_{\mathcal{F}}^{(DM)}} \left\{ \sup_{u \in \mathcal{U}_{\mathcal{G}^{(DM)}}} E[g(u,\tilde{v}) - z_{up}(u,\tilde{v})] \right\} \\ \inf_{v \in \mathcal{V}_{\mathcal{F}}^{(A)}} E[g(\tilde{u},v)] &= \sup_{z_{low} \in \mathcal{Z}_{\mathcal{F}}^{(A)}} \left\{ \inf_{v \in \mathcal{V}_{\mathcal{G}^{(A)}}} E[g(\tilde{u},v) - z_{low}(\tilde{u},v)] \right\} \end{aligned} \tag{8}$$

In particular, if $(\tilde{u}, \tilde{v})$ are the optimal controls for both the players i.e. (u^*, v^*), then

$$\sup_{z_{low} \in \mathcal{Z}_{\mathcal{F}}^{(A)}} \left\{ \inf_{v \in \mathcal{V}_{\mathcal{G}^{(A)}}} E[g(\tilde{u},v) - z_{low}(\tilde{u},v)] \right\} = E[g(u^*,v^*)] = \inf_{z_{up} \in \mathcal{Z}_{\mathcal{F}}^{(DM)}} \left\{ \sup_{u \in \mathcal{U}_{\mathcal{G}^{(DM)}}} E[g(u,\tilde{v}) - z_{up}(u,\tilde{v})] \right\} \tag{9}$$

If the primal problem is bounded, then both the dual problems are bounded and has an optimal solutions that achieve the bound

Proof: In Appendix (B)

Theorem 2 *(Complementary Slackness)*
Let (u^*, v^*) and (z^*_{up}, z^*_{low}) be feasible solutions for the primal and dual problems respectively with information relaxations $\mathcal{G}^{(DM)}, \mathcal{G}^{(A)}$ of $\mathcal{F}^{(DM)}$ and $\mathcal{F}^{(A)}$ respectively. A necessary and sufficient condition for these to be optimal solutions for their respective problems is that $E[z^*_{up}(u^*, v^*)] = 0$, $E[z^*_{low}(u^*, v^*)] = 0$ and

$$\inf_{v \in \mathcal{V}_{\mathcal{G}^{(A)}}} E[g(u^*,v) - z^*_{low}(u^*,v)] = E[g(u^*,v^*) - z^*_{up}(u^*,v^*)] \quad = \sup_{u \in \mathcal{U}_{\mathcal{G}^{(DM)}}} E[g(u,v^*) - z^*_{up}(u,v^*)] \tag{10}$$

Theorem 3 *(Ideal Penalty)*
(Gradient Penalty)

Proposition 1 *(Structural Policies)*
(Other Properties)

- Find the controls $(\tilde{u}_t, \tilde{v}_t)$ from one of the heuristic policies mentioned above.
- Use $(\tilde{u}_t, \tilde{v}_t)$ to calculate Dual Upper Bound.
- use $(\tilde{u}_t, \tilde{v}_t)$ to calculate Dual Lower Bound.
- Check the quality of the heuristic policy

3 Model - 2 (Games with Uncertain Transition Kernels)

- Formulation
- Talk in the context of special measures as in Garud. (just survey the main ideas). Talk how the infimum could be solved using them

4 Heuristic Policies

4.1 Fictitious plays

4.2 Value iteration/policy iteration

Discuss as in Garud.

4.3 Approximate dynamic programming

- Value iteration/policy iteration
- Temporal difference learning
- ALP
- Pathwise

5 Simulation - Numerical Experiments

- Zero sum games
- multiple stopping zero sum games
- pricing covertibles
- convertibles with multiple fancy
- auctions (stochastic games)
- multi period control

References

[1] D.P. Bertsekas. Dynamic programming and optimal control.

[2] D.P. Bertsekas and J.N. Tsitsiklis. Neuro-dynamic programming: an overview. In *Decision and Control, 1995., Proceedings of the 34th IEEE Conference on*, volume 1, pages 560–564. IEEE, 1995.

[3] D.B. Brown and J.E. Smith. Dynamic portfolio optimization with transaction costs: Heuristics and dual bounds. Technical report, Citeseer, 2010.

[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] S. Chandramouli and M. Haugh. A unified approach to multiple stopping and duality. 2011.

[6] MHA Davis and I. Karatzas. A deterministic approach to optimal stopping. *Probability, Statistics and Optimisation (ed. FP Kelly). NewYork Chichester: John Wiley & Sons Ltd*, pages 455–466, 1994.

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

[8] G.N. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, pages 257–280, 2005.

[9] F.A. Longstaff and E.S. Schwartz. Valuing american options by simulation: A simple least-squares approach. *Review of Financial Studies*, 14(1):113, 2001.

[10] J. Neveu. *Discrete-parameter martingales*, volume 10. Elsevier, 1975.

[11] A. Nilim and L.E. Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, pages 780–798, 2005.

[12] W.B. Powell. *Approximate Dynamic Programming: Solving the curses of dimensionality*, volume 703. Wiley-Blackwell, 2007.

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

[14] J.K. Satia and R.E. Lave Jr. Markovian decision processes with uncertain transition probabilities. *Operations Research*, pages 728–740, 1973.

[15] H. Topaloglu. Computation and dynamic programming. 2010.

[16] J.N. Tsitsiklis and B. Van Roy. Regression methods for pricing complex american-style options. *Neural Networks, IEEE Transactions on*, 12(4):694–703, 2001.

[17] C.C. White III and H.K. Eldeib. Markov decision processes with imprecise transition probabilities. *Operations Research*, pages 739–749, 1994.

6 Appendix B: Proofs

Proof: Weak duality Given feasible controls $\tilde{u}_t = \tilde{u}_t(x_t)$ and $\tilde{v}_t = \tilde{v}_t(x_t, \tilde{u}_t)$

(Dual_Upper):

$$\inf_{v_t \in V_t(x_t,u_t)} E_t \left[g_t(x_t, u_t, v_t) + J^*_{t+1}(x_{t+1}) \right] \leq E_t \left[g_t(x_t, u_t, \tilde{v}_t) + J^*_{t+1}(\tilde{x}_{t+1}) \right] \tag{11}$$

where $\tilde{x}_{t+1} = f_t(x_t, u_t, \tilde{v}_t)$

Now, take $\sup_{u_t \in U_t}$ on both sides, we get,

$$\begin{aligned} J^*_t(x_t) &= \sup_{u_t \in U_t(x_t)} \inf_{v_t \in V_t(x_t,u_t)} E_t \left[g_t(x_t, u_t, v_t) + J^*_{t+1}(x_{t+1}) \right] \\ &\leq \sup_{u_t \in U_t(x_t)} E_t \left[g_t(x_t, u_t, \tilde{v}_t) + J^*_{t+1}(\tilde{x}_{t+1}) \right] \end{aligned} \tag{12}$$

Now, just follow the information relaxation ideas for a usual DP to find a bound on the above quantity. Following, perfect information relaxation, we will have

$$J^*_t(x_t) \leq E_t \left[\sup_{u_t \in U_t(x_t)} g_t(x_t, u_t, \tilde{v}_t) + J^*_{t+1}(\tilde{x}_{t+1}) - \Delta \tilde{J}_{t+1}(\tilde{x}_{t+1}) \right] \tag{13}$$

where $\Delta \tilde{J}_{t+1}(\tilde{x}_{t+1}) = \tilde{J}(\tilde{x}_{t+1}) - E_t[\tilde{J}(\tilde{x}_{t+1})]$ is the penalty discussed in Brown, Smith, Sun.

(Dual_Lower):

Using ideas of perfect information relaxation again, we have,

$$\inf_{v_t \in V_t(x_t,u_t)} E_t \left[g_t(x_t, u_t, v_t) + J^*_{t+1}(x_{t+1}) \right] \geq E_t \left[\inf_{v_t \in V_t(x_t,u_t)} g_t(x_t, u_t, v_t) + J^*_{t+1}(x_{t+1}) - \Delta J_t \right] \tag{14}$$

Taking $\sup_{u_t \in U_t}$ on both sides,

$$\begin{aligned} J^*_t(x_t) &= \sup_{u_t \in U_t(x_t)} \inf_{v_t \in V_t(x_t,u_t)} E_t \left[g_t(x_t, u_t, v_t) + J^*_{t+1}(x_{t+1}) \right] \\ &\geq \sup_{u_t \in U_t(x_t)} E_t \left[\inf_{v_t \in V_t(x_t,u_t)} g_t(x_t, u_t, v_t) + J^*_{t+1}(x_{t+1}) - \Delta J_t \right] \\ &\geq E_t \left[\inf_{v_t \in V_t(x_t,\tilde{u}_t)} g_t(x_t, \tilde{u}_t, v_t) + J^*_{t+1}(x_{t+1}) - \Delta J_t \right] \end{aligned} \tag{15}$$

The last inequality follows because $\tilde{u}_t$ is just an arbitrary policy which need not attain the supremum above.