Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC PapersOnLine 50-1 (2017) 6140–6147

An Efficient Regression Approach to Solving the Dual Problems of Dynamic Programs ⋆

Helin Zhu * **Fan Ye** ** **Enlu Zhou** ***

* *Uber, San Francisco, CA 94103 USA (e-mail: zhuhelin1990@gmail.com).*
** *Morgan Stanley, New York, NY 10019 USA (e-mail: cnfanye@gmail.com)*
*** *Georgia Institute of Technology, Atlanta, GA 30322 USA (e-mail: enlu.zhou@isye.gatech.edu)*

Abstract: In recent years, information relaxation and duality in dynamic programs have been studied extensively, and the resulted primal-dual approach has become a powerful procedure in solving dynamic programs by providing lower-upper bounds on the optimal value function. Theoretically, with the so called value-based optimal dual penalty, the optimal value function could be recovered exactly via strong duality; however, in practice, generating tight dual bounds usually requires good approximations of the optimal dual penalty, which could be time-consuming due to the conditional expectation terms that need to be estimated via nested simulation. In this paper, we will develop an efficient regression approach to approximating the optimal dual penalty in a non-nested manner, by exploring the structure of the feasible dual penalty space. The resulted approximation maintains to be a dual feasible penalty, leading to a valid dual bound on the optimal value function. We show that the proposed approach is computationally efficient, and the resulted dual penalty leads to a numerically tractable dual problem.

Keywords: Information relaxation, dynamic programs, optimal dual penalty, regression, non-nested.

1. INTRODUCTION

Markov decision process (MDP) is commonly used for modeling complex dynamic decision making problems under uncertainties. The objective usually is to find an optimal policy that maximizes the expected accumulated reward (or minimizes the expected accumulated cost) in the long run. In theory, the optimal policy and value function can be found via Bellman dynamic programming; in practice, however, this approach suffers from "the curse of dimensionality". Facing this issue, there is abundant literature that focuses on developing good approximate dynamic programming methods and constructing good suboptimal policies, see Bertsekas (2007), etc. In principle, given a policy, Monte Carlo simulation could be used to evaluate the policy and generate a lower bound estimator on the optimal value function; furthermore, it scales well with the problem dimension. Nevertheless, in lack of the exact optimal value function or its upper bounds, the optimality of the policy is difficult to measure.

The duality theory, developed independently by Rogers (2007) and Brown et al. (2010), addresses this issue by providing upper bounds on the optimal value function. In particular, the authors formulate and solve the dual problems of general dynamic programs (DP). Furthermore, if the duality gap, i,e., the difference between the lower bound induced by a policy and the upper bound, is small enough, then the policy could be claimed to be sufficiently good.

The main idea of duality theory is to relax the non-anticipativity constraint on the feasible policies of a dynamic program, i.e., to allow the decision maker (DM) to choose actions based on the outcomes of future uncertainties, and penalize the DM for the access to future information. Brown et al. (2010) further show that there exist an optimal dual penalty such that the resulted dual problem recovers the optimal value function. That is, strong duality holds. Following this line of research, there have been many new methodologies and applications in recent years. Brown and Smith (2011) propose to use gradient dual penalties for convex DPs, in order to preserve convexity in the dual problem. Ye and Zhou (2015) generalize duality theory to controlled Markov diffusion in a continuous-time setting, etc.

There are also new developments in computational methods that aim at approximating the optimal dual penalty or generating good dual penalties. In general, the optimal dual penalty could not be computed exactly because it involves optimal value functions that are not available and conditional expectation terms that need to be estimated.

⋆ This work was supported by National Science Foundation under Grants CMMI-1413790 and CAREER CMMI-1453934, and Air Force Office of Scientific Research under Grant YIP FA-9550-14-1-0059.

10.1016/j.ifacol.2017.08.2024

A simple alternative is to replace the optimal value functions with approximate ones, and use nested simulation to estimate the conditional expectation terms; however, this approach often requires substantial computational effort. Various methods have been proposed to improve the accuracy and efficiency of the approximation, including the non-nested simulation approaches by Belomestny et al. (2009) and Zhu et al. (2015) in American option pricing, the pathwise optimization techniques by Desai et al. (2011) and Ye and Zhou (2012), etc. However, we note that two key things are missing in most of the existing approaches: 1) the structure of the space of feasible dual penalties, which is referred to as the dual penalty space, is not well-studied; 2) the information such as the suboptimal policy and the sample paths generated in solving the primal problem is not well utilized.

Motivated by the above observation,our goal in this paper is to develop an efficient and tractable approach for approximating the optimal dual penalty. Our approach is based on the analysis on structure of the dual penalty space, which is a well-defined function space. In particular, we find a functional basis of the space and expand the optimal dual penalty with respect to (w.r.t.) the basis. Then we propose a regression approach to estimate the resulted coefficients. The advantages of the proposed approach are three-folds: 1) it circumvents nested simulation, which is usually required in approximating the optimal dual penalty; 2) it incurs minimal extra simulation or computational costs since it reuses the suboptimal policy and samples generated in the primal problem; 3) it yields a feasible dual penalty, leading to a valid upper bound on the optimal value function.

The remainder of the paper is organized as follows. In Section 2, we review the information relaxation duality theory for general DPs. We present the framework of regression approach to approximating the optimal dual penalty in Section 3. In Section 4, we apply the proposed framework to a high-dimensional dynamic trading problem. Conclusions are provided in Section 5.

2. DYNAMIC PROGRAMS AND DUAL FORMULATIONS

On a general probability space $(\Omega, \mathbb{P}, \mathscr{F})$, consider a finite-horizon MDP as follows. Time is indexed by $\mathcal{T} = \{0, 1, ..., N\}$. The state x follows the dynamics

$$x_{n+1} = f(x_n, a_n, z_{n+1}), \quad n = 0, 1, \cdots, N-1, \quad (1)$$

where f is an $\mathscr{F}$-measurable deterministic function, $x_n \in \mathscr{X}_n$ denotes the state, $a_n \in \mathscr{A}_n$ denotes the action, z_{n+1} is the random noise and $\{z_n : n = 1, ..., N\}$ are assumed to be independently and identically distributed (i.i.d.) random variables with probability (law) measure ρ on support space $\Xi \subseteq \mathbb{R}^d$. The evolution of information is described by the natural information filtration $\mathcal{F} = \{\mathscr{F}_n : n = 0, ..., N\}$ with $\mathscr{F}_N = \mathscr{F}$, where $\mathscr{F}_n$ is the σ-algebra consisting of all the measurable events at time n. Loosely speaking, $\mathscr{F}_n$ denotes the information available to the DM at time n. In particular, each z_n is $\mathscr{F}_n$-measurable.

We use a mapping $\alpha_n(\cdot)$ from the state space $\mathscr{X}_n$ to the action space $\mathscr{A}_n$, i.e., $\alpha_n : \mathscr{X}_n \to \mathscr{A}_n$, to denote the decision rule at time n. A policy/strategy $\boldsymbol{\alpha} := (\alpha_0, \alpha_1, ..., \alpha_{N-1})$, which consists of a sequence of decision rules, is called non-anticipative/$\mathcal{F}$-adapted, if each decision rule $\alpha_n(\cdot)$ is $\mathscr{F}_n$-measurable. Intuitively, it means that the DM chooses a_n only based on the information available to him/her, he/she shall not choose a_n based on future information. We use $\mathbb{A}_{\mathcal{F}}$ to denote the set of all non-anticipative policies, and $\mathbb{A}$ to denote the set of all policies (including the anticipative ones); clearly $\mathbb{A}_{\mathcal{F}} \subseteq \mathbb{A}$. Furthermore, we assume there is an $\mathscr{F}_n$-measurable immediate reward at time $n = 0, ..., N-1$, denoted by $r_n(x_n, a_n)$, and an $\mathscr{F}_N$-measurable terminal reward, denoted by $r_N(x_N)$.

Given $x_0 \in \mathscr{X}_0$, the objective of the DM is to select a non-anticipative policy $\boldsymbol{\alpha} \in \mathbb{A}_{\mathcal{F}}$ that maximizes the accumulated reward over all the horizons, i.e.,

$$(P): \ V_0(x_0) \triangleq \sup_{\boldsymbol{\alpha} \in \mathbb{A}_{\mathcal{F}}} \mathbb{E}_0 \left[\sum_{n=0}^{N-1} r_n(x_n, \alpha_n(x_n)) + r_N(x_N) \right], \quad (2)$$

where $V_0(x_0)$ is the optimal value function, and $\mathbb{E}_n[\cdot]$ denotes an expectation taken w.r.t. $\mathscr{F}_n$.

It is well known that problem (2) can be recursively solved in theory via Bellman backward dynamic programming

$$\begin{cases} V_N(x_N) \triangleq r_N(x_N), \\ V_n(x_n) \triangleq \sup_{a_n \in \mathscr{A}_n} \{ r_n(x_n, a_n) + \mathbb{E}_n[V_{n+1}(x_{n+1})] \}. \end{cases}$$

In practice, however, the above Bellman recursion can hardly be solved exactly in most cases, due to the curse of dimensionality. Therefore, we often have to settle with suboptimal policies and lower bounds on the optimal value function $V_0(x_0)$ induced by the suboptimal policies. In the absence of the exact value of $V_0(x_0)$ or a benchmark upper bound on $V_0(x_0)$, the quality of a suboptimal policy could hardly be measured.

Brown et al. (2010) address this issue by solving a dual problem of the (primal) DP (2), and providing an upper bound on $V_0(x_0)$. Therefore, the quality of a suboptimal policy could be empirically measured by examining the duality gap. If the duality gap is sufficiently small, then the suboptimal policy could be claimed to be near-optimal. To be more specific, let us first define a *feasible dual penalty*.

Definition 1. We say $M(\boldsymbol{\alpha}, \mathbf{z})$, a functional of policy $\boldsymbol{\alpha} \in \mathbb{A}$ and noise sequence $\mathbf{z} := (z_1, ..., z_N)$, is a feasible dual penalty if

$$\mathbb{E}_0[M(\boldsymbol{\alpha}, \mathbf{z})] = 0, \quad \forall \boldsymbol{\alpha} \in \mathbb{A}_{\mathcal{F}}.$$

Put in another way, a penalty function $M(\boldsymbol{\alpha}, \mathbf{z})$ is dual feasible if it does not penalize any non-anticipative policy in expectation. We further use $\mathbb{M}_{\mathcal{F}}$ to denote the set of all feasible dual penalties, i.e., $\mathbb{M}_{\mathcal{F}}$ is the dual penalty space.

Remark 2. Definition 1 is slightly different from the one in Brown et al. (2010), in which a dual penalty is called feasible if $\mathbb{E}_0[M(\boldsymbol{\alpha}, \mathbf{z})] \leq 0, \ \forall \boldsymbol{\alpha} \in \mathbb{A}_{\mathcal{F}}$. We use Definition 1 instead of the original definition to ensure that $\mathbb{M}_{\mathcal{F}}$ is a well-defined function space (vector space). Note that using Definition 1 does not exclude any good feasible dual penalties in Brown et al. (2010). Specifically, if $M(\boldsymbol{\alpha}, \mathbf{z})$ is a feasible dual penalty in Brown et al. (2010), i.e., $\mathbb{E}_0[M(\boldsymbol{\alpha}, \mathbf{z})] \leq 0, \ \forall \boldsymbol{\alpha} \in \mathbb{A}_{\mathcal{F}}$. Then $\mathbb{E}_0[(M(\boldsymbol{\alpha}, \mathbf{z}) - \mathbb{E}_0[M(\boldsymbol{\alpha}, \mathbf{z})])] = 0, \ \forall \boldsymbol{\alpha} \in \mathbb{A}_{\mathcal{F}}$. That is, $(M(\boldsymbol{\alpha}, \mathbf{z}) - \mathbb{E}_0[M(\boldsymbol{\alpha}, \mathbf{z})])$ is a feasible dual penalty under Definition 1, and it always induces an upper bound as tight