Robust Decomposable Markov Decision Processes Motivated by Allocating School Budgets

Nedialko B. Dimitrov[a], Stanko Dimitrov[b], Stefanka Chukova[c]

[a] *Operations Research Department*
Naval Postgraduate School, USA
[b] *Management Sciences*
University of Waterloo, Canada
[c] *School of Mathematics, Statistics and Operations Research*
Victoria University of Wellington, New Zealand

Abstract

Motivated by an application to school funding, we introduce the notion of a robust decomposable Markov decision process (MDP). A robust decomposable MDP model applies to situations where several MDPs, with the transition probabilities in each only known through an uncertainty set, are coupled together by joint resource constraints. Robust decomposible MDPs are different than both decomposible MDPs, and robust MDPs and can not be solved by a direct application of the solution methods from either of those areas. In fact, to the best of our knowledge, there is no known method to tractably compute optimal policies in robust, decomposible MDPs. We show how to tractably compute good policies for this model, and apply the derived method to a stylized school funding example.

Keywords: Markov processes, Dynamic Programming-optimal control, School Funding

1. Introduction

Allocation of school funding is critical to improving school performance. Unfortunately, there is no consensus on how this limited resource should be allocated. As such, the allocation of school funding is a recurring topic of political discussion (Shutt, 1979; Fensterwald, 2013; Blume, 2013; Garber, 1997). For example, the state of California passed an initiative in 2012 to raise taxes specifically to fund schools. In 2013, a major debate was how to allocate this schools funding: should it be allocated by population or should poorer schools receive more funding?

Generally, there are two groupings of the scientific literature on school budget allocation. The first grouping proposes formulas that distribute funding based on school characteristics and past performance (King and Mathers, 1997; Miles and Roza, 2006; BenDavid-Hadar and Ziderman, 2011), while the second grouping is empirical in attempting to quantify the impact of increased levels of funding on school performance (Epple and Romano, 2003; Hanushek, 1996; Rubenstein et al., 2007). In this work, we propose

Email addresses: ned@nps.edu (Nedialko B. Dimitrov), sdimitro@uwaterloo.ca (Stanko Dimitrov), stefanka@gmail.com (Stefanka Chukova)

an alternate method of allocating funds to schools, based on control theory, that does not directly fall into either grouping. Our method extends previous work on using Markov decision processes (MDPs) for budget allocation. Motivated by school budget allocation, the main contributions of this paper are 1) to introduce the concept of robust, decomposable MDPs 2) to propose a computationally tractable method of computing good policies in such MDPs and 3) to illustrate, through a stylized example, that funding allocation based on these MDPs outperforms strategies inspired by a current real-world policy, No Child Left Behind.

In the remainder of this section we discuss the existing literature on school funding and improving school performance. We then motivate the robust, decomposable MDP model in words in Section 2. Related control theory work is discussed in Section 3. Sections 4 and 5 mathematically present the model and show how it can tractably produce a funding policy. In Section 6 we compare the funding policy generated from our model to a stylized version of an existing policy, namely No Child Left Behind. We conclude the paper with some discussion in Section 7.

1.1. School Funding: Literature Review

King and Mathers (1997) propose an approach for school improvement using performance-based accountability and financial rewards. The authors consider several performance indicators, including the students achievements—e.g., in English/language or mathematics, students attendance, the dropout and retention rates, etc.—for inclusion in a school performance measure. They provide an extensive review of programs adopted in four US states, in which based on the state identified performance measures, the schools could qualify for reward money.

Miles and Roza (2006) discuss student-weighted resource allocation, which uses student need as the building block of school budgeting. The authors develop a student-weighted index for each school that takes into account the type of students in the school. The student-weighted index is the ratio between two dollar amounts: the actual expenditures at a given school and the expected expenditures, which are computed using district-weighted average expenditures for each type of student. Based on several examples, the study concludes that student-weighted resource allocation provides greater resource equity among schools within districts.

BenDavid-Hadar and Ziderman (2011) propose a new budget allocation formula, designed to achieve an equitable distribution of educational achievement. In addition to needs-based elements, similar to Miles and Roza (2006), the suggested composite allocation formula includes an improvement component, so that the schools receive budgetary allocations based on a new incentive measure. The effect of the new budget allocation formula is demonstrated on data from the Israeli schooling system. Datasets relating students' academic achievement to student background variables, teacher profiles, and school characteristics are analyzed aiming to identify appropriate needs-based formula components. The main approach in this study is the Ordinary Least Squares regression analysis, a method widely used in the literature to measure the relationships between student performance and student background characteristics, see (Gould et al., 2004; Carnoy, 2007; Jenkins et al., 2006).

There are studies, as Epple and Romano (2003), where the focus is on the differences in student ability, household income, and peer groups within the schooling system and the consequences of those differences on educational benefits. Epple and Romano find

that significant heterogeneity in schools and outcomes occurs even under policies that equalize finance and small expenditure variation has small effects. The study suggests that greater funding variation amongst schools may help improve school performance.

A study by Hanushek (1996) concludes that school resource variations are not closely related to variations in student outcomes, *i.e.*, that aggressive spending programs are not always good investment programs unless coupled with other fundamental reforms. In other words, how a budget is utilized appears to be much more important than the size of the budget. This result suggests that the distribution choice of a finite budget across schools in a school district, or districts in a state, may be more important than the size of the overall budget.

Rubenstein et al. (2007) focus on the resource allocation to individual schools within a given school district. They explore resource allocation across schools in large districts based on factors that reflect differential school costs or factors that are related to the distribution of resources. Their empirical work, based on linear regression, shows that schools with higher percentages of poor pupils often receive more money and have more teachers per pupil, but the teachers tend to be less educated and less paid.

In our model, to be presented below, we assume that schools can be assessed either absolutely or relatively to one another. This is not an unreasonable assumption as school children are tested annually, for example in New Jersey (New Jersey Department of Education, Office of Student Achievement and Accountability, 2011). However, school assessment need not be through an annual exam and can be through any method that produces a performance measure for schools. Bessent et al. (1982) propose one such measure and Borhan and Jemain (2012) propose another, this is not an exhaustive list of all school assessment methods.

Bessent et al. (1982) view a school as an enterprise in which the professional staff provide the operating conditions for converting quantifiable resources or inputs into pupil learning outputs. The resources are determined by budgets, teacher assignments, and student assignments; learning outputs are determined by pupil performance on standardized tests. A series of mathematical programs are designed to determine the school efficiency in utilizing the available resources.

Borhan and Jemain (2012) propose an approach to assess school performance in an hierarchical setting by using standardized test results. The approach is based on a belief structure to represent an assessment as a distribution. This method, allows for the assessment and the ranking of schools.

To the best of our knowledge, there are two basic approaches to allocating school funding. The first are policies that are the result of a political process, such as No Child Left Behind. The second are policies suggested academically that distribute funding based on school characteristics and past performance (King and Mathers, 1997; Miles and Roza, 2006; BenDavid-Hadar and Ziderman, 2011). We suggest the application of optimization methods to school budget allocation. The main challenge with a direct application of optimization methods is parameterization. Specifically, a typical optimization-based funding allocation method requires data on the impact of funding decisions. To alleviate the need for data, we create a robust optimization-based funding allocation model. Specifically, the model does not require exact or probabilistic assessments of the impact of funding decisions. Instead, it simply requires a range of potential impacts. We believe that an expert decision maker would be able to generate these ranges or it may be possible to derive these ranges based on historical funding decisions. For example,

New Zealand maintains data on the country-wide school funding and performance on annual exams (New Zealand Ministry of Education, 2013). It may be possible to use this data to statistically derive ranges of impacts resulting from funding decisions. For the purposes of this study, we use a stylized example to compare our funding method to a policy inspired by No Child Left Behind.

2. The Model

In this paper, we use a robust, decomposable, finite horizon MDP to compute funding distribution policies for a school district. While the tools derived are generally applicable to robust, decomposable, finite horizon MDPs, they are motivated by an application to school district funding. In this section, we begin with an English language description of the basic problem and motivation for the model we develop.

Consider a *school district* composed of a set of *schools*. While we use this phrasing, school district could be interpreted in a state-level sense, as a group of smaller districts. In either case, we are concerned with a large institution—which we call the school district—composed of smaller institutions—which we call the schools.

Each school has a *proficiency level*, which, in practice, is determined annually by the school's students' performance on a standardized exam. The school district wants to maximize the proficiency of its schools. The main tool the school district has to improve the proficiency of its schools is distributing a limited amount of funding amongst the schools. Of course, not all schools are the same. Schools have varying numbers of students, and varying capabilities to effectively use allocated funding. The school district makes an annual decision on how to allocate its limited funding amongst the individual schools. Realistically, the school district can only make a funding allocation plan for a decade or so in the future.

Each school transitions into a state of higher or lower proficiency randomly, but the transition probabilities are based on the funding the school receives from the district. Unfortunately, it is impossible for the school district to know these transition probabilities—which depend on the school in question, its current proficiency state, and its level of funding. Even historical data on the school's transitions would not help identify specific transition probabilities, since the situation at an individual school changes quickly, as compared to the annual funding decision cycle. Instead, the school district only has an uncertainty set—a range of possible transition probabilities—for the school.

We formalize the school district's funding allocation problem into a mathematical model as follows. Each individual school can be modeled as an MDP, where the states describe the school's proficiency, the actions describe the school's funding level, and the rewards describe the benefit the school district receives from the school's proficiency level. The rewards for each school capture the school's importance to the district, which could, for example, be proportional to the number of students in the school. The exact transition probabilities for each school's MDP are not known, but instead for each state action pair, we know an uncertainty set describing a set of possible transition probabilities. We term an individual school's MDP as a *little MDP*.

The school district's funding allocation problem is also an MDP. The school district's MDP has states that are the cross product of the states of the individual schools. The individual schools MDPs are coupled by the joint funding allocation decision. The only actions available to the school district are those that satisfy the district's common budget. But, given a distribution of the common funding, the transitions at each of the schools

are independent. The school district has a finite horizon, about a decade, for which to plan funding allocation decisions. We term the school district's MDP as the *big MDP*.

The proposed method answers the basic question: *How should the school district distribute its limited funding to maximize the discounted total proficiency of its schools over its planning horizon?* We model this decision using a robust, decomposable, finite horizon MDP. The term robust comes from the fact that the transition probabilities for each school are not known, but are only known through uncertainty sets. The term decomposable comes from the fact that the big MDP is made of a cross product of little MDPs, coupled only by the limited funding to take actions across them. The term finite horizon comes from the approximately decade long planning period for the district. We formalize the robust, decomposable, finite horizon MDP with mathematical notation in Section 4. In the next section, we discuss some related control theory work.

3. Related Work

Markov decision processes have provided a powerful modeling tool in solving problems related to planning and decision-making under uncertainty, see (Puterman, 2005). However, often in practice an MDP model ends up with very large state and action spaces. The computational difficulties in analyzing MPDs with high dimensionality stimulated research on developing techniques to deal with this problem. Typical approaches to address high dimensionality in MDPs are: function approximation, reachability considerations and aggregation techniques (Meuleau et al., 1998; Powell, 2007).

Nilim and Ghaoui (2005) study the sensitivity of the optimal solutions to Markov decision problems with respect to the state transition probabilities. Typically, in practice, the MDP transition probabilities are unknown and the errors in their estimations often impose limitations in using MPDs as a modeling tool. The authors consider a finite-state, finite-action MDP, and model the uncertainty in the transition matrices by using uncertainty sets. They successfully use a modification of the classical dynamic programming algorithm to solve this robust MDP problem.

Adelman and Mersereau (2008) focus on stochastic dynamic programming problems that are suitable for relaxation via decomposition. The problems they consider consist of a number of subproblems that are independent of each other except for a set of coupling constraints on the action space. They use Lagrangian relaxation and linear programming to approximate the dynamic programming formulation of the problem and provide comparisons between the two relaxations and the optimal solution. Also, they conclude with a useful discussion on how to select the appropriate relaxation.

The model we develop, while it has pieces from both Adelman and Mersereau and Nilim and El Ghaoui, is novel. Specifically, solution methods for our model do not follow directly from either paper. This is demonstrated by the fact that the big MDP is not solvable by standard MDP methods, even using an exponentially sized state space, as the corresponding big MDP of Adelman and Mersereau. Moreover, it can not be solved by the methods of Nilim and El Ghaoui, because the uncertainty sets we address have fundamentally different structure than their robust MDPs. We provide additional discussion on this in Appendix B.

Sisikoglu et al. (2011) develop a learning algorithm for solving a discounted homogeneous MDP with unknown transition probabilities. These transition probabilities are learned either through simulation or direct observation of the system in real time. The authors compare the performance of their algorithm with other traditional learning meth-

ods.

Meuleau et al. (1998) focus on a technique for computing approximately optimal solutions to stochastic resource allocation problems modeled as MDPs with very large state and action spaces. They assume that the problems are composed of multiple tasks whose utilities are independent, and that the actions taken with respect to a task do not influence the status of any other task, *i.e.*, each task can be modeled as an MDP. Overall, the tasks' MDPs are weakly coupled by resource constraints: actions selected for one MDP restrict the actions available to others. The authors propose heuristic techniques for dealing with several classes of constraints that use the solutions for individual MDPs to construct an approximate global solution.

Glazebrook et al. (2011) consider the allocation of a divisible resource, such as manpower, money, or equipment, to a collection of tasks requiring it. The authors develop a notion of indexability for problems of dynamic resource allocation where the resource concerned may be assigned more flexibility than is allowed, for example, in classical multi-armed bandits. In contrast to the previous work listed in the above paragraphs, our model does not assume that we know the transition probabilities, or have access to a simulator for the MDP. Instead, our model assumes we know a range of possible values, the uncertainty set, for the transition probabilities.

We are not the first to use MDPs in constrained budget allocation. MDPs have been used as a modeling tool for solving budget allocation problems in many different applications areas. We review a few of these applications below.

Zhang (2006) considers the budget allocation problem for a building facilities management system. The aim of his optimization model, based on an MDP, is to maximize the overall performance of the building network over the planned time horizon by optimizing the set of annual management actions on all building elements. The optimization is subject to various constraints, such as annual budget and minimum performance requirement for a building or system.

Parvin et al. (2012) develop a model for trading off resources in testing, prevention, and cure of two-stage contagious diseases, such as AIDS and cervical cancer. Under a constrained budget, policymakers are faced with the decision of how to allocate budget for prevention (via vaccinations), subsidized treatment, and examination to detect the presence of initial stages of the contagious disease. Their model, based on an MDP, aims to facilitate this decision-making exercise.

Tirenni et al. (2007) develop a decision-support system that offers a scientific framework for optimal planning and budgeting of targeted marketing campaigns to maximize return on marketing investments. MDPs are used to model customer dynamics and to find optimal marketing policies that maximize the value generated by a customer over a given time horizon. Lifetime value optimization is achieved through dynamic programming algorithms that identify which marketing actions (cross-selling, up-selling, and loyalty marketing campaigns), transition customers to better value and loyalty states. They illustrate their approach with a case study, aiming to optimize marketing planning and budgeting for Finnair's frequent-flyer program. In contrast to the previous work on MDPs in budget allocation, we consider a problem where transition probabilities are known only through an uncertainty set.

4. Mathematical Formulation of the Model

In this section we introduce the mathematical formulation of the model described in Section 2. We formally define the notation and specify a formulation for a robust, decomposable, Markov decision process. While our discussion focuses on a finite horizon problem due to the application of funding allocation in a school district, the results and notation can be extended to an infinite horizon discounted case.

4.1. Notation

For a vector a indexed by elements of a set A, we refer to the coordinate corresponding to $i \in A$ with $a[i]$. We use $\langle \cdot \rangle_{i \in A}$ to mean a sequence, with one element for each element i in the set A.

The set I indexes all the little MDPs, each corresponding to an individual school. In the remainder of the paragraph we define notation for little MDP i, for $i \in I$. The set of possible states is $\mathcal{S}_i$ and the set of actions in state $s \in \mathcal{S}_i$ is $\mathcal{A}_i(s)$. The reward for taking action $a \in \mathcal{A}_i(s)$ in state $s \in \mathcal{S}_i$ is $r_i(s, a)$. There is a cost associated with taking action a in state s in period t denoted by $C_t^i(s, a) \in \mathbb{R}^m$. The cost is a vector over m resources that may be allocated to the little MDP; in the schools example the resources may be funding, man hours, staff, *etc.* The transition probability of moving to state u is $p_i(u \mid s, a)$, given current state and action pair (s, a). The transition probability is unknown, and is defined by the uncertainty set $\mathbb{P}_{s,a}^i \subset \mathbb{R}^{|\mathcal{S}_i|}$. When we consider a numerical example, we will consider a polyhedral set $\mathbb{P}_{s,a}^i$, however we do not put these restrictions in the proofs and exposition below.

The big MDP is composed of a set of little MDPs. Intuitively the only link between the little MDPs is the budget allocation decision, and otherwise they are independent. Formally, the state space of the big MDP, denoted by $\boldsymbol{S} = \bigotimes_{i \in I} \mathcal{S}_i$, is the Cartesian product of the states of the little MDPs. A vector $\boldsymbol{s} \in \boldsymbol{S}$ is indexed by the set I, and we denote the state of little MDP i in the vector $\boldsymbol{s}$ as $\boldsymbol{s}[i]$. There is a finite budget, $\boldsymbol{b}_t \in \mathbb{R}^m$, for planning period t. The budget is a vector over m resources that may be allocated to each little MDP. The set of actions for state $\boldsymbol{s}$ in period t is:

$$\bar{\boldsymbol{A}}_t(\boldsymbol{s}) = \left\{ \boldsymbol{a} \in \bigotimes_{i \in I} \mathcal{A}_i(\boldsymbol{s}[i]) \mid \boldsymbol{C}_t(\boldsymbol{s}, \boldsymbol{a}) \leq \boldsymbol{b}_t \right\},$$

where $\boldsymbol{C}_t(\boldsymbol{s}, \boldsymbol{a}) = \sum_{i \in I} C_t^i(\boldsymbol{s}[i], \boldsymbol{a}[i])$. Intuitively the only actions available in the big MDP are those within budget. We also define an unrestricted action set as:

$$\boldsymbol{A}(\boldsymbol{s}) = \left\{ \boldsymbol{a} \in \bigotimes_{i \in I} \mathcal{A}_i(\boldsymbol{s}[i]) \right\}.$$

The unrestricted action set may include actions that are budget infeasible in some states. We define a reward for each state action pair, $\boldsymbol{s} \in \boldsymbol{S}, \boldsymbol{a} \in \boldsymbol{A}(\boldsymbol{s})$, as $\boldsymbol{r}(\boldsymbol{s}, \boldsymbol{a}) = \sum_{i \in I} r_i(\boldsymbol{s}[i], \boldsymbol{a}[i])$. Intuitively, the reward is additive across little MDPs.

The set of possible transition probabilities for the big MDP in state $\boldsymbol{s} \in \boldsymbol{S}$ given

action $\boldsymbol{a} \in \boldsymbol{A}(\boldsymbol{s})$ is defined as $\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}$. Formally, we have:

$$\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}} = \left\{ \boldsymbol{p} \in \mathbb{R}^{|\boldsymbol{S}|} \mid \boldsymbol{p} = \langle \prod_{i \in I} p^i[\boldsymbol{u}[i]] \rangle_{\boldsymbol{u} \in \boldsymbol{S}} \text{ where } p^i \in \mathbb{P}^i_{\boldsymbol{s}[i],\boldsymbol{a}[i]} \text{ for all } i \in I \right\}.$$

Intuitively, an element in $\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}$ is a vector in $\mathbb{R}^{|\boldsymbol{S}|}$ that gives the probabilities of transitioning to each state $\boldsymbol{u} \in \boldsymbol{S}$. If little MDP i's transition probabilities are p^i, then the probability to transition to state $\boldsymbol{u} \in \boldsymbol{S}$ is the product of the probabilities that each little MDP transitions to state $\boldsymbol{u}[i]$, $\prod_{i \in I} p^i[\boldsymbol{u}[i]]$. This structure is related to the rectangularity assumption of Nilim and Ghaoui (2005). In other words, given a $\boldsymbol{s}$ and $\boldsymbol{a}$ the little MDPs transition independently.

4.2. Solving the big MDP

Solving the robust big MDP—computing the value for each big MDP state using some variant of back-propagation—involves solving complex non-linear, non-convex optimization problems to find the worst-case transition probabilities. To our knowledge no tractable algorithm exists to solve such problems exactly. We discuss a direct solution approach further in Appendix B. Therefore, in this section, we take an alternate solution approach based on Lagrangian relaxation.

We build an approximation for the value function of the big MDP, and discuss using that approximation to compute policies. The robust MDP Bellman recursion (Nilim and Ghaoui, 2005) for the big MDP in state $\boldsymbol{s}$ at time period $t-1$ is:

$$\boldsymbol{V}_{t-1}(\boldsymbol{s}) = \max_{\boldsymbol{a} \in \bar{\boldsymbol{A}}_t(\boldsymbol{s})} \left[\boldsymbol{r}(\boldsymbol{s},\boldsymbol{a}) + \beta \min_{\boldsymbol{p} \in \mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}} \sum_{\boldsymbol{u} \in \boldsymbol{S}} \boldsymbol{p}[\boldsymbol{u}] \boldsymbol{V}_t(\boldsymbol{u}) \right],$$

where $\boldsymbol{V}_{t-1}(\boldsymbol{s})$ is the value of state $\boldsymbol{s}$ in time period $t-1$. We also assume that at the end of the planning horizon $t = T$ the value function is given and decomposes, $\boldsymbol{V}_T(\boldsymbol{u}) = \sum_{i \in I} V^i_T(\boldsymbol{u}[i])$, where $V^i_T(\boldsymbol{u}[i])$ is the value of the little MDP i in state $\boldsymbol{u}[i]$ in period T. The robust Bellman recursion differs from the standard Bellman recursion (Puterman, 2005) because instead of taking the expectation with respect to known transition probabilities, it takes the expectation with respect to worst case transition probabilities within the uncertainty set, $\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}$, defined for each state action pair. The correctness of the robust Bellman recursion is established by Nilim and Ghaoui (2005). The above recursion uses the restricted action set $\bar{\boldsymbol{A}}_t(\boldsymbol{s})$. If instead the recursion used the unrestricted action set, $\boldsymbol{A}(\boldsymbol{s})$, and the value function $\boldsymbol{V}_t(\boldsymbol{s})$ split into a sum of functions, one for each state of each little MDP, then the recursion would decompose into $|I|$ separate recursions, one for each little MDP. Such a decomposition is carried out by Adelman and Mersereau (2008) for weakly coupled, but not robust, MDPs. The advantage of the decomposition is an exponential reduction in the number of value function values that require computation. For example, if the big MDP consists of 10 little MDPs, each with 5 states, then prior to decomposition we require 5^{10} value function computations, while with decomposition we require only $5 \cdot 10 = 50$ value function computations.

To facilitate decomposition, we perform a Lagrangian relaxation on the coupling budget constraints, $\boldsymbol{C}_t(\boldsymbol{s},\boldsymbol{a}) \leq \boldsymbol{b}_t$, of the big MDP. An alternate, approximate dynamic programming (ADP), approach is discussed in Appendix A. We consider the following

robust Lagrangian relaxation of the robust Bellman recursion above:

$$\boldsymbol{V}_{t-1}^{\lambda_{t-1}}(\boldsymbol{s}) = \max_{\boldsymbol{a}\in\boldsymbol{A}(\boldsymbol{s})} \left[\boldsymbol{r}(\boldsymbol{s},\boldsymbol{a}) + \boldsymbol{\lambda_{t-1}} \cdot (\boldsymbol{b}_{t-1} - \boldsymbol{C}_{t-1}(\boldsymbol{s},\boldsymbol{a})) + \beta \min_{\boldsymbol{p}\in\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}} \sum_{\boldsymbol{u}\in\boldsymbol{S}} \boldsymbol{p}[\boldsymbol{u}] \boldsymbol{V}_t^{\lambda_t}(\boldsymbol{u}) \right]. \tag{1}$$

Intuitively, we now consider unrestricted action set, $\boldsymbol{A}(\boldsymbol{s})$, but penalize actions that violate the budget constraints using the Lagrange multiplier vector, $\boldsymbol{\lambda_t}$. We still have to show that the value function, $\boldsymbol{V}_t^{\lambda_t}(\boldsymbol{s})$, decomposes into a sum of value functions, one for each little MDP, and that the Lagrangian relaxation gives an upper bound on the original value function, $\boldsymbol{V}_t(\boldsymbol{s})$.

Lemma 4.1 shows how to decompose (1) into separate optimization problems over the little MDPs. To facilitate this lemma, the assumption that the big MDP value function is decomposable at $t = T$ is key. We also set $\boldsymbol{V}_T^{\lambda_T}(\boldsymbol{s}) = \boldsymbol{V}_T(\boldsymbol{u})$ and $V_T^{\lambda_T,i}(s) = V_T^i(s)\,\forall i \in I$, setting the value functions and Lagrangian approximation equal at the final period, T.

Lemma 4.1.

$$\boldsymbol{V}_t^{\lambda_t}(\boldsymbol{s}) = \sum_{i\in I} V_t^{\lambda_t,i}(\boldsymbol{s}[i])$$

for t from 1 to T, where

$$V_{t-1}^{\lambda_{t-1},i}(s) = \frac{\boldsymbol{\lambda_{t-1}} \cdot \boldsymbol{b}_{t-1}}{|I|} + \max_{a\in\mathcal{A}_i(s)} \left[r_i(s,a) - \boldsymbol{\lambda_{t-1}} \cdot C_{t-1}^i(s,a) + \beta \min_{p^i\in\mathbb{P}_{s,a}^i} \sum_{u\in\mathcal{S}_i} p^i[u] V_t^{\lambda_t,i}(u) \right]$$

Proof. We proceed by induction on t. The claim is true for $t = T$ by the assumption on the form of $\boldsymbol{V}_T(\boldsymbol{u})$. We assume that the claim is true for t and prove it for $t-1$. We start with the definition of the robust Lagrangian Bellman recursion:

$$\boldsymbol{V}_{t-1}^{\lambda_{t-1}}(\boldsymbol{s}) = \max_{\boldsymbol{a}\in\boldsymbol{A}(\boldsymbol{s})} \left[\boldsymbol{r}(\boldsymbol{s},\boldsymbol{a}) + \boldsymbol{\lambda_{t-1}} \cdot (\boldsymbol{b}_{t-1} - \boldsymbol{C}_{t-1}(\boldsymbol{s},\boldsymbol{a})) + \beta \min_{\boldsymbol{p}\in\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}} \sum_{\boldsymbol{u}\in\boldsymbol{S}} \boldsymbol{p}[\boldsymbol{u}] \boldsymbol{V}_t^{\lambda_t}(\boldsymbol{u}) \right].$$

Substituting the definitions of $\boldsymbol{r}(\boldsymbol{s},\boldsymbol{a})$ and $\boldsymbol{C}_{t-1}(\boldsymbol{s},\boldsymbol{a})$, and distributing and moving summations we have:

$$\begin{aligned}\boldsymbol{V}_{t-1}^{\lambda_{t-1}}(\boldsymbol{s}) = \max_{\boldsymbol{a}\in\boldsymbol{A}(\boldsymbol{s})} \Bigg[& \sum_{i\in I} r_i(\boldsymbol{s}[i],\boldsymbol{a}[i]) \\ & + \sum_{i\in I} \left(\frac{\boldsymbol{\lambda_{t-1}} \cdot \boldsymbol{b}_{t-1}}{|I|} - \boldsymbol{\lambda_{t-1}} \cdot C_{t-1}^i(\boldsymbol{s}[i],\boldsymbol{a}[i]) \right) + \beta \min_{\boldsymbol{p}\in\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}} \sum_{\boldsymbol{u}\in\boldsymbol{S}} \boldsymbol{p}[\boldsymbol{u}] \boldsymbol{V}_t^{\lambda_t}(\boldsymbol{u}) \Bigg].\end{aligned}$$

Applying the induction hypothesis and the definition of vectors in $\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}$ gives:

$$\begin{aligned}\boldsymbol{V}_{t-1}^{\lambda_{t-1}}(\boldsymbol{s}) = \max_{\boldsymbol{a}\in\boldsymbol{A}(\boldsymbol{s})} \Bigg[& \sum_{i\in I} r_i(\boldsymbol{s}[i],\boldsymbol{a}[i]) + \sum_{i\in I} \Bigg(\frac{\boldsymbol{\lambda_{t-1}} \cdot \boldsymbol{b}_{t-1}}{|I|} - \boldsymbol{\lambda_{t-1}} \\ & \cdot C_{t-1}^i(\boldsymbol{s}[i],\boldsymbol{a}[i]) \Bigg) + \beta \min_{\boldsymbol{p}\in\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}} \sum_{\boldsymbol{u}\in\boldsymbol{S}} \left(\prod_{i\in I} p^i[\boldsymbol{u}[i]] \right) \left(\sum_{i\in I} V_t^{\lambda_t,i}(\boldsymbol{u}[i]) \right) \Bigg].\end{aligned}$$

By reordering the summation and product within the minimization gives:

$$\boldsymbol{V}_{t-1}^{\lambda_{t-1}}(\boldsymbol{s}) = \max_{\boldsymbol{a}\in\boldsymbol{A}(\boldsymbol{s})} \left[\sum_{i\in I} r_i(\boldsymbol{s}[i],\boldsymbol{a}[i]) + \sum_{i\in I} \left(\frac{\boldsymbol{\lambda_{t-1}}\cdot\boldsymbol{b}_{t-1}}{|I|} - \boldsymbol{\lambda_{t-1}} \right.\right.$$
$$\left.\left. \cdot\, C_{t-1}^{i}(\boldsymbol{s}[i],\boldsymbol{a}[i]) \right) + \beta \min_{\boldsymbol{p}\in\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}} \left(\sum_{i\in I} \sum_{\boldsymbol{u}\in\boldsymbol{S}} V_t^{\lambda_t,i}(\boldsymbol{u}[i]) \prod_{j\in I} p^j[\boldsymbol{u}[j]] \right) \right].$$

We define $\boldsymbol{S}_{-i} = \bigotimes_{j\in I-i} \mathcal{S}_j$, implying that $\boldsymbol{S} = \mathcal{S}_i \bigotimes \boldsymbol{S}_{-i}$, and reorder summations further to get

$$\boldsymbol{V}_{t-1}^{\lambda_{t-1}}(\boldsymbol{s}) = \max_{\boldsymbol{a}\in\boldsymbol{A}(\boldsymbol{s})} \left[\sum_{i\in I} r_i(\boldsymbol{s}[i],\boldsymbol{a}[i]) + \sum_{i\in I} \left(\frac{\boldsymbol{\lambda_{t-1}}\cdot\boldsymbol{b}_{t-1}}{|I|} - \boldsymbol{\lambda_{t-1}}\cdot C_{t-1}^{i}(\boldsymbol{s}[i], \right.\right.$$
$$\left.\left. \boldsymbol{a}[i]) \right) + \beta \min_{\boldsymbol{p}\in\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}} \left(\sum_{i\in I} \sum_{s\in\mathcal{S}_i} V_t^{\lambda_t,i}(s) p^i[s] \sum_{\boldsymbol{w}\in\boldsymbol{S}_{-i}} \prod_{j\in I-i} p^j[\boldsymbol{w}[j]] \right) \right].$$

We use the fact that $1 = \sum_{\boldsymbol{w}\in\boldsymbol{S}_{-i}} \prod_{j\in I-i} p^j[\boldsymbol{w}[j]]$ to get

$$\boldsymbol{V}_{t-1}^{\lambda_{t-1}}(\boldsymbol{s}) = \max_{\boldsymbol{a}\in\boldsymbol{A}(\boldsymbol{s})} \left[\sum_{i\in I} r_i(\boldsymbol{s}[i],\boldsymbol{a}[i]) + \sum_{i\in I} \left(\frac{\boldsymbol{\lambda_{t-1}}\cdot\boldsymbol{b}_{t-1}}{|I|} \right.\right.$$
$$\left.\left. - \boldsymbol{\lambda_{t-1}}\cdot C_{t-1}^{i}(\boldsymbol{s}[i],\boldsymbol{a}[i]) \right) + \beta \min_{\boldsymbol{p}\in\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}} \left(\sum_{i\in I} \sum_{s\in\mathcal{S}_i} V_t^{\lambda_t,i}(s) p^i[s] \right) \right].$$

From the definition of $\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}$ we have

$$\boldsymbol{V}_{t-1}^{\lambda_{t-1}}(\boldsymbol{s}) = \max_{\boldsymbol{a}\in\boldsymbol{A}(\boldsymbol{s})} \left[\sum_{i\in I} r_i(\boldsymbol{s}[i],\boldsymbol{a}[i]) + \sum_{i\in I} \left(\frac{\boldsymbol{\lambda_{t-1}}\cdot\boldsymbol{b}_{t-1}}{|I|} \right.\right.$$
$$\left.\left. - \boldsymbol{\lambda_{t-1}}\cdot C_{t-1}^{i}(\boldsymbol{s}[i],\boldsymbol{a}[i]) \right) + \beta \min_{p^i\in\mathbb{P}^i_{\boldsymbol{s}[i],\boldsymbol{a}[i]},\forall i\in I} \left(\sum_{i\in I} \sum_{s\in\mathcal{S}_i} V_t^{\lambda_t,i}(s) p^i[s] \right) \right].$$

The minimization problem inside the recursion has an objective function that separates into one term for each little MDP $i \in I$. In addition, the definition of $\boldsymbol{p} \in \mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}$ allows for an independent choice for each $p^i \in \mathbb{P}^i_{\boldsymbol{s}[i],\boldsymbol{a}[i]}$. These facts imply that the inner minimization problem reduces to a sum of minimization problems and we have

$$\boldsymbol{V}_{t-1}^{\lambda_{t-1}}(\boldsymbol{s}) = \max_{\boldsymbol{a}\in\boldsymbol{A}(\boldsymbol{s})} \left[\sum_{i\in I} r_i(\boldsymbol{s}[i],\boldsymbol{a}[i]) + \sum_{i\in I} \left(\frac{\boldsymbol{\lambda_{t-1}}\cdot\boldsymbol{b}_{t-1}}{|I|} \right.\right.$$
$$\left.\left. - \boldsymbol{\lambda_{t-1}}\cdot C_{t-1}^{i}(\boldsymbol{s}[i],\boldsymbol{a}[i]) \right) + \sum_{i\in I} \beta \min_{p^i\in\mathbb{P}^i_{\boldsymbol{s}[i],\boldsymbol{a}[i]}} \left(\sum_{s\in\mathcal{S}_i} V_t^{\lambda_t,i}(s) p^i[s] \right) \right].$$

Now the outer maximization problem has an objective that is a sum with one term for each little MDP, and it is over an unrestricted action set, giving independent choice of action in each little MDP. Thus the outer maximization problem reduces to a sum of maximization problems and we have

$$\boldsymbol{V}_{t-1}^{\lambda_{t-1}}(\boldsymbol{s}) = \sum_{i \in I} \max_{a \in \mathcal{A}_i(\boldsymbol{s}[i])} \Bigg[r_i(\boldsymbol{s}[i], \boldsymbol{a}[i]) + \left(\frac{\boldsymbol{\lambda_{t-1}} \cdot \boldsymbol{b}_{t-1}}{|I|} - \boldsymbol{\lambda_{t-1}} \cdot C_{t-1}^i(\boldsymbol{s}[i], \boldsymbol{a}[i]) \right) + \beta \min_{p^i \in \mathbb{P}^i_{\boldsymbol{s}[i], \boldsymbol{a}[i]}} \left(\sum_{s \in \mathcal{S}_i} V_t^{\lambda_t, i}(s) p^i[s] \right) \Bigg].$$

Taking out constant terms from the outer maximization we have

$$\boldsymbol{V}_{t-1}^{\lambda_{t-1}}(\boldsymbol{s}) = \sum_{i \in I} \left\{ \frac{\boldsymbol{\lambda_{t-1}} \cdot \boldsymbol{b}_{t-1}}{|I|} + \max_{a \in \mathcal{A}_i(\boldsymbol{s}[i])} \Bigg[r_i(\boldsymbol{s}[i], \boldsymbol{a}[i]) - \boldsymbol{\lambda_{t-1}} \cdot C_{t-1}^i(\boldsymbol{s}[i], \boldsymbol{a}[i]) + \beta \min_{p^i \in \mathbb{P}^i_{\boldsymbol{s}[i], \boldsymbol{a}[i]}} \left(\sum_{s \in \mathcal{S}_i} V_t^{\lambda_t, i}(s) p^i[s] \right) \Bigg] \right\}.$$

Substituting the definition of $V_{t-1}^{\lambda_{t-1}, i}(\boldsymbol{s}[i])$, we have the statement of the lemma

$$\boldsymbol{V}_{t-1}^{\lambda_{t-1}}(\boldsymbol{s}) = \sum_{i \in I} V_{t-1}^{\lambda_{t-1}, i}(\boldsymbol{s}[i]).$$

□

Lemma 4.2 shows that $\boldsymbol{V}_t^{\lambda_t}(\boldsymbol{s}) \geq \boldsymbol{V}_t(\boldsymbol{s})$ for all $\boldsymbol{s} \in \boldsymbol{S}$, $t = 1, \ldots, T$, showing that the Lagrangian relaxation does indeed provide an upper bound on the robust MDP value function.

Lemma 4.2. *For any $\boldsymbol{\lambda_t} \geq 0$, $\boldsymbol{V}_t^{\lambda_t}(\boldsymbol{s}) \geq \boldsymbol{V}_t(\boldsymbol{s})$ for all $\boldsymbol{s} \in \boldsymbol{S}$, $t = 1, \ldots, T$.*

Proof. We prove the statement by induction on t. The lemma statement is true for $t = T$, because $\boldsymbol{V}_T^{\lambda_t}(\boldsymbol{s}) = \boldsymbol{V}_T(\boldsymbol{s})$ in the final period.

We assume the statement for t and prove it for $t - 1$. For all $\boldsymbol{a} \in \bar{\boldsymbol{A}}_{t-1}(\boldsymbol{s})$, by definition, $\boldsymbol{b}_{t-1} - \boldsymbol{C}_{t-1}(\boldsymbol{s}, \boldsymbol{a}) \geq 0$ must hold. Because $\boldsymbol{\lambda_{t-1}} \geq 0$ we have:

$$\boldsymbol{V}_{t-1}(\boldsymbol{s}) \leq \max_{\boldsymbol{a} \in \bar{\boldsymbol{A}}_{t-1}(\boldsymbol{s})} \left[\boldsymbol{r}(\boldsymbol{s}, \boldsymbol{a}) + \boldsymbol{\lambda_{t-1}} \cdot (\boldsymbol{b}_{t-1} - \boldsymbol{C}_{t-1}(\boldsymbol{s}, \boldsymbol{a})) + \beta \min_{\boldsymbol{p} \in \mathbb{P}_{\boldsymbol{s}, \boldsymbol{a}}} \sum_{\boldsymbol{u} \in \boldsymbol{S}} \boldsymbol{p}[\boldsymbol{u}] \boldsymbol{V}_t(\boldsymbol{u}) \right].$$

Because $\bar{\boldsymbol{A}}_{t-1}(\boldsymbol{s}) \subseteq \boldsymbol{A}(\boldsymbol{s})$,

$$\boldsymbol{V}_{t-1}(\boldsymbol{s}) \leq \max_{\boldsymbol{a} \in \boldsymbol{A}(\boldsymbol{s})} \left[\boldsymbol{r}(\boldsymbol{s}, \boldsymbol{a}) + \boldsymbol{\lambda_{t-1}} \cdot (\boldsymbol{b}_{t-1} - \boldsymbol{C}_{t-1}(\boldsymbol{s}, \boldsymbol{a})) + \beta \min_{\boldsymbol{p} \in \mathbb{P}_{\boldsymbol{s}, \boldsymbol{a}}} \sum_{\boldsymbol{u} \in \boldsymbol{S}} \boldsymbol{p}[\boldsymbol{u}] \boldsymbol{V}_t(\boldsymbol{u}) \right].$$

By the induction hypothesis and the definition of $\boldsymbol{V}_t^{\lambda_t}(\boldsymbol{u})$,

$$\boldsymbol{V}_{t-1}(\boldsymbol{s}) \leq \max_{\boldsymbol{a} \in \boldsymbol{A}(\boldsymbol{s})} \left[\boldsymbol{r}(\boldsymbol{s}, \boldsymbol{a}) + \boldsymbol{\lambda_{t-1}} \cdot (\boldsymbol{b}_{t-1} - \boldsymbol{C}_{t-1}(\boldsymbol{s}, \boldsymbol{a})) + \beta \min_{\boldsymbol{p} \in \mathbb{P}_{\boldsymbol{s}, \boldsymbol{a}}} \sum_{\boldsymbol{u} \in \boldsymbol{S}} \boldsymbol{p}[\boldsymbol{u}] \boldsymbol{V}_t^{\lambda_t}(\boldsymbol{u}) \right] = \boldsymbol{V}_{t-1}^{\lambda_{t-1}}(\boldsymbol{s}),$$

proving the lemma statement. □

4.3. Linear programming approximation

The results of Section 4.2 show that performing Lagrangian relaxation decomposes the problem and gives an upper bound on the value function of the big MDP. Using the recursing in Lemma 4.2 we can solve for that upper bound given any specific value of the Lagrange multiplier, $\boldsymbol{\lambda_t}$. In this section we show how to compute the optimal value for the Lagrange multiplier, in other words, the values that give the tightest upper bound.

We begin by writing the optimization problem to compute the Lagrangian relaxed value function in Lemma 4.2. The following optimization problem is similar to the optimization problem used to compute the value function for a standard MDP (Puterman, 2005). It has three main differences: 1) it is for a finite horizon problem, expressed by constraints for each time period 2) it is for a robust problem, expressed by the minimization operator within each constraint 3) it is for a decomposable problem, expressed by the separate variables for each $i \in I$. We let $\boldsymbol{\alpha} \in \mathbb{R}^{|\boldsymbol{S}|}$ to denote a probability distribution on the initial states of the big MDP, however all we require below is that it is a non-negative vector. The following optimization problem computes the expected payoff of the big MDP under the Lagrangian relaxed value function:

$$\min_{\boldsymbol{\lambda_t}, V_t^{\lambda_t,i}(\cdot), \forall i,t} \sum_{\boldsymbol{s}\in\boldsymbol{S}} \boldsymbol{\alpha}[\boldsymbol{s}] \sum_{i\in I} V_1^{\lambda_1,i}(\boldsymbol{s}[i]) \tag{2}$$

$$V_{t-1}^{\lambda_{t-1},i}(s) \geq \frac{\boldsymbol{\lambda_{t-1}} \cdot \boldsymbol{b}_{t-1}}{|I|} + r_i(s,a) - \boldsymbol{\lambda_{t-1}} \cdot C_{t-1}^i(s,a) + \beta \min_{p^i \in \mathbb{P}^i_{s,a}} \sum_{u\in\mathcal{S}_i} p^i[u] V_t^{\lambda_t,i}(u)$$

$$\forall s \in \mathcal{S}_i, a \in \mathcal{A}_i(s), t \in \{2,\ldots,T\}$$

Now, lets consider the following lemma which appears as Lemma 1 (Nilim and Ghaoui, 2005), the proof of which we provide here for completeness:

Lemma 4.3. *Consider the problem:*

$$\begin{aligned} \eta \quad = \quad & \min_{v_1,\ldots,v_{T-1}} q \cdot v_1 \\ & v_t \geq g_t(v_{t+1}), \quad \forall t = 1,\ldots,T-1, \end{aligned}$$

where v_T is a given vector of constants, and the v_i for $i = 1,\ldots,T-1$ are vectors of variables all of the same size. If we have that $q \geq 0$ and $u \leq w \Longrightarrow g_t(u) \leq g_t(w)$, where $q, u, w, g_t(\cdot)$ are vectors or vector valued functions, then the optimal solution is

$$v_t = g_t(v_{t+1}) \quad \forall t = 1,\ldots,T-1.$$

Proof. The proof here comes from the component-wise monotonicity of the g functions. In particular, let the proposed optimal solution in the lemma statement be v_t^*. It is feasible by construction. But, also, for any solution, we have $v_{T-1} \geq g_{T-1}(v_T) = g_{T-1}(v_T^*) = v_{T-1}^*$, because all solutions have one and the same value at T (no decisions are made and all values are fixed constants). From the monotonicity of g_{T-2}, we have

$v_{T-2} \geq g_{T-2}(v_{T-1}) \geq g_{T-2}(v^*_{T-1}) = v^*_{T-2}$. Continuing this argument in an inductive way, we have $v_1 \geq v^*_1$, which combined with $q \geq 0$ proves the result. □

So, if we can express problem (2) in the form of the problem in Lemma 4.3, and if we can prove that the corresponding g functions are component wise monotone, then we can derive an optimal solution to problem (2).

So, we re-write problem (2) in an appropriate form. Let v_t denote the vector with components $V_t^{\lambda_t,i}(\cdot), \forall i$. Let v_t^i denote the coordinates of v_t that correspond to little MDP i. We can then rewrite problem (2) as

$$\begin{aligned} \eta \quad = \quad & \min_{v_1,\ldots,v_{T-1}} q \cdot v_1 \\ & v_t \geq g_t(v_{t+1}), \quad \forall t = 1,\ldots,T-1, \end{aligned} \tag{3}$$

where $g_t(v_{t+1})$ returns the vector v that minimizes:

$$\begin{aligned} g_t(v_{t+1}) \quad = \quad & \operatorname*{argmin}_{v} \min_{\boldsymbol{\lambda_t}} 1 \cdot v \\ & v^i[s] \geq \frac{\boldsymbol{\lambda_t} \cdot \boldsymbol{b}_t}{|I|} + r_i(s,a) - \boldsymbol{\lambda_t} \cdot C_t^i(s,a) + \beta \min_{p^i \in \mathbb{P}^i_{s,a}} \sum_{u \in \mathcal{S}_i} p^i[u] v^i_{t+1}[u] \\ & \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall i \in I, s \in \mathcal{S}_i, a \in \mathcal{A}_i(s). \end{aligned} \tag{4}$$

The vector q in (3) equates to $\boldsymbol{\alpha}[\boldsymbol{s}]$ in problem (2). Intuitively, the main step in the transformation is that we moved the optimization over $\boldsymbol{\lambda_t}$ in problem (2) inside the definition of g_t in (3).

Theorem 4.4. *The optimal objective function values of* (2) *and* (3)-(4) *are the same. In addition, any optimal solution to* (3)-(4) *is also optimal in* (2).

Proof. Any $v_1, v_2, \ldots, v_{T-1}$ and $\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \ldots, \boldsymbol{\lambda_{T-1}}$ feasible in (3)-(4) is also feasible in (2) with the same objective function value.

Consider any feasible solution to (2), $\widehat{\boldsymbol{\lambda_t}}$ and $\widehat{V_t^{\lambda_t,i}}(\cdot)$. We group the values $\widehat{V_t^{\lambda_t,i}}(\cdot)$ across little MDPs into vectors $\widehat{v_1}, \widehat{v_2}, \ldots, \widehat{v_{T-1}}$. We define $v_t^{\min} = g_t(\widehat{v_{t+1}})$. For $t = T$, we have that $g_{T-1}(\widehat{v_T}) = v_{T-1}^{\min} \leq \widehat{v_{T-1}}$ The first equality follows by the definition of $v_{T-1}^{\min}$, the inequality follows from the definition of $g_{T-1}(\cdot)$ and the fact that $\widehat{v_T}$ is feasible in (2). We assume the monotonicity of the g_t functions momentarily, to be shown in Lemma 4.5, and proceed by induction on t. The induction hypothesis is $g_{t-1}(\widehat{v_t}) \leq \widehat{v_{t-1}}$. By induction: $g_{t-2}(v_{t-1}^{\min}) \leq g_{t-2}(\widehat{v_{t-1}}) = v_{t-2}^{\min} \leq \widehat{v_{t-2}}$. The first inequality follows from applying the monotonic $g_{t-2}(\cdot)$ to both sides of the induction hypothesis and substituting the definition of $v_{t-1}^{\min}$. The second equality is the definition of $v_{t-2}^{\min}$. The third inequality follows from the fact that $\widehat{v_{t-2}}$ is feasible in (2). By this inductive argument, we have $v_1^{\min} \leq \widehat{v_1}$, and $g_t(v_{t+1}^{\min}) \leq v_t^{\min}$ for $t = 1, \ldots T-1$. Thus we constructed a solution to (3)-(4) with a smaller value as $q \geq 0$. □

The only thing left to show is the monotonicity of the $g_t(\cdot)$ function defined in (4). That monotonicity is required in two places. First, we used it in the proof of Theorem 4.4. Second, it is required by Lemma 4.3 to provide a solution method for problem (3)-(4).

Lemma 4.5. *Let $g_t(v)$ be as defined in (4). Then $u \leq w \implies g_t(u) \leq g_t(w)$.*

Proof. Let ${\boldsymbol{\lambda_t}}^w, p^{i,w}$ be values of $\boldsymbol{\lambda_t}$ and p^i that produce the argmin for $g_t(w)$. Plug the same values into $g_t(u)$, and we'll get a component-wise smaller vector. This is because

$$\beta \min_{p^i \in \mathbb{P}^i_{s,a}} \sum_{s \in \mathcal{S}_i} p^i[s]u^i[s] \leq \beta \sum_{s \in \mathcal{S}_i} p^{i,w}[s]u^i[s] \leq \beta \sum_{s \in \mathcal{S}_i} p^{i,w}[s]w^i[s] = \beta \min_{p^i \in \mathbb{P}^i_{s,a}} \sum_{s \in \mathcal{S}_i} p^i[s]w^i[s],$$

where the first inequality comes from plugging a specific value into the minimization problem; the second inequality follows from the fact that w is component wise greater than u, and the last equality follows from the definition of $p^{i,w}$. Thus we have: $\beta \min_{p^i \in \mathbb{P}^i_{s,a}} \sum_{s \in \mathcal{S}_i} p^i[s]u^i[s] \leq \beta \min_{p^i \in \mathbb{P}^i_{s,a}} \sum_{s \in \mathcal{S}_i} p^i[s]w^i[s]$. Plugging ${\boldsymbol{\lambda_t}}^w$ into the definition of $g_t(u)$, shows that $g_t(u)$ is component wise smaller than $g_t(w)$, thus completing the proof. □

5. Solving the Model

The following steps compute strategies for the robust, decomposable, finite horizon MDP. We begin by sketching the process from the top-down, starting with the goals and getting to specifics. We then conclude with the explicit steps to compute a strategy. At an intuitive level, the process involves a variant of back-propagation, followed by a computation of strategies on an as-needed basis.

Computing a strategy is an optimization problem, if we are given the values of the future states as input. Thus, we would like to compute the values of the future states, $\boldsymbol{V}_{t+1}(\boldsymbol{s})$ if we are in time period t. Computing these values is computationally difficult, for example, a big MDP composed of 10 little MDPs each with 5 states requires the computation of 5^{10} values in each time period. Instead we compute upper bounds on the values, based on the Lagrangian relaxation of the robust, decomposable MDP, $\boldsymbol{V}^{\lambda_{t+1}}_{t+1}(\boldsymbol{s})$. By Lemma 4.1 we only have to compute $5 \cdot 10 = 50$ values for the Lagrangian relaxation of the big MDP composed of 10 little MDPs with 5 states each, instead of 5^{10}. In fact, we compute the smallest upper bounds, optimizing over the Lagrange multipliers. To do this, we would like to solve problem (2). That problem, however, includes non-linear terms in the constraints. By Lemmas 4.3 and 4.5, we can solve problem (4) instead of problem (2) to get the values $\boldsymbol{V}^{\lambda_{t+1}}_{t+1}(\boldsymbol{s})$ and the optimal Lagrange multipliers. Finally, to solve problem (4), we observe that the expression $\beta \min_{p^i \in \mathbb{P}^i_{s,a}} \sum_{u \in \mathcal{S}_i} p^i[u]v^i_{t+1}[u]$ can be optimized over $p^i \in \mathbb{P}^i_{s,a}$ to produce a constant, since the values v^i_{t+1} are specified as input to $g_t(v_{t+1})$. This gives the complete set of steps required to produce a strategy for the robust, decomposable, finite horizon MDP:

1. Begin by specifying $I, \mathcal{S}_i, \mathcal{A}_i(s), r_i(s,a), C^i_t(s,a), T, \boldsymbol{b_t}$ and $\boldsymbol{V}_T(\boldsymbol{s})$ for $i \in I$, $s \in \mathcal{S}_i$, $a \in \mathcal{A}_i(s)$, $t \in \{1, \ldots, T-1\}$. This defines the robust, decomposable, finite horizon MDP and assigns values to the horizon states.
2. Back-propagate, repeatedly solving problem (4) to compute values $\boldsymbol{V}^{\lambda_t}_t(\boldsymbol{s})$.
   (a) Start with $t = T-1$ and $v_T = \boldsymbol{V}_T(\boldsymbol{s})$.
   (b) Solve the inner problems $\min_{p^i \in \mathbb{P}^i_{s,a}} \sum_{u \in \mathcal{S}_i} p^i[u]v^i_T[u]$. This step assumes we can optimize over the uncertainty sets $\mathbb{P}^i_{s,a}$. We save the optimal solution $p^{i*}_{s,a}$.
   (c) Solve the outer problem, problem (4) to find $g_{T-1}(v_T)$, by substituting the objective values found in the previous step into the constraints of problem (4).

(d) This gives values for v_{T-1}, which specify $\boldsymbol{V}_{T-1}^{\lambda_{T-1}}(\boldsymbol{s})$. Repeat these steps to continue the back-propagation until $t = 1$.

3. Using the computed values for $\boldsymbol{V}_t^{\lambda_t}(\boldsymbol{s})$, compute strategies for the robust, decomposable, finite horizon MDP. These strategies can be computed on an as-needed fashion. In other words, given a state $\boldsymbol{s}$, which specifies $\boldsymbol{s}[i]$, and time period t, we can solve the MIP:

$$\begin{aligned}
\max_{x^i_{\boldsymbol{a}[i]}} \quad & \sum_{i \in I} \sum_{\boldsymbol{a}[i] \in \mathcal{A}_i(\boldsymbol{s}[i])} x^i_{\boldsymbol{a}[i]} (r_i(\boldsymbol{s}[i], \boldsymbol{a}[i]) + \sum_{u \in \mathcal{S}_i} p^{i*}_{\boldsymbol{s}[i], \boldsymbol{a}[i]}[u] \boldsymbol{V}_{t+1}^{\lambda_{t+1}}(u)) \qquad (5) \\
\text{s.t.} \quad & \sum_{\boldsymbol{a}[i] \in \mathcal{A}_i(\boldsymbol{s}[i])} x^i_{\boldsymbol{a}[i]} = 1, \quad \forall i \in I \\
& \sum_{i \in I} \sum_{\boldsymbol{a}[i] \in \mathcal{A}_i(\boldsymbol{s}[i])} C_t^i(s^i, \boldsymbol{a}[i]) x^i_{\boldsymbol{a}[i]} \leq \boldsymbol{b}_t \\
& x^i_{\boldsymbol{a}[i]} \in \{0, 1\}, \quad \forall i \in I, \boldsymbol{a}[i] \in \mathcal{A}_i(\boldsymbol{s}[i]).
\end{aligned}$$

In (5), $x^i_{\boldsymbol{a}[i]} = 1$ if action $\boldsymbol{a}[i]$ is taken in time period t in little MDP i and zero otherwise. The objective maximizes the total expected reward from taking the selected actions, $\boldsymbol{a}[i]$, across all little MDPs. The first constraint ensures that only one action is taken in each little MDP, the second constraint ensures that all actions across all the little MDPs are within budget, and the last constraint ensures that $x^i_{\boldsymbol{a}[i]}$ is binary. We observe that solving the above program with the true state values, $\boldsymbol{V}_{t+1}(\boldsymbol{s})$, as opposed to the upper bounds, $\boldsymbol{V}_{t+1}^{\lambda_{t+1}}(\boldsymbol{s})$, would yield an optimal strategy.

Even for a small example problem, solving the robust big MDP directly, computing the value for each big MDP state, as opposed to the Lagrangian relaxation we propose, involves solving a complex non-linear, non-convex optimization problem to find the worst-case transition probabilities. Because a generic algorithm to optimally solve non-linear, non-convex problems is not known, it is unlikely that the robust big MDP can be solved directly. Intuitively, the big MDP has a transition uncertainty set that is significantly different than that of Nilim and Ghaoui (2005), and the tools employed there are not applicable. We discuss a direct solution further in Appendix B. In contrast, the method described in this section is tractable and requires solving T linear programs in step 2 and T integer programs, on an as-needed fashion, with some additional optimization problems to compute worst-case transition probabilities.

6. Computational Results

In this section we apply the developed method to determine a funding policy for a stylized school district composed of four schools: a small wealthy (SW), a small impoverished (SI), a large wealthy (LW), and a large impoverished school (LI). The district has a finite budget per planning period, and would like to find a funding policy for a finite planning horizon, in our case 12 periods, that maximizes the school district's total expected reward over the planning horizon. It helps to think of each period representing a single year, and we are tasked with finding a funding policy for a 12 year period. Once a school receives its funding for a year, that school may choose to do with the budget anything it sees fit, such as hire more teachers, acquire more equipment, or fund

different programs. The implicit assumption in this model is that each school needs a certain amount of money to keep itself sustained, and additional funding may help it improve. Our results may be extended for any finite number of years, however we think that political and educational policies change so often that even 12 years may be too long for a single school district.

For our stylized example, a school receives income in two ways: 1) directly from the school district and 2) directly from the community. Some communities are able to collect more funds due to the makeup of the community than others, thus making some schools wealthier than others. These wealthy schools may have enough community funds to operate and fund all of their programs, and that makes them less responsive to any additional funds they may receive from the school district. With this perspective, in the stylized example, we assume wealthy schools' transition probabilities are independent of the district's funding actions. On the other hand, impoverished schools do not receive much funding through the local community and must depend on the school district for the majority of their funding.

Further, we assume that schools may be small or large. Small schools have fewer students than larger schools and thus need a lower amount of resources to function. However, if large schools are not performing well, to be defined later on in this section, it means that more students are also not performing well. Thus, the district receives a larger negative reward from large schools not performing well than from small schools.

Each of the four schools may be in one of five states: *failing, poor, average, good, excellent*. A school's state is determined by its performance on a yearly exam. Depending on how the students of that school perform, the school is given a performance grade that has the same value as its state. As mentioned previously, the school district would like to maximize its performance over the planning horizon, and the school district receives a reward for each school depending on the school's state at the end of each year. If a large schools is the *good* state, then it means that more students are in the *good* state than when a small school is in the *good* state. Therefore, the reward the district receives also depends on the size of the school. In this example, a district receives rewards of -10, -5, 0, 5, 10 for small schools and rewards of -20, -10, 0, 10, 20 for large schools being in states *failing, poor, average, good, excellent* respectively. Further, as large schools have more students than small schools, they also require more funding to keep their programs running. Each year, the school district can only decide if it will give a school *small, medium*, or *large* funding level. For the stylized example, large schools cost the district 0, 1, 3 and small schools cost the district 0, 1, 2 to have funding levels of *small, medium*, or *large*, respectively.

Each year the school district has to decide what action it should take, *i.e.*, what funding should the district give to each school. However, given the current state of schools and the funding decision taken, it is not clear to the school district what state a schools will be in the following year. Not only does the district not know the exact state, but the district also does not have an exact distribution on the future states given a state, action pair. The district may, at best, have a range, upper and lower bounds, on the likelihood of each future state given a state, action pair. It is with this range that we construct an uncertainty set on the probability of future states. We give a complete parameterization of the stylized example in Appendix D.

One possible comparison for this stylized example, is to compare policies derived from the robust, decomposable, finite horizon MDP to an *average case* policy, derived by as-

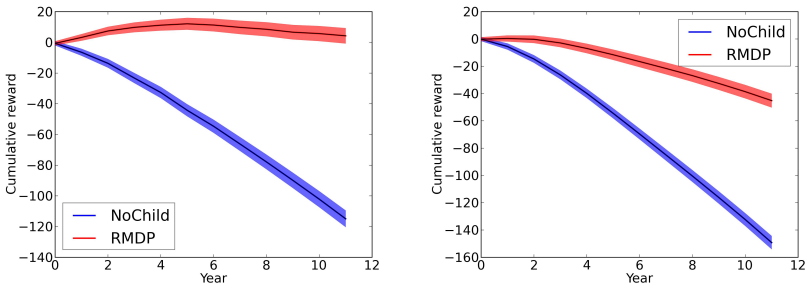
suming fixed transition probabilities. Although Nilim and Ghaoui (2005) do not consider coupled MDPs, their robust aircraft routing example highlights that average case policies may perform worse than a robust policy. The main purpose of our example is illustrate our method and compare it, at least in a stylized setting, to a school funding policy inspired by the real-world. An example that shows the robust policy performing better than the average policy is possible, though we do not think such an example will add to the understanding of the proposed method, especially since such a demonstration has already been done by Nilim and El Ghaoui. For completeness, we present a comparison of the average case policy to the robust policy for our example in Appendix C.

We compare funding policies resulting from the robust, decomposable, finite horizon MDP to a baseline policy inspired by a funding policy used in practice today, No Child Left Behind. No Child Left Behind, as implemented by the state of New Jersey (New Jersey Department of Education, Office of Student Achievement and Accountability, 2011), is rather complex, so the baseline that we use is a simplified abstraction of what is used in practice. We model No Child Left Behind by having a school be eligible to receive *large* funding if that school has reduced its performance this year relative to the previous year and its current performance is less than *good*. Of all the schools that are eligible to receive *large* funding, those schools with the lowest current state will receive funding first, and if multiple schools have the same current state, then larger schools are given preference. After *large* funding is allocated to some schools, the remaining budget is used to fund schools at the *medium* level, with schools with the lowest state given preference and large schools given preference over small schools.

In the remainder of this section we compare the funding policy determined using our method—referred to *RMDP* for the remainder of this section—to that of the policy inspired by No Child Left Behind—referred to *iNoChild* for the remainder of this section—in the four school stylized example described above. We consider two possible situations: 1) State evolution based on the worst case transition probabilities as defined by the uncertainty set and 2) State evolution based on a collection of random draws from the uncertainty set, *i.e.*, a point within the uncertainty set for the possible transition probabilities. The developed method, *RMDP*, is expected to perform well in the first case as the funding policy itself is designed assuming worst case transition probabilities. This may give this policy an unfair advantage over *iNoChild*, therefore, we also consider random draws from the feasibility set.

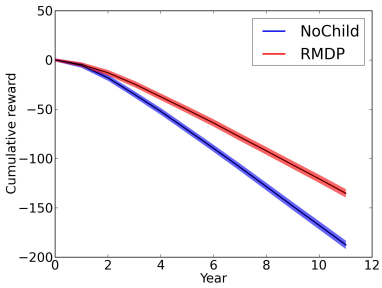
In Figure 1 we plot the cumulative expected reward the school district receives assuming worst case transition probabilities. *RMDP*, whose policy is targeted to worst case transitions, outperforms *iNoChild* across all budget values. In this figure, we omitted the plots of budgets of 5 and 4, as they are qualitatively similar to that of budgets of 6 and 3, respectively. Finally, for budgets of 2 and 1, shown in Figures 1(c) and 1(d), we note that though both policies provide negative cumulative reward, *RMDP* still performs at least as well, statistically speaking, as *iNoChild*, even in the first few years of the planning horizon.

In Figure 2 we plot the cumulative reward the school district receives assuming random draws from the transition probability uncertainty set. It is not surprising that the expected cumulative reward is higher for these types of transitions, relative to the worst case transitions. What may be surprising is that *RMDP* is still performing better than *iNoChild*. The only thing to note is that *RMDP* tends to perform a little worse on average terms, though equivalently, from a statistical perspective, than *iNoChild* when

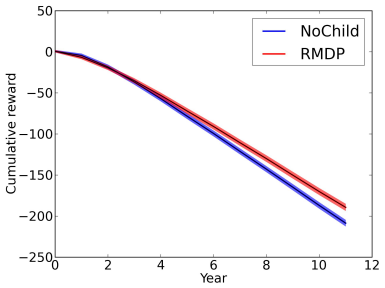


(a) Budget is 6

(b) Budget is 3

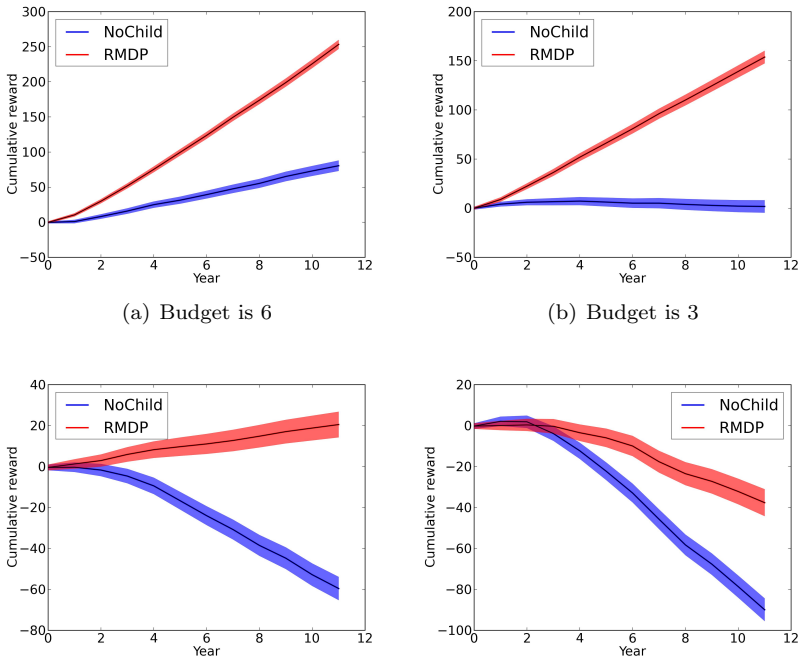


(c) Budget is 2



(d) Budget is 1

Figure 1: The cumulative expected rewards for the school district assuming worst case transition probabilities, with 95% confidence interval about each point



(a) Budget is 6

(b) Budget is 3

(c) Budget is 2

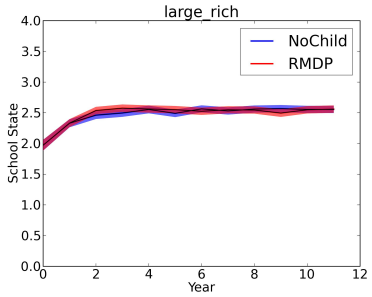
(d) Budget is 1

Figure 2: The cumulative expected rewards for the school district assuming random draws from the transition probability uncertainty sets, with 95% confidence interval about each point

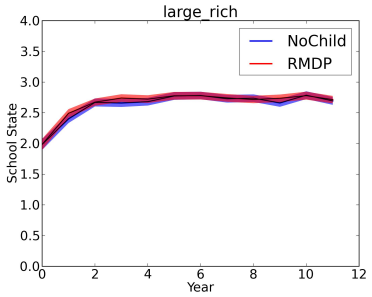
the budget is low, 2 or 1 in the initial two years. This may be an artifact from the fact that all schools start initially in the *average* state in our stylized example. Or, it may indicate that *RMDP* is too conservative initially as it considers performance over the entire planning horizon.

As mentioned earlier, funding to wealthy schools does not influence their transition probabilities. With this perspective, we would like any funding strategy to allocate funding to impoverished schools, and not spend resources on schools that do not respond to additional funding. In a sense, this part of the stylized example can be viewed as a sanity check. We expect *RMDP* to perform as well as *iNoChild*, in terms of school state, for all wealthy schools. Figures 3, 4 and 5 display the school states over the planning horizon under worst case transition probabilities for both *RMDP* and *iNoChild*. In Figure 3 we show that for both large and small wealthy schools, *RMDP* performs as well as *iNoChild*.

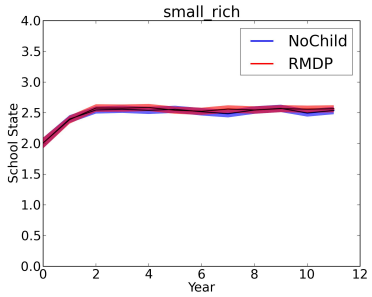
In Figure 4, we notice that *RMDP* spends resources in maintaining and improving the state of impoverished schools, agreeing with the sanity check. Qualitatively, the *RMDP* funding policy performs better then *iNoChild* for all impoverished schools for budget values of 5 and 6, under these budget constraints both the large and small impoverished school may be funded at the same time. However, for budgets of 4 and 3, it is impossible to give *large* funding or *medium* funding to both the large and the small school in the same



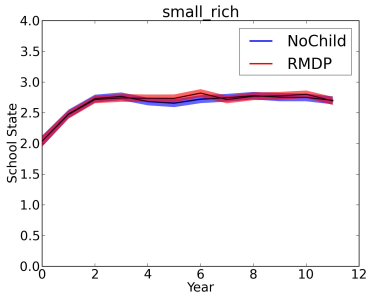
(a) Large school, worst case probability



(b) Large school, average case probability



(c) Small school, worst case probability



(d) Small school, average case probability

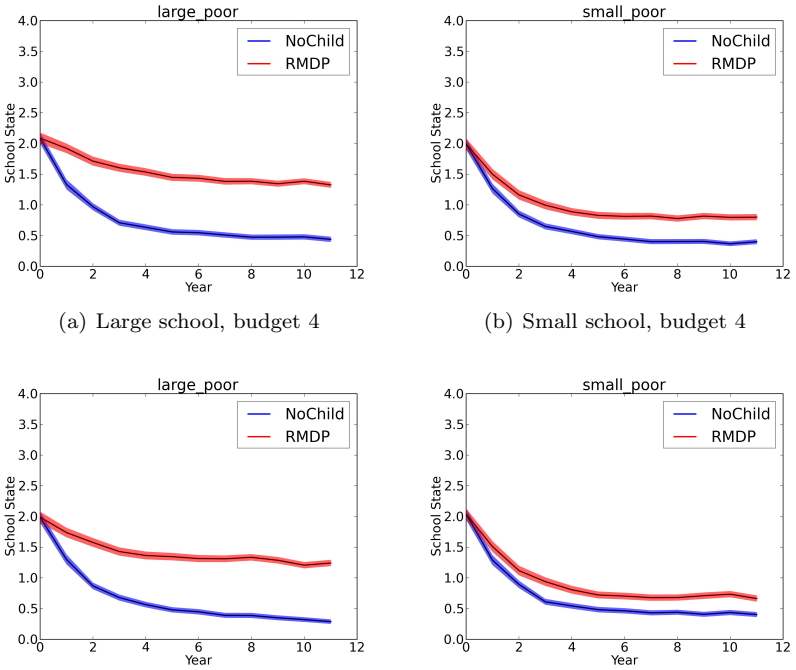
Figure 3: The state of small and large wealthy schools for worst case and average case probabilities with a budget of 4. *RMDP* performs equivalently to *iNoChild*.

year. As shown in Figure 4(c), the large impoverished school receives more support, and thus has a better state than the small school. This policy makes sense as more students are affected from funding decisions in the large school than in the small school. We observe the same behavior when we consider an average case probability realization.

Interestingly once the budget level is 2, in Figure 5, we notice that *RMDP* supports the small impoverished school more, as it can actually provide *large* funding to the small school where it can not for the large school. With a budget of 1, the *RMDP* policy again diverts more support to the large school and tries to provide *medium* funding year to year. We observe these trends for the worst case probability realizations in Figure 5(c), as with the previous figure, similar results hold for the average case of probability realizations.

7. Discussion and Conclusion

In this paper we propose a method to compute an approximate value function for robust, decomposable, finite horizon MDPs and show how to tractably derive a feasible policy from the approximate value function. The method is motivated by an application to school funding allocation and we apply it to a stylized school district. We compare our funding policies to those of a policy inspired by No Child Left Behind, and show that our policies perform well in comparison. At least based on the stylized example, this indicates that if decision makers make funding allocation decisions based on uncertain transition probabilities, it may lead to better results than making those decisions simply



(a) Large school, budget 4

(b) Small school, budget 4

(c) Large school, budget 3

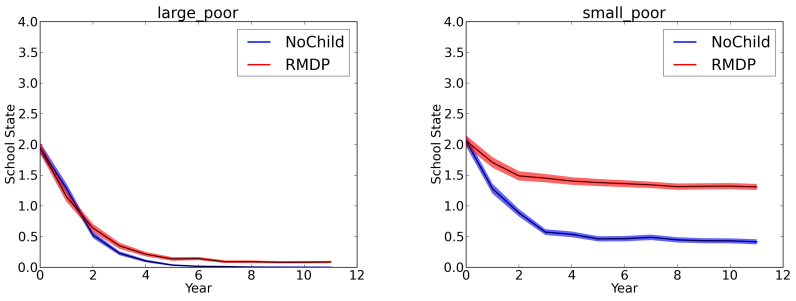
(d) Small school, budget 3

Figure 4: The state of small and large impoverished schools for worst case probabilities with a budget levels of 4 and 3. The large school receive *large* funding, and are thus in a better school state than the small school.

on the last observed transition as in the policy inspired by No Child Left Behind.

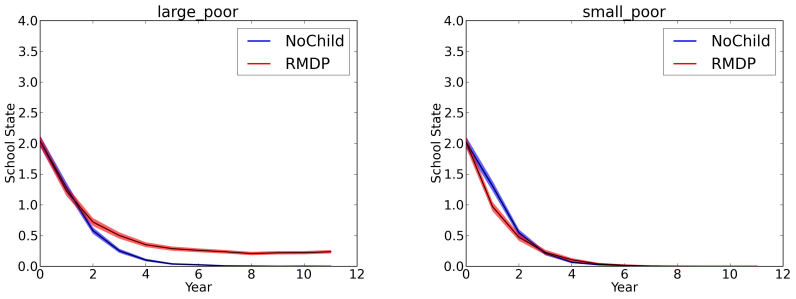
All school funding decisions are ultimately the result of a political process, and it is unrealistic to expect that the results of an optimization model would be used directly. Instead, the model we propose could be used to gain insight into the types of funding decisions that could lead to good outcomes. One way the model can be parameterized is by the decision maker experts directly, through specifying ranges of possible impacts to funding decisions. Another possible way to parameterize the model is through analyzing large data sets on school funding. New Zealand, for example, maintains a history of school funding and school performance on national exams (New Zealand Ministry of Education, 2013). It may be possible to statistically derive uncertainty sets from such a large data set, based on school characteristics. A real-world application of an optimization model to school funding allocation is an outstanding open problem. Inspired by this problem, the main contribution of this paper is the definition of, and suggested solution approach to robust, decomposable MDPs. These MDPs represent a fundamentally new type of control problem, different from both the decomposable MDPs and robust MDPs studied in the past.

There are a number of additional theoretical problems for future research. Currently, to the best of our knowledge, there is no known tractable method to compute optimal strategies for robust, decomposable MDPs. In addition, for realistic applications, both



(a) Large school, budget 2

(b) Small school, budget 2



(c) Large school, budget 1

(d) Small school, budget 1

Figure 5: The state of small and large impoverished schools for worst case probabilities with a budget levels of 2 and 1. With a budget of 2, small schools receive more funding, and thus have a better state than large schools. However, as the budget level drops, we notice that large schools again have a better state than small schools.

the rewards and the transition probabilities between little MDPs may be coupled. For example, a school district may be interested in equity–having all the schools at about the same level. Such considerations create difficulties for the decomposition approaches we take. The decomposition approaches ultimately work by decomposing the value function of the big MDP. If equity is a consideration, ultimately the value of a state in the big MDP is very much dependent on the states of all the little MDPs. The development of solution methods for such situations would be an interesting, though we believe difficult, theoretical contribution.

References

Adelman, D., A. J. Mersereau. 2008. Relaxations of weakly coupled stochastic dynamic programs. *Operations Research* **56** 712–727.

BenDavid-Hadar, I., A. Ziderman. 2011. A new model for equitable and efficient resource allocation to schools: the israeli case. *Education Economics* **19** 341–362.

Bessent, A., W. Bessent, J. Kennington, B. Reagan. 1982. An application of mathematical programming to assess productivity in the houston independent school district. *Management Science* **28** 1355 – 1367.

Blume, H. 2013. Gov. Jerry Brown pitches education budget in East L.A. school. *The Los Angeles Times*.

Borhan, M., A. A. Jemain. 2012. Assessing schools' academic performance using a belief structure. *Social Indicators Research* **106** 187–197.

Carnoy, M. 2007. *Cubas academic advantage: Why students in Cuba do better in school*. Stanford University Press.

Epple, D. N., R. Romano. 2003. *The Economics of School Choice: Neighborhood Schools, Choice, and the Distribution of Educational Benefits*, chap. 8. University of Chicago Press, College Station, Texas, 227 – 286.

Fensterwald, J. 2013. Brown commits $1 billion for Common Core, sticks with funding formula. *EdSource*.

Garber, Jim. 1997. Taxes and school funding problems. *The Bryan Times*. URL http://news.google.com/newspapers?nid=799&dat=19970531&id=IK8wAAAAIBAJ&sjid=700DAAAAIBAJ&pg=4689,2465829.

Glazebrook, K. D., D. J. Hodge, C. Kirkbride. 2011. General notions of indexability for queueing control and asset management. *Annals of Applied Probability* **21** 876–907.

Gould, E., V. Lavy, D. M. Paserma. 2004. Immigrating to opportunity: Estimating the effect of school quality using a natural experiment on ethiopians in israel. *The Quarterly Journal of Economics* **119** 489–526.

Hanushek, E. A. 1996. Measuring instrument in education. *The Journal of Economic Perspectives* **10** 9 – 30.

Jenkins, A., R. Levacic, A. Vignoles. 2006. Estimating the relationship between school resources and pupil attainment at GCSE. Tech. Rep. Research Report 727, London: Department of Education and Skills.

King, Richard A., Judith K. Mathers. 1997. Improving Schools Through Performance-Based Accountability and Financial Rewards. *Journal Of Educational Finance* **23** 147–176.

Meuleau, N., M. Hauskrecht, Kee-eung Kim, L. Peshkin, L. P. Kaelbling, T. Dean, C. Boutilier. 1998. Solving very large weakly coupled Markov decision processes. *In Proceedings of the Fifteenth National Conference on Artificial Intelligence*. 165–172.

Miles, K., M. Roza. 2006. Understanding student - weighted allocation as a means to greater school resource equity. *Peabody Journal of Education* **81** 39–62.

New Jersey Department of Education, Office of Student Achievement and Accountability. 2011. Understanding accountability in New Jersey for 2011 state assessments http://www.nj.gov/education/title1/accountability/ayp/1112/understanding.pdf. Tech. rep.

New Zealand Ministry of Education. 2013. Welcome to Education Counts. http://www.educationcounts.govt.nz. Accessed on 2013-12-28.

Nilim, A., L. E. Ghaoui. 2005. Robust control of Markov decision processes with uncertain transition matrices. *Operations Research* **53** 780–798.

Parvin, H., P. Goel, N. Gautam. 2012. An analytic framework to develop policies for testing, prevention, and treatment of two-stage contagious diseases. *Annals of Operations Research* **196** 707–735.

Powell, W. B. 2007. *Approximate Dynamic Programming: Solving the curses of dimensionality*, vol. 703. Wiley-Interscience.

Puterman, Martin L. 2005. *Markov Decision Processes: Discrete Stochastic Dynamic Programming (Wiley Series in Probability and Statistics)*. 1st ed. Wiley-Interscience. URL http://www.worldcat.org/isbn/0471727822.

Rubenstein, R, A E Schwartz, L Stiefel, H B H Amor. 2007. From districts to schools: The distribution of resources across schools in big city school districts. *Economics of Education Review* **26** 532 – 545.

Shutt, Dave. 1979. School Fund Allocation Revision is Overdue. *Toledo Blade*. URL http://news.google.com/newspapers?nid=1350&dat=19790614&id=tj9PAAAAIBAJ&sjid=rAIEAAAAIBAJ&pg=6238,3630857.

Sisikoglu, E., M. A. Epelman, R. L. Smith. 2011. A sampled fictitious play based learning algorithm for infinite horizon Markov decision processes. S. Jain, R. R. Creasey, J. Himmelspach, K. P. White, M. Fu, eds., *In Proceedings Winter Simulation Conference*. 4086 – 4097.

Tirenni, G., A. Labbi, C. Berrospi, A. Elisseeff, T. Bhose, K. Pauro, S. Pöyhönen. 2007. The 2005 ISMS practice prize winner–customer equity and lifetime management (CELM) Finnair case study. *Marketing Science* **26** 553–565.

Zhang, X. 2006. Markov-based optimization model for building facilities management. *Journal of construction engineering and management* **132** 1203–1211.

Appendix A. Approximate Dynamic Programming

We begin by assuming that the big MDP value function has the form $\boldsymbol{V}_t(\boldsymbol{s}) = \theta_t + \sum_{i \in I} V_t^i(\boldsymbol{s}[i])$ and writing the standard LP form for finding the values of each state:

$$\begin{aligned} \boldsymbol{H}^{\theta}(\boldsymbol{\alpha}) \quad = \quad & \min_{V_t^i(\cdot), \forall i,t} \theta_1 + \sum_{\boldsymbol{s} \in \boldsymbol{S}} \sum_{i \in I} \boldsymbol{\alpha}_i[\boldsymbol{s}[i]] V_1^i(\boldsymbol{s}[i]) \qquad (A.1) \\ & \theta_{t-1} + \sum_{i \in I} V_{t-1}^i(\boldsymbol{s}[i]) \geq \sum_{i \in I} r_i(s,a) + \beta \min_{\boldsymbol{p} \in \mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}} \sum_{\boldsymbol{u} \in \boldsymbol{S}} \prod_{i \in I} p^i[\boldsymbol{u}[i]] \left[\theta_t + \sum_{i \in I} V_t^i(\boldsymbol{u}[i]) \right] \\ & \qquad \forall \boldsymbol{s} \in \mathcal{S}, \boldsymbol{a} \in \bar{\boldsymbol{A}}_t(\boldsymbol{s}), t \in \{2, \ldots, T\}. \end{aligned}$$

Moving the constant θ_t out of the optimization objective, and using the same sequence of algebraic manipulations as in Lemma 4.1 we can rewrite the above as

$$\begin{aligned} \boldsymbol{H}^{\theta}(\boldsymbol{\alpha}) \quad = \quad & \min_{V_t^i(\cdot), \forall i,t} \theta_1 + \sum_{\boldsymbol{s} \in \boldsymbol{S}} \sum_{i \in I} \boldsymbol{\alpha}_i[\boldsymbol{s}[i]] V_1^i(\boldsymbol{s}[i]) \qquad (A.2) \\ & \theta_{t-1} - \beta\theta_t + \sum_{i \in I} V_{t-1}^i(\boldsymbol{s}[i]) \geq \sum_{i \in I} r_i(s,a) + \beta \sum_{i \in I} \min_{p^i \in \mathbb{P}^i_{\boldsymbol{s}[i], \boldsymbol{a}[i]}} \left(\sum_{u \in \mathcal{S}_i} V_t^i(u) p^i[u] \right) \\ & \qquad \forall \boldsymbol{s} \in \mathcal{S}, \boldsymbol{a} \in \bar{\boldsymbol{A}}_t(\boldsymbol{s}), t \in \{2, \ldots, T\}. \end{aligned}$$

Proposition 3 of Adelman and Mersereau (2008) holds, by the same reasoning as in their paper, giving that the parameters $\boldsymbol{\theta}$ does not change the objective function value of the above LP. In other words, $\boldsymbol{H}^{\boldsymbol{\theta}}(\boldsymbol{\alpha}) = \boldsymbol{H}^{\boldsymbol{\theta'}}(\boldsymbol{\alpha})$ for any two values $\boldsymbol{\theta}$ and $\boldsymbol{\theta'}$.

In addition, the remaining ADP results of Adelman and Mersereau also hold in the robust, decomposable, finite horizon MDP above. For example, their Proposition 4, stating that the ADP value approximation of each state is bigger than the true value of the state can be shown by an inductive argument similar to that in Theorem 4.4. Similarly, the equivalents of their Theorem 1 and Corollary 1 can be shown by taking an optimal solution to model (2), creating the same relationship between $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ as in their paper, and showing that it is feasible in model (A.2).

Because the same theoretical results hold in the robust, decomposable, finite horizon MDP, the ADP approximation gives more accurate value function approximations than the Lagrangian relaxation. However, the key problem with utilizing the ADP approximation is solving model (A.2). It has an exponential number of constraints, and in addition, unlike in Adelman and Mersereau, it has a non-linear term inside the constraints. Thus, standard column generation or row generation approaches can not be applied.

We sketch two possible computational approaches to solving model (A.2), however we believe both are complex and practically not scalable. The first is the standard coordinate-wise optimization approach often taken in non-linear programming. In this approach, we alternate between solving for the values of $V_t^i(\cdot)$ given fixed values of the p^i, then solving for the p^i given fixed values of the $V_t^i(\cdot)$. While this solution approach is common, we have no theoretical guarantees that it converges or produces optimal solutions. On the other hand, there may be applications where this approach works well.

Our second approach involves both row generation, for the constraints based on $\boldsymbol{s} \in \mathcal{S}, \boldsymbol{a} \in \bar{\boldsymbol{A}}_t(\boldsymbol{s})$ and column generation for decision variables p^i. The basic idea behind the approach is re-writing model (A.2) in the form

$$
\begin{aligned}
\boldsymbol{H}^{\theta}(\boldsymbol{\alpha}) \quad = \quad & \min_{V_t^i(\cdot), z, x} \ \theta_1 + \sum_{\boldsymbol{s} \in \boldsymbol{S}} \sum_{i \in I} \boldsymbol{\alpha}_i[\boldsymbol{s}[i]] V_1^i(\boldsymbol{s}[i]) \\
& \theta_{t-1} - \beta \theta_t + \sum_{i \in I} V_{t-1}^i(\boldsymbol{s}[i]) \geq \sum_{i \in I} r_i(s, a) + \beta \sum_{i \in I} z_{t, \boldsymbol{s}[i], \boldsymbol{a}[i]} \qquad \text{(A.3a)} \\
& \qquad \forall \boldsymbol{s} \in \mathcal{S}, \boldsymbol{a} \in \bar{\boldsymbol{A}}_t(\boldsymbol{s}), t \in \{2, \ldots, T\} \\
& z_{t, \boldsymbol{s}[i], \boldsymbol{a}[i]} \geq \sum_{u \in \mathcal{S}_i} V_t^i(u) \widehat{p^i}[u] - M(1 - x_{\widehat{p^i}}) \qquad \text{(A.3b)} \\
& \qquad \forall \boldsymbol{s} \in \mathcal{S}, \boldsymbol{a} \in \bar{\boldsymbol{A}}_t(\boldsymbol{s}), t \in \{2, \ldots, T\}, \widehat{p^i} \in \mathbf{Ext}(\mathbb{P}^i_{\boldsymbol{s}[i], \boldsymbol{a}[i]}) \\
& \sum_{\widehat{p^i} \in \mathbf{Ext}(\mathbb{P}^i_{\boldsymbol{s}[i], \boldsymbol{a}[i]})} x_{\widehat{p^i}} = 1 \qquad \text{(A.3c)} \\
& \qquad \forall \boldsymbol{s} \in \mathcal{S}, \boldsymbol{a} \in \bar{\boldsymbol{A}}_t(\boldsymbol{s}), t \in \{2, \ldots, T\} \\
& x_{\widehat{p^i}} \in \{0, 1\} \qquad \text{(A.3d)} \\
& \qquad \forall \boldsymbol{s} \in \mathcal{S}, \boldsymbol{a} \in \bar{\boldsymbol{A}}_t(\boldsymbol{s}), t \in \{2, \ldots, T\}, \widehat{p^i} \in \mathbf{Ext}(\mathbb{P}^i_{\boldsymbol{s}[i], \boldsymbol{a}[i]}),
\end{aligned}
$$

where $\mathbf{Ext}(\mathbb{P}^i_{\boldsymbol{s}[i], \boldsymbol{a}[i]})$ is the set of extreme points of the uncertainty set $\mathbb{P}^i_{\boldsymbol{s}[i], \boldsymbol{a}[i]}$.

The idea behind reformulation (A.3) is to assume the uncertainty sets are polyhedrons and rewrite the optimization over the uncertainty sets as a selection of a single extreme point. In particular, constraints (A.3b)-(A.3d) select a single extreme point, and the variables $z_{t, \boldsymbol{s}[i], \boldsymbol{a}[i]}$ compute the objective value of inner optimization problems. The reformulation removes the non-linearities in model (A.2), however it also introduces many more constraints. However, we can use a combination of row and column generation to solve model (A.3) to optimality.

Suppose we start with an incomplete version of model (A.3), in the sense that it only includes some subset of the extreme points in constraints (A.3b)-(A.3d) as well as only some of the constraints (A.3a) based on a subset of pairs $(\boldsymbol{s}, \boldsymbol{a})$. Suppose that the incomplete version of the model is small enough that we can solve the mixed integer program (MIP) to optimality, finding solution values $\widehat{V_t^i(\cdot)}, \widehat{z}$. We can check if the solution we received is optimal to the full model with the following two steps:

1. Solve the problems

$$
\min_{p^i \in \mathbb{P}^i_{\boldsymbol{s}[i], \boldsymbol{a}[i]}} \left(\sum_{u \in \mathcal{S}_i} \widehat{V_t^i(u)} p^i[u] \right).
$$

   Let the optimal value, say $z^*_{t, \boldsymbol{s}[i], \boldsymbol{a}[i]}$, occurs at an extreme point $\widehat{p^i}$. The values $z^*_{t, \boldsymbol{s}[i], \boldsymbol{a}[i]}$ give us what the true values of the z variables should have been, had we included all extreme points in the formulation, and the extreme point, $\widehat{p^i}$ gives us a new extreme point to include in constraints (A.3b)-(A.3d). This is the column generation part of the solution approach.

2. Solve the problem

$$\max_{\forall \boldsymbol{s}\in\mathcal{S},\boldsymbol{a}\in\boldsymbol{A}(t)\boldsymbol{s},t\in\{2,\ldots,T\}} \sum_{i\in I} r_i(s,a) - \sum_{i\in I} V^i_{t-1}(\boldsymbol{s}[i]) + \beta \sum_{i\in I} z^*_{t,\boldsymbol{s}[i],\boldsymbol{a}[i]}$$
$$\sum_{i\in I} C^i_t(\boldsymbol{s}[i],\boldsymbol{a}[i]) \leq \boldsymbol{b}_t.$$

This problem is the same knapsack problem that Adelman and Mersereau (2008) solve for the computational approach in their paper. If the objective function value of the problem is greater than zero, it generates a new pair $(\boldsymbol{s},\boldsymbol{a})$ to add to our incomplete version of constraint (A.3a). If the objective function value is zero or less, then all the constraints in the full formulation (A.3) are satisfied. This is the row generation part of the solution approach.

As we repeat the above two steps, if it is ever the case that we do not generate new extreme points in step one, or new constraints in step two, then we know we have found an optimal solution to the full problem. This approach will terminate in a finite number of iterations, simply because there are only finitely many extreme points and constraints we can generate. However, it could potentially take a long time, as with many row and column generation schemes.

Appendix B. Finding an Exact Solution Directly

Here, we illustrate that a direct method to solve the big MDP involves solving non-linear, non-convex optimization problems. The direct problem to solve for the values of all states of the big MDP is

$$\min_{\boldsymbol{V}_t(\cdot),\forall t} \sum_{\boldsymbol{s}\in\boldsymbol{S}} \boldsymbol{\alpha}[\boldsymbol{s}]\boldsymbol{V}_1(\boldsymbol{s}) \tag{B.1}$$
$$\boldsymbol{V}_{t-1}(\boldsymbol{s}) \geq \boldsymbol{r}(\boldsymbol{s},\boldsymbol{a}) + \beta \min_{\boldsymbol{p}\in\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}} \sum_{\boldsymbol{u}\in\boldsymbol{S}} \boldsymbol{p}[\boldsymbol{u}]\boldsymbol{V}_t(\boldsymbol{u})$$
$$\forall t \in \{2,\ldots,T\}, \boldsymbol{s}\in\boldsymbol{S}, \boldsymbol{a}\in\bar{\boldsymbol{A}}_t(\boldsymbol{s}).$$

In general, there is no guarantee that if the value function in the last time period is decomposible, $\boldsymbol{V}_T(\boldsymbol{u}) = \sum_{i\in I} V^i_T(\boldsymbol{u}[i])$, then the value funciton of the previous time period $\boldsymbol{V}_{T-1}(\boldsymbol{u})$ will be as well. In fact, that is the entire reason for the Lagrangian relaxation and ADP approaches, which work to achieve this decomposibility. Given a non-decomposible value function, consider the problem of finding the worst case transition probabilities $\min_{\boldsymbol{p}\in\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}} \sum_{\boldsymbol{u}\in\boldsymbol{S}} \boldsymbol{p}[\boldsymbol{u}]\boldsymbol{V}_t(\boldsymbol{u})$. Expanding the definition of $\mathbb{P}_{\boldsymbol{s},\boldsymbol{a}}$, we can re-write this problem as:

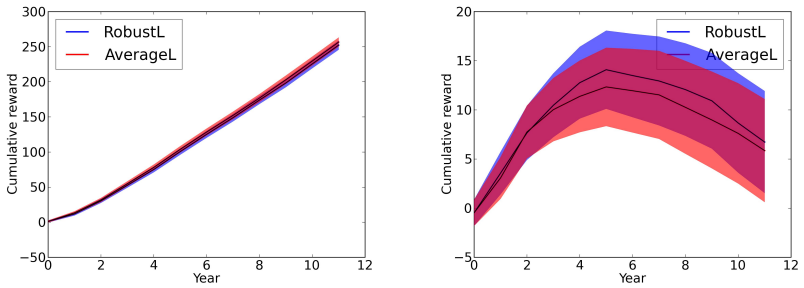
$$\min_{p^i\forall i} \sum_{\boldsymbol{u}\in\boldsymbol{S}} \left[\prod_{i\in I} p^i[\boldsymbol{u}[i]]\right] \boldsymbol{V}_t(\boldsymbol{u})$$
$$p^i \in \mathbb{P}^i_{\boldsymbol{s}[i],\boldsymbol{a}[i]}.$$

The product in the objective function makes this problem non-linear, and non-convex even if the individual uncertainty sets for the little MDPs, $\mathbb{P}^i_{\boldsymbol{s}[i],\boldsymbol{a}[i]}$ are polyhedrons.

Ultimately, the uncertainty set of the big MDP has a different structure than that used by Nilim and Ghaoui (2005), and so no known method exists to solve the big MDP exactly. To the best of our knowledge, there is no solver currently available that always solves (B.1) to optimality, and thus we are not able to compare our decomposition methods to an optimal solution.

Appendix C. Computational experiment

In this section we compare the performance of the robust policy to the *average policy*, the policy determined by considering the average transition probabilities between states. This policy substitutes the polyhedral uncertainty set with the expected value of a uniform distribution over all points in the polyhedron. The average policy is determined using the decomposition approach introduced by Adelman and Mersereau (2008). In Figure C.6 we show the cumulative awards for both policies, where RobustL, is the robust policy, and AverageL, is the average policy. The cumulative rewards shown in Figure C.6



(a) Cumulative Reward assuming average case draws.

(b) Cumulative Reward assuming worst case draws.

Figure C.6: Cumulative Rewards for a budget of 5 using average and robust policies

are qualitatively the same across all funding levels. As such, it is natural to ask if the robust policy adds any value over the average policy. The observed performance may be an artifact of the example we consider. Intuitively, the transition probability uncertainty is small, *i.e.*, the average transition probability is not much different from the worst case transition probability. One way to address this issue is to have one or more cases with high negative payoff but relatively low probability of occurring. In addition, if our probability uncertainty sets are not convex, then the average transition probability could lie outside the set. Creating an example where robust policies perform differently than average policies is outside the scope of our contribution. As we mention in Section 6, the purpose of our schools example is to convey that our algorithm may be used to solve a plausible practical problem. Nilim and Ghaoui (2005) already demonstrate that average case policies may perform worse than robust policies.

Appendix D. Parameterization of Stylized Example

In this section we describe the parameters used in our stylized computational experiment in Section 6. In the example we consider four schools: small wealthy (SW), small impoverished (SI), large wealthy (LW), and large impoverished (LI). Each school can be in one of five states: *failing*, *poor*, *average*, *good*, or *excellent*. Large schools provide the

school district rewards of -20, -10, 0, 10, 20 for being in each of the five respective states. Small schools provide the district rewards of -10, -5, 0, 5, 10 for being in each of the five respective states. At each planning epoch, the district determines the funding level for each school to be either *small*, *medium*, or *large* The cost for each funding level is 0, 1, and 2 for small schools, and 0, 1, and 3 for large schools, respectively.

Wealthy schools' transitions are independent of the actions taken by the school district, thus the transition uncertainty sets are only a function of a school's current state. In Table D.1 we list the box uncertainty sets for wealthy schools.

current state \ next state	failing	poor	average	good	excellent
failing	[0, 0.1]	[0.1, 0.5]	[0, 0.3]	[0, 0.1]	[0, 0.01]
poor	[0.05, 0.1]	[0, 0.01]	[0.1, 0.5]	[0, 0.3]	[0, 0.1]
average	[0, 0.05]	[0.05, 0.1]	[0, 0.2]	[0.1, 0.5]	[0, 0.2]
good	[0, 0.01]	[0, 0.1]	[0.1, 0.3]	[0, 0.5]	[0, 0.3]
excellent	[0, 0.01]	[0, 0.05]	[0, 0.2]	[0.2, 0.5]	[0, 0.4]

Table D.1: The box uncertainty sets for wealthy schools.

Unlike wealthy schools, impoverished schools respond to external funding levels, and thus the box uncertainty sets are a function of both the current state and the action. The uncertainty sets are listed in Table D.2.

current state, action \ next state	failing	poor	average	good	excellent
failing, small	[0.7,1]	[0, 0.3]	[0, 0.1]	[0, 0.05]	[0,0]
failing, medium	[0.4,1]	[0.2, 0.3]	[0, 0.15]	[0, 0.1]	[0, 0.05]
failing, large	[0.1, 0.2]	[0.4,1]	[0.2, 0.5]	[0, 0.15]	[0, 0.1]
poor, small	[0.7,1]	[0, 0.3]	[0, 0.1]	[0, 0.05]	[0,0]
poor, medium	[0.1,1]	[0.3, 0.5]	[0, 0.2]	[0, 0.1]	[0, 0.001]
poor, large	[0, 0.1]	[0, 0.3]	[0.45, 0.8]	[0, 0.2]	[0, 0.1]
average, small	[0.1, 0.3]	[0.3, 0.6]	[0, 0.2]	[0, 0.1]	[0,0]
average, medium	[0, 0.2]	[0.1, 0.4]	[0, 0.5]	[0, 0.3]	[0, 0.1]
average, large	[0, 0.1]	[0.05, 0.6]	[0, 0.8]	[0, 0.7]	[0, 0.3]
good, small	[0, 0.2]	[0, 0.3]	[0.3, 0.6]	[0, 0.4]	[0, 0.1]
good, medium	[0, 0.1]	[0, 0.2]	[0.1, 0.4]	[0, 0.6]	[0, 0.3]
good, large	[0, 0.05]	[0, 0.1]	[0.05, 0.2]	[0, 0.8]	[0, 0.6]
excellent, small	[0, 0.1]	[0, 0.2]	[0, 0.4]	[0.2, 0.8]	[0, 0.4]
excellent, medium	[0, 0.05]	[0, 0.1]	[0, 0.2]	[0.1, 0.6]	[0.3, 0.7]
excellent, large	[0,0]	[0, 0.05]	[0, 0.1]	[0.05, 0.15]	[0.5, 0.99]

Table D.2: The box uncertainty sets for impoverished schools.

Using the parameters above we solve the stylized example with the method outlined in Section 4 to generate the figures in Section 6.