

# A Moment and Sum-of-Squares Extension of Dual Dynamic Programming with Application to Nonlinear Energy Storage Problems

Marc Hohmann<sup>a,\*</sup>, Joseph Warrington<sup>b</sup>, John Lygeros<sup>b</sup>

<sup>a</sup>*Urban Energy Systems Group, Empa, Swiss Federal Laboratories for Materials Science and Technology, Überlandstrasse 129, 8600 Dübendorf, Switzerland*

<sup>b</sup>*Automatic Control Laboratory, ETH Zurich, Physikstrasse 3, 8092 Zürich, Switzerland*

---

## Abstract

We present a finite-horizon optimization algorithm that extends the established concept of Dual Dynamic Programming (DDP) in two ways. First, in contrast to the linear costs, dynamics, and constraints of standard DDP, we consider problems in which all of these can be polynomial functions. Second, we allow the state trajectory to be described by probability distributions rather than point values, and return approximate value functions fitted to these. The algorithm is in part an adaptation of sum-of-squares techniques used in the approximate dynamic programming literature. It alternates between a forward simulation through the horizon, in which the moments of the state distribution are propagated through a succession of single-stage problems, and a backward recursion, in which a new polynomial function is derived for each stage using the moments of the state as fixed data. The value function approximation returned for a given stage is the point-wise maximum of all polynomials derived for that stage. This contrasts with the piecewise affine functions derived in conventional DDP. We prove key convergence properties of the new algorithm, and validate it in simulation on two case studies related to the optimal operation of energy storage devices with nonlinear characteristics. The first is a small borehole storage problem, for which multiple value function approximations can be compared. The second is a larger problem, for which conventional discretized dynamic programming is intractable.

*Keywords:* Control, Dual dynamic programming, Moment/SOS techniques, Long-term energy storage management

---

## 1. Introduction

Dual Dynamic Programming (DDP) (Pereira & Pinto, 1991), also referred to as nested Benders decomposition, is a means of solving multi-stage optimization problems in which constraints on decision variables are coupled only across adjacent stages. The most common application is in a linear, stochastic setting, where it is referred to as Stochastic Dual Dynamic Programming (SDDP). The algorithm relies on a Benders decomposition argument to generate increasingly tight lower

---

\*Corresponding author

bounds on the optimal cost-to-go at each stage. Convergence to optimality of these bounds and of forward state trajectories has been studied in Philpott & Guan (2008) for the linear case, and Girardeau et al. (2015) for the general nonlinear case. Inexact approaches featuring suboptimal cuts and/or forward state trajectories were studied in Zakeri et al. (2000) and Guigues (2018), and a number of other extensions have been developed, notably for risk-averse decision making (Guigues & Römisch, 2012) and multi-stage integer problems (Zou et al., 2018).

In a multi-stage setting, value functions allow single-stage decisions to be taken without explicit consideration of the remainder of the time horizon. This is relevant in many energy applications featuring storage of some kind, where short-term decisions must often be made in the presence of long-term effects driven by slower, for example seasonal, dynamics (see Abgottspon, 2015; Dariyanakis et al., 2017). A locally-tight approximation of the cost-to-go allows relatively efficient trade-offs between short- and long-term costs to be made, even when an exogenous disturbance, or modelling error, may have caused the system state to deviate somewhat from a previously computed trajectory. The value function approximations generated by (S)DDP often have this property, and can therefore be well suited to this purpose.

However, a shortcoming common to many nested decomposition approaches, including (S)DDP, is that they are only applicable to systems with linear dynamics, costs, and constraints, or with “benign” (convex) nonlinearities (Girardeau et al., 2015). Many problems to which (S)DDP could otherwise be applied feature nonconvex, in particular polynomial, relationships between variables. Examples of polynomial nonlinearities in the energy domain include hydro storage planning with head effects (Cerisola et al., 2012), district heating networks (Jiang et al., 2014), borehole management using heat pumps (Atam et al., 2015), and alternating-current (AC) power system optimization (Taylor, 2015). Although in some cases it is possible to apply a convex approximation, for example McCormick envelopes for bilinear functions Cerisola et al. (2012), this may not offer acceptable modelling accuracy.

For low-dimensional nonlinear systems, it is possible in a very broad range of cases to compute a near-optimal value function by discretizing the state and input spaces and performing the standard Dynamic Programming (DP) recursion (Bertsekas, 1995). This approach has been applied to seasonal borehole storage problems in De Ridder et al. (2011) and Atam et al. (2015), but it becomes impractical for systems with more than only a few states and inputs due to exponential memory and computation requirements. It is therefore desirable to extend the existing theory of DDP to handle nonlinear systems, in order to take advantage of DDP’s relative scalability.

Other Approximate Dynamic Programming (ADP) (Powell, 2011) approaches address the drawbacks of discretized DP by using relaxations of the dynamic programming principle, most commonly in an infinite-horizon setting. Recent approaches such as Wang et al. (2014), Summers

et al. (2012), and Beuchat et al. (2017) propose tractable approximations to the Linear Programming (LP) formulation of ADP (Hernández-Lerma & Hernández-Hernández, 1994), in which the computation of a value function is cast as an (infinite-dimensional) LP. The authors of Savorgnan et al. (2009), Kamoutsi et al. (2017), and Lasserre et al. (2008) formulate a Generalized Moment Problem (GMP) over occupation measures, of which this LP formulation is a dual. They derive tractable approximations of the GMP and LP formulation in the form of moment relaxations and Sum-of-Squares (SOS) programs for approximate control synthesis of polynomial systems. In these approaches, the optimal control problem is solved for a specified initial state distribution. It should also be noted that GMPs have gained interest recently in the energy domain outside of DP, due to their ability to find global solutions of the AC optimal power flow problem (Ghaddar et al., 2016; Molzahn & Hiskens, 2015).

In this paper, we develop an approach that brings the advantages of the LP formulation of ADP to DDP, in that it handles polynomial costs, dynamics, and constraints, and fits the value function to trajectories emanating from an initial state *distribution*, in contrast to the single initial state used in conventional DDP. As with conventional DDP, the algorithm performs an iterative sequence of forward simulations and backward recursions. The forward simulation consists of moment problems approximating the occupation measure of candidate trajectories, while the backward recursion is composed of SOS programs, dual to the moment problems, that generate under-approximators of the value function. The output of our proposed algorithm is a collection of functions for each stage, the point-wise maximum of which under-approximates the true value function. This yields a richer class of approximations than the Moment/SOS approaches of Lasserre et al. (2008) and Savorgnan et al. (2009) for polynomial dynamical systems, which rely on a single, high-order polynomial to increase accuracy. The methods developed in O’Donoghue et al. (2011) and Beuchat et al. (2017) also generate a point-wise maximum under-approximation in an iterative fashion, but do not use the primal side over moments of the occupation measure to refine the approximate value functions.

Specifically, we make the following contributions:

- We extend the well-known DDP framework to generic polynomial dynamical systems using moment/SOS techniques. We define an algorithm, Moment DDP, that generates increasingly tight lower bounds on each stage’s value function, and corresponding moments of the state distribution at each stage. This algorithm generates value function estimates that are valid for a *probability distribution* of initial states, encompassing the single initial state (or Dirac distribution) from conventional DDP as a special case.
- We prove that (i) the upper and lower cost bounds generated by the algorithm converge to at least the optimal cost of a relaxation of the finite-horizon decision problem and at most

the optimal cost of the original GMP, and (ii) this relaxation becomes tight in the limit as the order of the moment relaxation increases.

- We describe the stochastic extension of Moment DDP, and give conditions under which the uncertainty can be accommodated within the same framework.
- We demonstrate Moment DDP numerically with a nonlinear seasonal geothermal borehole dispatch problem based on real measurement data. Furthermore, we report successful application of the algorithm to a higher-dimensional system that is computationally too demanding for conventional discretized DP.

Section 2 states the class of finite-horizon polynomial problems considered in our framework, and presents a finite-horizon discrete-time SOS approach to ADP inspired by recent optimal control literature. Section 3 describes the Moment DDP algorithm, and Section 4 states and proves its key convergence properties. Section 5 presents numerical results for two nonlinear borehole systems of different state dimensions. Section 6 concludes and gives an outlook for future research.

### 1.1. Notation and preliminaries

The sets  $\mathbb{R}$ ,  $\mathbb{N}$  and  $\mathbb{N}^+$  denote the real numbers, non-negative and positive integers respectively. For a compact real vector space  $\mathbf{S}$ , let  $\mathcal{M}(\mathbf{S})$  be the set of Borel measures on  $\mathbf{S}$  and  $\mathcal{C}(\mathbf{S})$  the set of bounded continuous functions on  $\mathbf{S}$ . Together they form a dual pair  $(\mathcal{M}(\mathbf{S}), \mathcal{C}(\mathbf{S}))$  with duality brackets  $\langle v, \mu \rangle = \int_{\mathbf{S}} v d\mu$  for  $v \in \mathcal{C}(\mathbf{S})$ . If  $v$  is polynomial, we write the duality bracket as an inner product  $\langle \mathbf{v}, \mathbf{m} \rangle$ , where the vector  $\mathbf{v}$  contains the coefficients of  $v$  and the vector  $\mathbf{m}$  the corresponding moments of  $\mu$ .  $\mathcal{M}(\mathbf{S})_+$  denotes the set of positive Borel measures on  $\mathbf{S}$ . A positive Borel measure  $\varphi$  supported on  $\mathbf{S}$  with  $\varphi(\mathbf{S}) = 1$  is called a Borel probability measure. A special case of a Borel probability measure is a Dirac measure  $\delta_x$  supported on a single point  $x \in \mathbf{S}$ . The operator  $\otimes$  defines the cross product of two probability measures. The expected value with respect to a Borel probability measure  $\varphi$  is defined as  $\mathbf{E}_{\varphi}(x) = \int_{\mathbf{S}} x d\varphi$ . For a Borel set  $A$ , we define  $1_A(x)$  as an indicator function equal to 1 if  $x \in A$  and 0 if  $x \notin A$ .

Let  $\mathbb{R}[x]_k$  be the ring of polynomials of degree at most  $k$  in some variable  $x \in \mathbb{R}^n$ , and let  $\deg(p)$  denote the degree of  $p$ . The notation  $\Sigma_{2k}[x]$  stands for the Sum-of-Squares polynomials of degree at most  $2k$  in  $x$ . Polynomial  $p(x) \in \Sigma_{2k}[x]$  if and only if there exist polynomials  $\xi_1(x), \dots, \xi_{N_{\xi}}(x)$  such that  $p(x) = \sum_{i=1}^{N_{\xi}} \xi_i(x)^2$ , which implies that  $p(x) \geq 0$  for all  $x$ . This is equivalent to there existing a symmetric, positive semidefinite matrix  $\mathbf{P}$  (we denote this  $\mathbf{P} \succeq 0$ ) such that  $p(x) \equiv \tilde{p}(x)^{\top} \mathbf{P} \tilde{p}(x)$ . In this definition,  $\tilde{p}(x) := (1, x_1, x_2, \dots, x_1 x_2, \dots, x_n^k)$  is the vector of all possible monomials in  $x$ , of degree up to  $k$ . An optimization over the elements of  $\mathbf{P}$ , with the linear matrix inequality (LMI) constraint that  $\mathbf{P} \succeq 0$ , therefore yields parameterizations of SOS polynomials as solutions. We refer to the degree of a SOS polynomial  $p(x)$  as  $2k$  since  $\deg(p)$  is always an even number.

The truncated *quadratic module* of degree  $k$ , generated by the polynomials  $h_i(x)$  of a semi-algebraic set  $\mathbf{S} := \{h_i(x) \geq 0, i = 1, \dots, N_h\}$ , is defined as

$$\mathbf{Q}_k(\mathbf{S}) := \sigma_0(x) + \sum_{i=1}^{N_h} \sigma_i(x) h_i(x), \quad (1)$$

where  $\sigma_0 \in \Sigma_{2k}[x]$  and  $\sigma_i \in \Sigma_{2k}[x]$ , with the restriction that  $\deg(\sigma_i h_i) \leq 2k$ . Such polynomials are guaranteed to be non-negative for all  $x \in \mathbf{S}$ .

## 2. Problem statement and background

### 2.1. Finite horizon problem

We consider a finite-horizon decision problem of the form (2), and the corresponding optimal value  $V_0^*(x_0)$  for given  $x_0$ :

$$V_0^*(x_0) := \min_{\{x_t\}_{t=1}^T, \{u_t\}_{t=0}^{T-1}} \sum_{t=0}^{T-1} l_t(x_t, u_t) + H(x_T) \quad (2a)$$

$$\text{s.t. } x_{t+1} = f_t(x_t, u_t), \quad t = 0, \dots, T-1, \quad (2b)$$

$$g_{t,j}(x_t, u_t) \geq 0, \quad j = 1, \dots, N_{g,t}, \quad t = 0, \dots, T-1, \quad (2c)$$

$$g_{T,j}(x_T) \geq 0, \quad j = 1, \dots, N_{g,T}. \quad (2d)$$

Vector  $x_t \in \mathbb{R}^{n_x}$  represents the state at stage  $t$ ,  $u_t \in \mathbb{R}^{n_u}$  is a vector of control inputs (or actions), and  $t = 0, \dots, T$  is the time index over a prediction horizon of length  $T \in \mathbb{N}^+$ . Stage costs are defined by functions  $l_t : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \rightarrow \mathbb{R}$  and the terminal cost function is  $H : \mathbb{R}^{n_x} \rightarrow \mathbb{R}$ . The dynamics are modelled by the function  $f_t(x_t, u_t) : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \rightarrow \mathbb{R}^{n_x}$ , and the constraint functions  $g_{t,j}(x_t, u_t) : \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} \rightarrow \mathbb{R}$  encode conservation laws and technical bounds on variables at each stage.

For later developments, we will assume that  $x_t$  includes an auxiliary state  $x_{c,t}$  on the interval  $[0, T]$  with update equation  $x_{c,t+1} = x_{c,t} + 1$ , thus representing the current time step  $t$  as a state.

With a minor abuse of notation, we say that constraints (2c) that are uncoupled from  $u_t$  define the state space  $\mathbf{X}_t := \{x_t \in \mathbb{R}^{n_x} : g_{t,j}(x_t) \geq 0, j = 1, \dots, N_{g_x,t}; x_{c,t} = t\}$ . Constraints (2c) that are uncoupled from  $x_t$  define the action space  $\mathbf{U}_t := \{u_t \in \mathbb{R}^{n_u} : g_{t,j}(u_t) \geq 0, j = N_{g_x,t} + 1, \dots, N_{g_u,t}\}$ . The feasible set of state and control decisions at time step  $t$  is defined as

$$\mathbf{C}_t := \{(x_t, u_t) \in \mathbb{R}^{n_x} \times \mathbb{R}^{n_u} : g_{t,j}(x_t, u_t) \geq 0, j = 1, \dots, N_{g,t}; x_{c,t} = t\}.$$

For any  $x_t \in \mathbf{X}_t$ , the set of admissible controls is defined as  $\mathbf{U}_t(x_t) := \{u_t : (x_t, u_t) \in \mathbf{C}_t\}$ . Since the sets  $\mathbf{C}_t$  and  $\mathbf{X}_t$  contain a constraint  $x_{c,t} = t$ , and all problem constraints will be defined for

states and inputs belonging to these time-indexed sets, we will drop the time subscripts from  $x$  and  $u$  to maintain clean notation, without loss of clarity. We will also refer to  $x_{c,t}$  as  $x_c$  under the same rationale.

Furthermore, we make the following assumptions:

**Assumption 1.** *Functions  $l_t(x, u)$ ,  $f_t(x, u)$ ,  $g_{t,j}(x, u)$  are polynomials for all  $t \in \{0, \dots, T-1\}$ , as is  $H(x)$ . The state and control decisions are bounded i.e.,  $\mathbf{C}_t$  and  $\mathbf{X}_t$  are compact.*

**Assumption 2.** *For all  $t = 0, \dots, T-1$ , for all  $x \in \mathbf{X}_t$  there exists at least one  $u \in \mathbf{U}_t(x)$  such that  $f_t(x, u) \in \mathbf{X}_{t+1}$ .*

The *value function*  $V_t^* : \mathbf{X}_t \rightarrow \mathbb{R}$  represents the sum of all costs incurred in problem (2) starting from state  $x_t$  at time instance  $t$ , if optimal control decisions are taken at all times from  $t$  to  $T-1$ . It is defined recursively by the well-known Bellman optimality condition at each stage  $t = 0, \dots, T-1$ :

$$V_t^*(x) := \min_{u \in \mathbf{U}_t(x)} \{l_t(x, u) + V_{t+1}^*(f_t(x, u))\}, \quad \forall x \in \mathbf{X}_t, \quad (3)$$

with the boundary condition  $V_T^*(x) = H(x)$  for all  $x \in \mathbf{X}_T$ .

## 2.2. Generalized moment problem

We now develop a finite-horizon discrete-time optimal control problem in the form of a GMP (Lasserre, 2014). Our formulation, an infinite-dimensional linear program over occupation measures, is a finite-horizon problem related to the GMP developed in Savorgnan et al. (2009). An occupation measure can be interpreted as a probability distribution describing the trajectory  $x$  and  $u$  of a dynamical system starting from a known initial state distribution.

Consider the (nonstationary) Markov control model formed by the tuple  $(\mathbf{X}_t, \mathbf{U}_t, \{\mathbf{U}(x)_t | x \in \mathbf{X}_t\}, f_t(x, u), l_t(x, u), H(x))$  for which we wish to find an optimal control policy  $\varrho^*$ . Note that for the purposes of the derivations which follow, nonstationary Markov control models can be represented using an equivalent stationary model using state augmentation (Hernández-Lerma, 1989, Section 1.3). Under Assumption 2, from (Hernández-Lerma & Lasserre, 2012, Theorem 3.2.1) there exists an optimal policy  $\varrho$  that is deterministic and can therefore be expressed in the form  $u = \varrho^*(x)$ . The state-action occupation measure at time step  $t$  for a given policy  $\varrho$  and initial state measure  $\nu_0$  is a Borel measure  $\mu_t \in \mathcal{M}(\mathbf{C}_t)_+$  on the feasible set  $\mathbf{C}_t$ , defined by

$$\mu_t(B) := \mathbf{E}_{\nu_0}^{\varrho}(1_B(x, u)) \quad (4)$$

for all Borel sets  $B$  of  $\mathbf{C}_t$ .  $\mathbf{E}_{\nu_0}^{\varrho}$  is the expected value under policy  $\varrho$  given some initial distribution  $\nu_0$  of the state. Measure  $\mu_t$  contains all information about the relationship between the state  $x$  and control input  $u$  (which depends on  $x$ ) at time step  $t$ .

Let  $\pi : \mathcal{M}(\mathbf{C}_t)_+ \rightarrow \mathcal{M}(\mathbf{X}_t)_+$  be the projection from state-action space onto the state alone.<sup>1</sup> Then the linear operator  $\mathcal{L}_t : \mathcal{M}(\mathbf{C}_t)_+ \rightarrow \mathcal{M}(\mathbf{X}_{t+1})_+$  maps the state-action occupation measure at time step  $t$  to the occupation measure projected onto the state space  $\mathbf{X}_{t+1}$  at time step  $t + 1$  under the dynamics  $f_t(x, u)$ :

$$\pi\mu_{t+1}(A) = \mathcal{L}_t\mu_t(A) = \int_{\mathbf{C}_t} 1_A(f_t(x, u))d\mu_t \quad (5)$$

for all Borel sets  $A$  of  $\mathbf{X}_{t+1}$ .<sup>2</sup> In words, the probability mass of the state distribution in set  $A$  at time  $t + 1$  is equal to the total contributions of mass brought into  $A$  by the dynamics, across all infinitesimal elements of the state-action distribution  $\mu_t$ . This operator therefore encodes consistency with the dynamics of successive state-action distributions  $(\mu_t, \mu_{t+1})$ .

Using these definitions, the following linear constraint describes all state-action probability measures  $\mu_0, \mu_1, \dots, \mu_{T-1}$  that are consistent with a control policy  $\varrho$ , the dynamics  $f_t(x, u)$ , and a free choice of terminal state measure  $(\nu_T \otimes \delta_T) \in \mathcal{M}(\mathbf{X}_T)_+$ :

$$\nu_0 \otimes \delta_0 + \sum_{t=0}^{T-1} \mathcal{L}_t\mu_t = \sum_{t=0}^{T-1} \pi\mu_t + \nu_T \otimes \delta_T. \quad (6)$$

We use  $\nu_t$  to denote a probability measure over all elements of vector  $x$  except the auxiliary time index state  $x_c$ , and  $\delta_t$  to denote the Dirac measure supported on  $t$  for  $x_c$ . Thus, measure  $\nu_0 \otimes \delta_0$  is an initial probability distribution on  $\mathbf{X}_0$ , where  $\delta_0$  accounts for  $x_c$  being supported on  $t = 0$ . Similarly,  $\nu_T \otimes \delta_T$  is the terminal Borel probability measure on  $\mathbf{X}_T$ . Note that the sum of measures on each side of (6) is supported on  $x_c = 0, 1, \dots, T$ , thus the single constraint encodes all  $T$ -step trajectories of the system.

We can now formulate the GMP (7), which is a  $T$ -step decision problem related to (2). Measures  $\mu_t$  and the terminal state measure  $\nu_T$  fully specify the solution of (2) for a given distribution  $\nu_0$  of the initial state  $x_0$ .

$$\rho^* := \min_{\{\mu_t\}_{t=0}^{T-1}, \nu_T} \sum_{t=0}^{T-1} \int_{\mathbf{C}_t} l_t(x, u)d\mu_t + \int_{\mathbf{X}_T} H(x)d(\nu_T \otimes \delta_T) \quad (7a)$$

$$\text{s.t. } \nu_0 \otimes \delta_0 + \sum_{t=0}^{T-1} \mathcal{L}_t\mu_t = \sum_{t=0}^{T-1} \pi\mu_t + \nu_T \otimes \delta_T, \quad (7b)$$

$$\mu_t \in \mathcal{M}(\mathbf{C}_t)_+, \nu_T \otimes \delta_T \in \mathcal{M}(\mathbf{X}_T)_+. \quad (7c)$$

<sup>1</sup>For any Borel measure  $\mu_t \in \mathcal{M}(\mathbf{C}_t)_+$  this is formally defined by  $(\pi\mu_t)(B) = \mu_t((\mathbb{R}^{n_u} \times B) \cap \mathbf{C}_t)$  for all Borel subsets  $B$  of  $\mathbf{X}_t$ .

<sup>2</sup>This operator was first defined in Lasota & Mackey (1994), and used for the infinite-horizon control application in Savorgnan et al. (2009).

**Theorem 1.** *The optimal value  $\rho^*$  of (7) is equal to the optimal cost  $V_0^*(x_0)$  of (2) when  $\nu_0$  is a Dirac measure on  $x_0$ , and equal to the expected value  $\mathbf{E}_{\nu_0}(V_0^*(x_0))$  when  $\nu_0$  is a probability measure.*

*Proof.* The finite-horizon problem (7), expressed as an equivalent stationary model (Hernández-Lerma, 1989, Section 1.3), is a special case of the infinite horizon GMP from Hernández-Lerma & Lasserre (2012) and Savorgnan et al. (2009). Problem (2) can be restated as an infinite-horizon problem by setting the cost functions for  $t > T$  to zero. Since the support of any  $\mu_t$  is limited to values of auxiliary state  $x_c$  on the interval  $[0, T - 1]$ , by definition of the measure  $\mu_t$ , we have  $\sum_{t=T+1}^{\infty} \pi \mu_t = 0$  and  $\sum_{t=T}^{\infty} \mathcal{L}_t \mu_t = 0$ . Thus, the infinite-horizon GMP presented in Savorgnan et al. (2009) reduces to (7). Due to Assumption 1 (which implies continuity of  $l_t(x, u)$  and  $f_t(x, u)$ , and compactness of  $\mathbf{C}_t$  and  $\mathbf{X}_t$ ), we have  $\rho^* = \mathbf{E}_{\nu_0}(V_0^*(x_0))$  by (Hernández-Lerma & Lasserre, 2012, Theorem 6.3.7).  $\square$

### 2.3. Value function approximation

To facilitate the decomposition approach in Section 3, we rewrite (7) by introducing state measures  $\nu_t \otimes \delta_t \in \mathcal{M}(\mathbf{X}_t)_+$  for  $t = 1, \dots, T - 1$ , and replacing the single dynamical constraint (7b) with  $T$  separate one-step constraints,

$$\nu_t \otimes \delta_t + \mathcal{L}_t \mu_t = \pi \mu_t + \nu_{t+1} \otimes \delta_{t+1}, \quad t = 0, \dots, T - 1. \quad (8)$$

The resulting GMP is equivalent to (7), since eliminating the measures  $\nu_t \otimes \delta_t \in \mathcal{M}(\mathbf{X}_t)_+$  using equalities (8) recovers constraint (7b). We now state the dual of this equivalent GMP, and show that the component of its solution for  $t = 0$  approximates the value function  $V_0^*(x)$  of (3) over the initial distribution  $\nu_0$ . Following the dualization process of Anderson & Nash (1987) for infinite-dimensional linear programs, we obtain (9). This is another infinite-dimensional linear program, in this case in the space of bounded continuous functions on  $\mathbf{X}_t$  for each time step  $t$ , denoted  $\mathcal{C}(\mathbf{X}_t)$ .

$$\theta^* := \max_{\{V_t \in \mathcal{C}(\mathbf{X}_t)\}_{t=0}^{T-1}} \int_{\mathbf{X}_0} V_0(x) d(\nu_0 \otimes \delta_0) \quad (9a)$$

$$\text{s.t. } l_t(x, u) - V_t(x) + V_t(f_t(x, u)) \geq 0, \quad \forall (x, u) \in \mathbf{C}_t, \quad t = 0, \dots, T - 1, \quad (9b)$$

$$V_{t+1}(x) \geq V_t(x), \quad \forall x \in \mathbf{X}_{t+1}, \quad t = 0, \dots, T - 2, \quad (9c)$$

$$H(x) \geq V_{T-1}(x), \quad \forall x \in \mathbf{X}_T. \quad (9d)$$

The integral  $d(\nu_0 \otimes \delta_0)$  reflects the initial state distribution  $\nu_0$  and initial value of the auxiliary state  $x_c$ , which is always 0. Thus the objective integrates  $V_0(x)$  over a “slice” of  $x$ -space at  $x_c = 0$ .

Note that each function  $V_t(x)$  in (9) is constrained at time steps  $t$  and  $t + 1$ , and that  $V_0(x), \dots, V_{T-1}(x), H(x)$  form a chain of coupled functions. Constraint (9b) is a relaxation of



the Bellman optimality condition for each pair of points  $(x, f(x, u))$  generated by an  $(x, u)$  pair in  $\mathbf{C}_t$ ; since  $x$  and  $f(x, u)$  have time index states  $x_c = t$  and  $x_c = t + 1$  respectively,  $V_t(x)$  is constrained in how it changes between time steps  $t$  and  $t + 1$ . Constraint (9c) upper-bounds  $V_t(x)$  by the value of the “next” value function  $V_{t+1}(x)$ , on  $x$  values with time index  $x_c = t + 1$ .

Since we have shown that the finite-horizon case is just a special case of the infinite-horizon formulation and Assumption 1 holds, Problem (9) is in fact the LP formulation of the dynamic programming problem for (2) and there is no duality gap between (7) and (9) (Hernández-Lerma & Lasserre, 2012, Theorem 6.3.8). It is straightforward to show<sup>3</sup> that for all feasible solutions of (9),  $V_t(x) \leq V_t^*(x)$  on  $\mathbf{X}_t$  for  $t = 0, \dots, T - 1$ .

### 3. Moment DDP

We now present an algorithm, termed Moment DDP, to find approximate solutions to (2) that are fitted to a probability distribution  $\nu_0$  of values of  $x_0$ . This is achieved by decomposing the multi-stage problems (7) and (9) into single stages and solving finite approximations of these problems. We first describe the backward recursion (Section 3.1) and forward simulation (Section 3.2), which are familiar concepts from existing DDP approaches, and then state the Moment DDP algorithm as a whole in Section 3.3.

Moment DDP uses the same stage-wise decomposition principle as conventional DDP, in that it simulates state trajectories in the forward simulation and then solves dual problems to generate lower-bounding functions in the backward recursion. However it is different in two important respects. First, the forward simulation consists of a sequence of single-stage problems over *moments* of the occupation measure instead of the point values or sampled uncertainty realizations used in conventional (S)DDP. These moments are a finite approximation of the original problem (7) over occupation measures. Second, the backward recursion, comprising dual SOS problems, generates *polynomial* rather than linear cuts, and under-approximates the value function most closely around the state distribution computed by the forward simulation. Analogously to conventional DDP, the cuts are used in the forward simulation as approximate cost-to-go functions to improve the candidate state trajectory. The sum of costs in the forward simulation (as estimated from the truncated moment series) represents an upper bound on the optimal cost attainable under the moment/SOS approximation, while the expected value (with respect to the given initial state distribution  $\nu_0$ ) of the value function obtained for  $t = 0$  represents a lower bound. The difference between the upper

---

<sup>3</sup>The optimal solutions  $\hat{V}_t(x)$  of (9) are *subsolutions* of the Bellman equation (3), i.e.  $\hat{V}_t(x) \leq l_t(x, u) + \hat{V}_t(f_t(x, u))$  on  $\mathbf{C}_t$  and  $\hat{V}_t(x) \leq \hat{V}_{t+1}(x)$  on  $\mathbf{X}_{t+1}$ , with  $\hat{V}_{T-1}(x) \leq H(x)$  on  $\mathbf{X}_T$ . As pointed out in Savorgnan et al. (2009), this leads to the fact that  $\hat{V}_0(x)$ , a maximizer, minimizes the quantity  $\int_{\mathbf{X}_0} |V_0^*(x) - \hat{V}_0(x)| d(\nu_0 \otimes \delta_0) = \int_{\mathbf{X}_0} V_0^*(x) - \hat{V}_0(x) d(\nu_0 \otimes \delta_0)$ .

and lower bounds is used as a convergence criterion for terminating the algorithm.

Alongside our general description of Moment DDP, we will use problem (7) with horizon  $T = 2$  to illustrate the decomposition into single-stage problems. The proof of convergence in Section 4 will also apply to the two-stage problem, with an induction argument used to extend this to arbitrary  $T$ .

### 3.1. The backward recursion

The backward recursion creates a new polynomial lower bounding function  $V_{t,z}(x)$  for the value function for  $t = T - 1, \dots, 0$ , analogous to the Benders cuts in conventional DDP. For each time step  $t$  and iteration  $z$ , the single-stage subproblem uses the following data:

- The lower-bounding functions already generated from earlier backward recursions (including the current one),  $V_{t+1,i}(x)$ ,  $i = 0, \dots, z$ , satisfying  $V_{t+1,i}(x) \leq V_{t+1}^*(x)$  for all  $x \in \mathbf{X}_{t+1}$ .
- The state measure  $\nu_t \otimes \delta_t \in \mathcal{M}(\mathbf{X}_t)$  from the last forward pass completed.

By the standard dynamic programming argument used in conventional DDP, the subproblem corresponds to the first stage of a version of problem (2.3) starting at step  $t$ :

$$\theta_t := \max_{V_{t,z} \in \mathcal{C}(\mathbf{X}_t)} \int_{\mathbf{X}_t} V_{t,z}(x) d(\nu_t \otimes \delta_t) \quad (10a)$$

$$\text{s.t. } l_t(x, u) - V_{t,z}(x) + V_{t,z}(f_t(x, u)) \geq 0, \quad \forall (x, u) \in \mathbf{C}_t, \quad (10b)$$

$$V_{t,z}(x) \leq \begin{cases} \max \{V_{t+1,0}(x), \dots, V_{t+1,z}(x)\}, & \forall x \in \mathbf{X}_{t+1}, \text{ if } t \in \{0, \dots, T-2\}, \\ H(x), & \forall x \in \mathbf{X}_{t+1}, \text{ if } t = T-1. \end{cases} \quad (10c)$$

This problem is illustrated in Fig. 1. Constraint (10b) restricts the change in the value function from time step  $t$  to time step  $t + 1$  according to the Bellman principle, and (10c) upper-bounds the value function at time step  $t + 1$  by the lower bounds already derived for stage  $t + 1$  of the problem.

Problem (10) is intractable owing to its infinite-dimensional decision space, but can be approximated using a polynomial parameterization of  $V_{t,z}(x)$ . We note that, except for the case  $t = T - 1$ , constraint (10c) is equivalent to

$$V_{t,z}(x) \leq y, \quad \forall (x, y) \in (\mathbf{X}_{t+1} \times \mathbb{R}) \cap \{(x, y) : y \geq V_{t+1,0}(x), \dots, y \geq V_{t+1,z}(x)\};$$

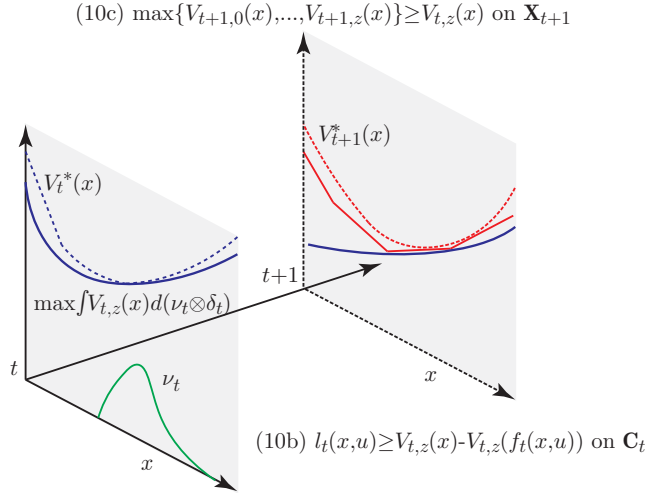


Figure 1: Illustration of the infinite-dimensional LP (10). The function  $V_{t,z}(x)$  (blue) is maximized over the state distribution  $\nu_t$  (green) at time step  $t$  subject to constraints (10b) and (10c), in order to approximate the value function  $V_t^*(x)$  (dashed blue). Constraint (10b) ensures  $V_{t,z}(x) \leq V_t^*(x)$  by limiting the values of  $V_{t,z}(x)$  at time step  $t$  such that transitions to step  $t+1$  incur costs that respect the Bellman inequality condition. Constraint (10c) bounds  $V_{t,z}(x)$  from above at  $t+1$  by the point-wise maximum (red) of lower-bounding functions computed in previous iterations for time step  $t+1$ . These are in turn under-approximations of the optimal value function (dashed red) at  $t+1$ .

this leads to the following SOS program for each time step  $t = T - 1, \dots, 0$ :

$$\theta_{t,z} := \max_{\mathbf{V}_{t,z}, \boldsymbol{\sigma}_{t,z}} \langle \mathbf{V}_{t,z}, \mathbf{q}_{t,z} \rangle \quad (11a)$$

$$\text{s.t. } l_t(x, u) - V_{t,z}(x) + V_{t,z}(f_t(x, u)) = \mathbf{Q}_k(\mathbf{C}_t), \quad (11b)$$

$$y - V_{t,z}(x) = \mathbf{Q}_k(\mathbf{Y}_{t+1,z}), \quad (11c)$$

$$\deg(V_{t,z})\kappa_t \leq 2k. \quad (11d)$$

The polynomial  $V_{t,z}(x)$  is represented by its vector of monomial coefficients  $\mathbf{V}_{t,z}$ , and the objective (10a) can thus be expressed as  $\langle \mathbf{V}_{t,z}, \mathbf{q}_{t,z} \rangle$ , where  $\mathbf{q}_{t,z}$  is a vector of moments of the state distribution  $\nu_t \otimes \delta_t$  returned at step  $t-1$  of the last forward pass completed.<sup>4</sup> The constraints (11b)-(11c) convert (10c)-(10b) into equality constraints using Putinar's Positivstellensatz (Putinar & Vasilescu, 1999) for compact semi-algebraic sets, in which the slacks are written as quadratic modules  $\mathbf{Q}_k(\mathbf{C}_t)$  and  $\mathbf{Q}_k(\mathbf{Y}_{t+1,z})$  that are non-negative by construction; see definition (1). The vector  $\boldsymbol{\sigma}_{t,z}$  contains all coefficients of the SOS polynomials introduced by the quadratic modules and is subject to additional LMI constraints not shown explicitly here, ensuring that the coefficients form valid SOS polynomials.<sup>5</sup> Constraints (11b) and (11c) are implemented by matching the coefficients of each

<sup>4</sup>In our proposed implementation, the first backward pass takes place before the first forward pass, hence the moments  $\mathbf{q}_{t,0}$  of the state trajectory must be initialized. The uniform distribution may be an appropriate choice when no information about the optimal state trajectory is available *a priori*.

<sup>5</sup>More precisely,  $\boldsymbol{\sigma}_{t,z}$  is a concatenation of the vectorizations of the matrix of coefficients  $\mathbf{P}$ , as described in Section 1.1, for all of the SOS polynomials  $\sigma_i$  within the quadratic modules  $\mathbf{Q}_k(\mathbf{C}_t)$  and  $\mathbf{Q}_k(\mathbf{Y}_{t+1,z})$ .

monomial on either side, i.e., using linear equality constraints linking the elements of  $\mathbf{V}_{t,z}$  and  $\boldsymbol{\sigma}_{t,z}$ . Since the definition of the quadratic module limits the degree of polynomial used to  $2k$ , and polynomials  $V_t$  are composed with polynomials  $f_t(x, u)$  in (11b), the degree of  $V_t$  must be restricted by (11d), where  $\kappa_t := \max_{i=1, \dots, n_x} (\deg(f_{t,i}(x, u)))$  is the highest-order polynomial found in the dynamics.

In constraint (11c) we introduced a new epigraph set  $\mathbf{Y}_{t+1,z}$ . For each time step  $t = T, \dots, 1$ ,  $\mathbf{Y}_{t,z}$  is defined by the  $z$  lower-bounding functions generated so far for that time step, and an upper bound  $\bar{y}$  on the epigraph variable  $y$ :

$$\mathbf{Y}_{t,z} := \begin{cases} \{(x, y) : x \in \mathbf{X}_t; y \in \mathbb{R}; y \leq \bar{y}; y \geq V_{t,i}(x), i = 0, \dots, z\}, & t = 1, \dots, T-1, \\ \{(x, y) : x \in \mathbf{X}_t; y \in \mathbb{R}; y \leq \bar{y}; y \geq H(x)\}, & t = T. \end{cases}$$

The parameter  $\bar{y} \in \mathbb{R}$  must be chosen in advance and ensures that, in combination with at least one lower-bounding value function, the epigraph set is compact.<sup>6</sup>

Since the function parameterization in (11) is contained in the feasible set of (10), it follows that  $\theta_{t,z}$  is upper bounded by the optimal value of (10). The approximation accuracy is known to improve as  $k$  increases (Korda et al., 2017).

Returning to the two-stage example, the backward recursion at iteration  $z$  for  $t = 1$  is a SOS problem of type (11):

$$\theta_{1,z} = \max_{\mathbf{V}_{1,z}, \boldsymbol{\sigma}_{1,z}} \langle \mathbf{V}_{1,z}, \mathbf{q}_{1,z} \rangle \quad (12a)$$

$$\text{s.t. } l_1(x, u) - V_{1,z}(x) + V_{1,z}(f_1(x, u)) = \mathbf{Q}_k(\mathbf{C}_1), \quad (12b)$$

$$H(x) - V_{1,z}(x) = \mathbf{Q}_k(\mathbf{X}_2), \quad (12c)$$

$$\deg(V_{1,z})\kappa_1 \leq 2k, \quad (12d)$$

We add the optimal solution  $\hat{V}_{1,z}$  of (12) to the epigraph set  $\mathbf{Y}_{1,z}$  and solve a SOS problem for

---

<sup>6</sup>We acknowledge that this is not an epigraph in the strict sense of the word, since it includes an upper bound on  $y$ . The value of  $\bar{y}$  used to define  $\mathbf{Y}_t$  must be larger than the greatest sum of costs from time steps  $t$  to  $T$  that can occur in any state trajectory. Since the state-input set is compact, the stage cost is bounded, and the number of stages is finite, it is generally straightforward to obtain such a bound.

$t = 0$ :

$$\theta_{0,z} = \max_{\mathbf{V}_{0,z}, \boldsymbol{\sigma}_{0,z}} \langle \mathbf{V}_{0,z}, \mathbf{q}_{0,z} \rangle \quad (13a)$$

$$\text{s.t. } l_0(x, u) - V_{0,z}(x) + V_{0,z}(f_0(x, u)) = Q_k(\mathbf{C}_0), \quad (13b)$$

$$y - V_{0,z}(x) = Q_k(\mathbf{Y}_{1,z}), \quad (13c)$$

$$\deg(V_{0,z})\kappa_0 \leq 2k. \quad (13d)$$

In the Moment DDP algorithm described in Section 3.3, the lower bound value  $\theta_{LB,z} = \theta_{0,z}$  is used in the termination criterion.

### 3.2. The forward simulation

The forward simulation finds, for each  $t = 1, \dots, T$ , an approximate solution to a single stage of the GMP (7), in which the state occupation measure  $\nu_t$  is inherited from the previous step's solution, and the cost-to-go is under-approximated by the lower-bounding functions  $V_{t+1,i}(x)$  generated in the backward recursions completed so far:

$$\rho_t := \min_{\mu_t, \nu_{t+1}} \int_{\mathbf{C}_t} l_t(x, u) d\mu_t + \int_{\mathbf{X}_{t+1}} \max_{i=0, \dots, z-1} V_{t+1,i}(x) d(\nu_{t+1} \otimes \delta_{t+1}), \quad (14a)$$

$$\text{s.t. } \nu_t \otimes \delta_t + \mathcal{L}\mu_t = \pi\mu_t + \nu_{t+1} \otimes \delta_{t+1}, \quad (14b)$$

$$\mu_t \in \mathcal{M}(\mathbf{C}_t)_+, \nu_{t+1} \otimes \delta_{t+1} \in \mathcal{M}(\mathbf{X}_{t+1})_+. \quad (14c)$$

As this problem is infinite-dimensional and therefore intractable, the approximation used is an optimization over a finite vector of moments of the state-action occupation measure  $\mu_t$  at time step  $t$ , and the state occupation measure  $\nu_{t+1}$  at time step  $t + 1$ .

We now explain how this finite-moment approximation of (14) is represented. Let  $\mu_{t,z}$  be the state-action occupation measure on  $\mathbf{C}_t$  for a single time step  $t$  at iteration  $z$ , and let  $m_{t,z}^{\alpha\gamma}$  be its  $(\alpha, \gamma)$  moment for non-negative integer vectors  $\alpha \in \mathbb{N}^{n_x}$  and  $\gamma \in \mathbb{N}^{n_u}$ , defined by

$$m_{t,z}^{\alpha\gamma} := \int_{\mathbf{C}_t} x^\alpha u^\gamma d\mu_{t,z}. \quad (15)$$

Following convention from related literature, the vector-valued exponents are interpreted as  $x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \dots x_{n_x}^{\alpha_{n_x}}$  and  $u^\gamma = u_1^{\gamma_1} u_2^{\gamma_2} \dots u_{n_u}^{\gamma_{n_u}}$ , with  $\sum_{i=1}^{n_x} \alpha_i + \sum_{i=1}^{n_u} \gamma_i \leq 2k$ .

We use the epigraph set  $\mathbf{Y}_{t+1,z-1}$  created in the previous backward recursion to accommodate the maximum in the second term of (14a). For each time step  $t = 1, \dots, T$  and iteration  $z$ , we

define the moments of the augmented state measure  $\nu_{t,z} \otimes \delta_t$  supported on the epigraph set  $\mathbf{Y}_{t,z-1}$ :

$$q_{t,z}^{\alpha\eta} := \int_{\mathbf{Y}_{t,z-1}} x^\alpha y^\eta d(\nu_{t,z} \otimes \delta_t), \quad (16)$$

where  $x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \dots x_{n_x}^{\alpha_{n_x}}$  and  $y$  is the scalar epigraph variable used in the definition of  $\mathbf{Y}_{t,z-1}$ , with  $\sum_{i=1}^{n_x} \alpha_i + \eta \leq 2k$ . We collect these moments into vectors  $\mathbf{m}_{t,z}$  and  $\mathbf{q}_{t,z}$  respectively for each iteration  $z$  of the DDP algorithm. The number of elements in  $\mathbf{m}_{t,z}$  is combinatorial, given by  $n_{\mathbf{m}} = \binom{n_x + n_u + k}{k}$ . Similarly, the vector  $\mathbf{q}_{t,z}$  has size  $n_{\mathbf{q}} = \binom{n_x + 1 + k}{k}$ . The moments  $(q_{t+1,z}^{\alpha 0})$  of the state distribution at time step  $t$ , recalling that the superscript 0 signifies that  $y$  is excluded, are used as initial conditions in time step  $t + 1$ .

As with conventional DDP, the forward problem in Moment DDP for each stage  $t = 0, \dots, T-1$  is dual to the backward problem (11). It takes the form of a semidefinite program (SDP) in terms of the moments (up to degree  $2k$ ) of  $\mu_{t,z}$  and  $\nu_{t+1,z} \otimes \delta_{t+1}$ :

$$\rho_{t,z} := \min_{\mathbf{m}_{t,z}, \mathbf{q}_{t+1,z}} L_{\mathbf{m}_{t,z}}(l_t) + L_{\mathbf{q}_{t+1,z}}(y) \quad (17a)$$

$$\text{s.t. } L_{\mathbf{m}_{t,z}}\left(x^\alpha - f_t(x, u)^\alpha\right) + q_{t+1,z}^{\alpha 0} = q_{t,z}^{\alpha 0}, \quad \alpha \in \mathbb{N}^{n_x}, \sum_{i=1}^{n_x} \alpha_i \leq \lfloor 2k/\kappa_t \rfloor, \quad (17b)$$

$$M_{k-d_{g_{t,j}}}(g_{t,j} \mathbf{m}_{t,z}) \succeq 0, \quad j = 1, \dots, N_{g,t}, \quad (17c)$$

$$M_{k-d_{v_{t+1,s}}}(v_{t+1,s} \mathbf{q}_{t+1,z}) \succeq 0, \quad s = 1, \dots, N_{g,t} + z + 1, \quad (17d)$$

$$M_k(\mathbf{m}_{t,z}) \succeq 0, M_k(\mathbf{q}_{t+1,z}) \succeq 0, \quad (17e)$$

where  $f(x, u)^\alpha$  is shorthand for  $f_1(x, u)^{\alpha_1} f_2(x, u)^{\alpha_2} \dots f_{n_x}(x, u)^{\alpha_{n_x}}$ .

In brief, the objective (17a) approximates the expected cost  $\mathbf{E}_{\mu_{t,z}}(l_t) + \mathbf{E}_{\nu_{t+1,z}}(y)$  as a linear combination of moments of  $\mu_{t,z}$  and  $\nu_{t+1,z}$ . The constraint (17b) represents a truncated form of the infinite-dimensional constraint (7b), which means that the state update equation is transformed into a set of linear equalities on the moments of the state-action measure  $\mu_{t,z}$  and state measure  $\nu_{t+1,z} \otimes \delta_{t+1}$ . Constraints (17c) and (17d) jointly represent ‘‘moment relaxations’’ of the support constraints (14c) on  $\mu_{t,z}$  and  $\nu_{t+1,z}$ , and constraints (17e) are used to ensure that the moment vectors are compatible with valid measures. We now explain the elements of (17) in detail.

The operator  $L_{\mathbf{m}_{t,z}} : \mathbb{R}[x, u] \rightarrow \mathbb{R}$  is a linear mapping associated with a measure  $\mu_{t,z}$  acting on a polynomial  $h \in \mathbb{R}[x, u]$ :

$$L_{\mathbf{m}_{t,z}}(h) := \sum_{\alpha\gamma} h^{\alpha\gamma} m_{t,z}^{\alpha\gamma}, \quad (18)$$

where  $m_{t,z}^{\alpha\gamma}$  are the moments of  $\mu_{t,z}$  as defined in (15) and  $h^{\alpha\gamma}$  represents the polynomial coefficient of  $x^\alpha u^\gamma$ , with vectors  $\alpha$  and  $\gamma$  interpreted in the same manner as for (15). Analogously,  $L_{\mathbf{q}_{t,z}} :$

$\mathbb{R}[x, y] \rightarrow \mathbb{R}$  is a linear mapping associated with the moments defined in (16):

$$L_{\mathbf{q}_{t,z}}(h) := \sum_{\alpha\eta} h^{\alpha\eta} q_{t,z}^{\alpha\eta}. \quad (19)$$

These operators are used to approximate the expected cost (14a) in terms of moments, so that  $\mathbf{E}_{\mu_{t,z}}(l_t) + \mathbf{E}_{\nu_{t+1,z}}(y) = \int_{\mathbf{C}_t} l_t d\mu_{t,z} + \int_{\mathbf{Y}_{t+1,z-1}} y d\nu_{t+1,z}$  becomes  $L_{\mathbf{m}_{t,z}}(l_t) + L_{\mathbf{q}_{t+1,z}}(y) = \sum_{\alpha\gamma} l_t^{\alpha\gamma} m_{t,z}^{\alpha\gamma} + q_{t+1,z}^{01}$ .

The same linear operator is used in constraint (17b) to enforce consistency of the change in moments from  $q_{t,z}^{\alpha 0}$ , which are fixed data from the previous stage, and  $q_{t+1,z}^{\alpha 0}$  under the dynamics.

The standard *moment matrices*  $M_k(\mathbf{m}_{t,z})$  and  $M_k(\mathbf{q}_{t,z})$  of degree  $k$  in (17e); and the localizing matrices  $M_{k-d_{g_{t,j}}}(g_{t,j}\mathbf{m}_{t,z})$  and  $M_{k-d_{v_{t+1,s}}}(v_{t+1,s}\mathbf{q}_{t+1,z})$  in (17c)-(17d) enforce a condition that ensures the generic vectors of moments are consistent with finite Borel measures on compact set.<sup>7</sup> They are derived by applying the linear mappings  $L_{\mathbf{m}_{t,z}}$  and  $L_{\mathbf{q}_{t,z}}$  to the square of any polynomial  $h$  of degree  $k$ :

$$L_{\mathbf{m}_{t,z}}(h^2) = \mathbf{h}^\top M_k(\mathbf{m}_{t,z})\mathbf{h} \geq 0, \quad L_{\mathbf{q}_{t,z}}(h^2) = \mathbf{h}^\top M_k(\mathbf{q}_{t,z})\mathbf{h} \geq 0, \quad (20)$$

where  $\mathbf{h}$  is the vector of coefficients of  $h$ . Thus, the moment matrix, which is linear in the elements of  $\mathbf{m}_{t,z}$  or  $\mathbf{q}_{t,z}$ , is constrained to be a symmetric positive semi-definite matrix; the two constraints of (17e) are therefore standard LMI constraints.

For notational convenience, we now write the constraints defining the epigraph set  $\mathbf{Y}_{t,z-1}$  as  $v_{t,s}(x, y) \geq 0, s = 1, \dots, N_{g_x} + z + 1$ . The *localizing matrices* (17c) and (17d), which are also standard in moment problems, enforce a moment relaxation of the support constraints  $g_{t,j}(x, u) \geq 0$  (which define set  $\mathbf{C}_t$ ) and  $v_{t+1,s}(x, u) \geq 0$  (which define set  $\mathbf{Y}_{t+1,z}$ ). These are positive semi-definite and of the form

$$\begin{aligned} L_{\mathbf{m}}(g_{t,j}h^2) &= \mathbf{h}^\top M_{k-d_{g_{t,j}}}(g_{t,j}\mathbf{m}_{t,z})\mathbf{h} \geq 0, \\ L_{\mathbf{q}}(v_{t+1,s}h^2) &= \mathbf{h}^\top M_{k-d_{v_{t+1,s}}}(v_{t+1,s}\mathbf{q}_{t+1,z})\mathbf{h} \geq 0, \end{aligned} \quad (21)$$

where  $d_{g_{t,j}} = \lceil \deg(g_{t,j})/2 \rceil$  and  $d_{v_{t+1,s}} = \lceil \deg(v_{t+1,s})/2 \rceil$ .

Thus, (17) is a relaxation of (14), in which each of the constraints has been enforced on only a finite series of moments of  $\mu_{t,z}$  and  $\nu_{t+1,z}$ . It therefore attains a lower optimal value than (14); recall that its dual, the SOS program (11), is a restriction of the infinite-dimensional LP shown in Fig. 1 and has a corresponding lower optimal value.

We now state a known result concerning the value of relaxation (17) as the order  $k$  is increased:

---

<sup>7</sup>In fact, this is a relaxation of the consistency condition, which is only guaranteed to hold for an infinite series of moments (Lasserre, 2014, Theorem 3.8).

**Lemma 1.** *Let Assumption 1 hold, and let the feasible set  $\mathbf{C}_t$  and epigraph set  $\mathbf{Y}_{t+1,z}$  satisfy Putinar's condition <sup>8</sup>. If  $\rho_t$  is the optimal solution of the infinite-dimensional GMP (14) at time step  $t$ , then as  $k \rightarrow \infty$  the optimal value of (17) approaches  $\rho_t$  asymptotically from below.*

*Proof.* Following Theorem 1 in Savorgnan et al. (2009), one can show that  $\rho_{t,z}$ , when evaluated for increasing values of the relaxation degree  $k$  used in constraint (17b), is a monotone non-decreasing sequence converging to  $\rho_t$ . This makes use of Putinar's Positivstellensatz, and the fact that measures on compact sets are uniquely determined by their infinite sequence of moments.  $\square$

In case of example (7) with  $T = 2$ , we start the forward simulation by solving a moment relaxation of degree  $2k$  for  $t = 0$ , a SDP of type (17) that includes the epigraph set  $\mathbf{Y}_{1,z-1}$  built from all the value function under-approximators  $\{V_{1,i}(x)\}_{i=0}^{z-1}$ :

$$\rho_{0,z} = \min_{\mathbf{m}_{0,z}, \mathbf{q}_{1,z}} L_{\mathbf{m}_{0,z}}(l_0) + L_{\mathbf{q}_{1,z}}(y) \quad (22a)$$

$$\text{s.t. } L_{\mathbf{m}_{0,z}}\left(x^\alpha - f_0(x, u)^\alpha\right) + q_{1,z}^{\alpha 0} = q_0^{\alpha 0}, \quad \alpha \in \mathbb{N}^{n_x}, \sum_{i=1}^{n_x} \alpha_i \leq \lfloor 2k/\kappa_0 \rfloor, \quad (22b)$$

$$M_{k-d_{g_{0,j}}}(g_{0,j}\mathbf{m}_{0,z}) \succeq 0, \quad j = 1, \dots, N_{g,t}, \quad (22c)$$

$$M_{k-d_{v_{1,s}}}(v_{1,s}\mathbf{q}_{1,z}) \succeq 0, \quad s = 1, \dots, N_{g_x,t} + z + 1, \quad (22d)$$

$$M_k(\mathbf{m}_{0,z}) \succeq 0, M_k(\mathbf{q}_{1,z}) \succeq 0, \quad (22e)$$

where we note that moments  $q_0^{\alpha 0}$  (defined in the same way as (16)) are fixed data derived from the initial state distribution  $\nu_0 \otimes \delta_0$ . If (22) and (13) are strictly feasible, there is no duality gap and  $\rho_{0,z} = \theta_{LB,z}$ .

The primal problem for  $t = 1$  is a moment relaxation with the optimal solution  $\hat{\mathbf{q}}_{1,z}$  of (22) as input data:

$$\rho_{1,z} = \min_{\mathbf{m}_{1,z}, \mathbf{q}_{2,z}} L_{\mathbf{m}_{1,z}}(l_1) + L_{\mathbf{q}_{2,z}}(H) \quad (23a)$$

$$\text{s.t. } L_{\mathbf{m}_{1,z}}\left(x^\alpha - f_1(x, u)^\alpha\right) + q_{2,z}^{\alpha 0} = \hat{q}_{1,z}^{\alpha 0}, \quad \alpha \in \mathbb{N}^{n_x}, \sum_{i=1}^{n_x} \alpha_i \leq \lfloor 2k/\kappa_1 \rfloor, \quad (23b)$$

$$M_{k-d_{g_{1,j}}}(g_{1,j}\mathbf{m}_{1,z}) \succeq 0, \quad j = 1, \dots, N_{g,t}, \quad (23c)$$

$$M_{k-d_{v_{2,s}}}(v_{2,s}\mathbf{q}_{2,z}) \succeq 0, \quad s = 1, \dots, N_{g_x,t}, \quad (23d)$$

$$M_k(\mathbf{m}_{1,z}) \succeq 0, M_k(\mathbf{q}_{2,z}) \succeq 0, \quad (23e)$$

---

<sup>8</sup>One can ensure that the sets  $\mathbf{C}_t$  and  $\mathbf{Y}_{t+1,z}$  satisfy Putinar's condition (see Definition 3.4 in Lasserre et al. (2008)) by including an additional ball constraint. For instance one can add  $g_{N_g+1}(x, u) = R^2 - \sum_i^{n_x} x_i^2 - \sum_i^{n_u} u_i^2 \geq 0$  with  $R \in \mathbb{R}$  to the definition of  $\mathbf{C}_t$ . The assumption that  $\mathbf{C}_t$  and  $\mathbf{Y}_{t+1,z}$  are both compact makes it straightforward to determine such an  $R$  in most cases.



---

**Algorithm 1** Moment DDP

---

**Input:** Horizon  $T$ , functions  $f_t(x, u)$ ,  $l_t(x, u)$ ,  $H(x)$ ,  $g_{t,j}(x, u)$ , tolerance  $\epsilon$ , initial moments  $\mathbf{q}_{t,0}$

**Output:** Upper bound  $\rho_{UB,z}$ , lower bound  $\theta_{LB,z}$ , epigraph sets  $\mathbf{Y}_{t,z}$ , trajectory moments  $\mathbf{q}_{t,z}$

**Indices:** Iteration  $z$ , time step  $t$

---

```
1:  $z \leftarrow 0$ 
2: Create set  $\mathbf{Y}_{T,0}$  parameterized by  $H(x)$ 
3: for  $t = T - 1, \dots, 1$  do ▷ Initial backward recursion: Section 3.1
4:   Solve (11) to obtain  $V_{t,0}(x)$ 
5:   Create set  $\mathbf{Y}_{t,0}$  parameterized by  $V_{t,0}(x)$ .
6: repeat ▷ Repeat procedure until predefined tolerance  $\epsilon$  is achieved
7:    $z \leftarrow z + 1$ 
8:   for  $t = 0, \dots, T - 1$  do ▷ Forward simulation: Section 3.2
9:     Solve (17) to obtain state moments  $\mathbf{q}_{t+1,z}$ 
10:    Compute  $\rho_{UB,z} = \sum_{t=0}^{T-1} L_{\hat{\mathbf{m}}_{t,z}}(l_t) + L_{\hat{\mathbf{q}}_{T,z}}(H)$  (optimal values of (17))
11:    for  $t = T - 1, \dots, 0$  do ▷ Backward recursion: Section 3.1
12:      Solve (11) to obtain  $V_{t,z}(x)$ 
13:       $\mathbf{Y}_{t,z} \leftarrow \mathbf{Y}_{t,z-1} \cap \{(x, y) : y \geq V_{t,z}(x)\}$ 
14:      Set  $\theta_{LB,z} = \theta_{0,z}$  (optimal value of (11) for  $t = 0$ )
15: until  $\rho_{UB,z} - \theta_{LB,z} < \epsilon$ 
```

---

The updated moments  $\hat{\mathbf{q}}_{1,z}$  computed by (22) can then be used in a subsequent backward recursion to generate a new approximate value function in the backward recursion. If (12) and (23) are strictly feasible, there is no duality gap and  $\rho_{1,z} = \theta_{1,z}$ . Using the optimal values of (22) and (23), we define the upper bound as  $\rho_{UB,z} = L_{\hat{\mathbf{m}}_{0,z}}(l_0) + L_{\hat{\mathbf{m}}_{1,z}}(l_1) + L_{\hat{\mathbf{q}}_{2,z}}(H)$  for use in the termination criterion of the algorithm described below.

### 3.3. Moment DDP algorithm

Moment DDP is stated formally in Algorithm 1, and we now remark on some aspects of its implementation.

Firstly, we note that the degree of the under-approximating value functions can in practice be chosen to be relatively low, since a single function need not be an active bound over the entire state space. This is illustrated in Fig. 7 in the Appendix, which shows the lower-bounding functions generated by a sequence of six backward recursions for the single storage example of Section 5, alongside the approximation generated by discretized DP.

Secondly, it can be attractive to preserve convexity of the lower-bounding functions added in the backward recursion, in order to reduce the cost of computing forward control actions. Following the approach of Lasserre & Thanh (2013), convexity can be imposed on polynomials by constraining the Hessian of the value function in (11) and adding additional variables to the primal (17). This may of course cause an additional reduction in the tightness of the value function approximation.

Thirdly, if the problem input data remains constant over multiple time steps  $t$ , it becomes rela-

tively straightforward to adapt a single stage of the forward and backward recursions in Algorithm 1 to span these steps. In this case, one can use a single polynomial to approximate a value function over the relevant interval on the time coordinate  $x_c$ . Value functions can then be extracted for a time step within a stage by setting  $x_c$  to the relevant value. Throughout this paper, however, we maintain equivalence between problem stages and time steps  $t$  in (2) for clarity of notation.

### 3.4. Extension to stochastic dynamics

The Moment DDP approach can be extended to stochastic polynomial dynamics, in which the state update is described by a function  $f_t(x, u, w)$ , without increasing the computational complexity significantly. Vector  $w$  denotes an independent disturbance following the distribution  $\omega_t$  supported on  $\mathbf{W}_t$ , of which the statistical moments can be computed; and entering polynomially into the state update.

If these conditions hold, moment and SOS relaxations can be formulated using the same procedure described for generic optimal control problems in Savorgnan et al. (2009). Specifically, the operator  $\mathcal{L}_t$  is replaced by a new linear operator  $\tilde{\mathcal{L}}_t : \mathcal{M}(\mathbf{C}_t)_+ \rightarrow \mathcal{M}(\mathbf{X}_{t+1})_+$  defined as

$$\pi\mu_{t+1}(A) = \tilde{\mathcal{L}}_t\mu_t(A) := \int_{\mathbf{C}_t} \int_{\mathbf{W}_t} 1_A(f_t(x, u, w)) d\omega_t d\mu_t, \quad (24)$$

for all Borel sets  $A$  of  $\mathbf{X}_{t+1}$ . For simplicity of exposition, however, we have excluded stochastic dynamics from the derivations and numerical examples in the present paper, and the only uncertainty we include arises from the initial state distribution.

## 4. Convergence properties

In this section, we analyze the convergence of Algorithm 1 using an instance of the GMP (7) with  $T = 2$ , and argue subsequently that the results extend to longer horizons. Lemma 2 states that if the upper bound is strictly larger than the lower bound, (a relaxation of) the epigraph set strictly tightens from one iteration to the next. Lemmas 3 and 4 bound the values of  $\theta_{LB,z}$  and  $\rho_{UB,z}$  used in the termination criterion. Finally, Theorem 2 concludes that the Moment DDP approach converges in finite iterations for any tolerance  $\epsilon > 0$ .

To facilitate these derivations, we say the moments  $\hat{\mathbf{q}}_{1,z}$  computed by the SDP relaxation (22) are elements of the *relaxed* epigraph set, which we define as

$$\begin{aligned} \tilde{\mathbf{Y}}_{1,z} := \{ \mathbf{q}_{1,z} \in \mathbb{R}^{n_{\mathbf{a}}} : & M_k(\mathbf{q}_{1,z}) \succeq 0; \\ & M_{k-d_{g_{j,1}}}(g_{j,1}\mathbf{q}_{1,z}) \succeq 0, \quad j = 1, \dots, N_{g_x}; \\ & M_{k-d_{V_{1,i}}}((y - V_{1,i})\mathbf{q}_{1,z}) \succeq 0, \quad i = 0, \dots, z - 1; \\ & M_{k-d_y}((\bar{y} - y)\mathbf{q}_{1,z}) \succeq 0 \}. \end{aligned} \quad (25)$$

**Lemma 2.** *If  $\theta_{LB,z} < \rho_{UB,z}$  at some iteration  $z$ , the relaxed epigraph set strictly tightens, i.e.  $\tilde{\mathbf{Y}}_{1,z} \subset \tilde{\mathbf{Y}}_{1,z-1}$ . Moreover  $\theta_{LB,z+1} \geq \theta_{LB,z}$ .*

*Proof.* Let  $(\hat{\mathbf{m}}_{0,z}, \hat{\mathbf{q}}_{1,z}, \hat{\mathbf{m}}_{1,z}, \hat{\mathbf{q}}_{2,z})$  be a solution computed by the moment relaxations (22) and (23) during the forward simulation. Let  $\langle \hat{\mathbf{V}}_{1,z}, \hat{\mathbf{q}}_{1,z} \rangle$  be the optimal value of the backward recursion program (12). By definitions of  $\theta_{LB,z}$  and  $\rho_{UB,z}$ , and strong duality between the second-stage problems (23) and (12), we have

$$\begin{aligned} \theta_{LB,z} &= L_{\hat{\mathbf{m}}_{0,z}}(l_0) + L_{\hat{\mathbf{q}}_{1,z}}(y) \quad \text{and} \quad \rho_{UB,z} = L_{\hat{\mathbf{m}}_{0,z}}(l_0) + L_{\hat{\mathbf{m}}_{1,z}}(l_1) + L_{\hat{\mathbf{q}}_{2,z}}(H) \\ &= L_{\hat{\mathbf{m}}_{0,z}}(l_0) + \langle \hat{\mathbf{V}}_{1,z}, \hat{\mathbf{q}}_{1,z} \rangle. \end{aligned}$$

Thus,  $\theta_{LB,z} < \rho_{UB,z}$  implies  $L_{\hat{\mathbf{q}}_{1,z}}(y) < \langle \hat{\mathbf{V}}_{1,z}, \hat{\mathbf{q}}_{1,z} \rangle$ . For the next iteration, we add the LMI constraint  $M_{k-d_{\hat{V}_{1,z}}}(y - \hat{V}_{1,z})\mathbf{q}_{1,z} \succeq 0$  to  $\tilde{\mathbf{Y}}_{1,z}$ , and it is straightforward to show (see (Molzahn & Hiskens, 2015, eq. (14)) for a similar example) that the first diagonal element of this matrix is the linear expression  $L_{\mathbf{q}_{1,z}}(y) - L_{\mathbf{q}_{1,z}}(\hat{V}_{1,z}) = q_{1,z}^{01} - \langle \hat{\mathbf{V}}_{1,z}, \mathbf{q}_{1,z} \rangle$ . Because this is on the diagonal of a matrix that is constrained to be positive semidefinite, it must be nonnegative. Thus the new set  $\tilde{\mathbf{Y}}_{1,z}$  contains the constraint that  $L_{\mathbf{q}_{1,z}}(y) \geq \langle \hat{\mathbf{V}}_{1,z}, \mathbf{q}_{1,z} \rangle$ .

The old moment vector  $\hat{\mathbf{q}}_{1,z}$  is now infeasible at iteration  $z + 1$ . Thus,  $\tilde{\mathbf{Y}}_{1,z}$  must be a strict subset of  $\tilde{\mathbf{Y}}_{1,z-1}$ . Since (22) is a minimization over a subset of the previous feasible set, the cost attained may be no lower than at the previous iteration.  $\square$

Let  $\rho_k^*$  be the optimal value of the undecomposed moment relaxation of (7) with  $T = 2$ :

$$\begin{aligned} \rho_k^* &:= \min_{\mathbf{m}_0, \mathbf{q}_1, \mathbf{m}_1, \mathbf{q}_2} L_{\mathbf{m}_0}(l_0) + L_{\mathbf{m}_1}(l_1) + L_{\mathbf{q}_2}(H) \\ &\text{s.t.} \quad (22\text{b})\text{--}(22\text{e}), (23\text{b})\text{--}(23\text{e}) \end{aligned} \tag{26}$$

The following lemmas bound the possible values of the lower and upper bounds returned by Algorithm 1:

**Lemma 3.** *At any iteration  $z$ ,  $\rho_{UB,z} \geq \rho_k^*$ , the optimal value of the undecomposed moment relaxation (26).*

*Proof.* Let  $(\hat{\mathbf{m}}_{0,z}, \hat{\mathbf{q}}_{1,z}, \hat{\mathbf{m}}_{1,z}, \hat{\mathbf{q}}_{2,z})$  be a solution computed by the moment relaxations (22) and (23) during the forward simulation. Examination of the constraints of (26) shows that this is a feasible but in general suboptimal solution, thus  $\rho_{UB,z} = L_{\hat{\mathbf{m}}_0}(l_0) + L_{\hat{\mathbf{m}}_1}(l_1) + L_{\hat{\mathbf{q}}_2}(H) \geq \rho_k^*$ .  $\square$

**Lemma 4.** *At any iteration  $z$ ,  $\theta_{LB,z} \leq \rho^*$ , the optimal value of the GMP (7).*

*Proof.* By inserting the optimal solution  $V_1^*(x)$  of the undecomposed LP (9) with  $T = 2$  into the epigraph of the first stage LP (10), it can be seen that the optimal value  $\theta_t$  of (10) is bounded

from above by the optimal values  $\theta^* = \rho^*$  of the undecomposed LPs (7) and (9) with  $T = 2$ . Since the SOS approximation (13) of the first stage LP 10 is more restricted, we have  $\theta_{LB,z} \leq \rho^*$ .  $\square$

Finally, we can state the following result concerning the convergence of Algorithm 1:

**Theorem 2.** *Given a tolerance  $\epsilon > 0$ , Algorithm 1 attains  $\rho_{UB,z} - \theta_{LB,z} \leq \epsilon$  in a finite number of iterations when applied to GMP (7) with  $T = 2$ . Moreover, the sequence  $\{\theta_{LB,z}\}$  converges to a value  $\hat{\theta}_{LB}$  satisfying  $\rho_k^* \leq \hat{\theta}_{LB} \leq \rho^*$ , where  $\rho^*$  is the optimal value of the original multi-stage GMP (7) and  $\rho_k^*$  is the optimal value of its degree- $k$  moment relaxation (26).*

*Proof.* Let  $\{\rho_{UB,z}\}$  and  $\{\theta_{LB,z}\}$  be sequences over  $z$  iterations. Assumption 1 (continuity and compactness) implies that the sequences  $\{\rho_{UB,z}\}$  and  $\{\theta_{LB,z}\}$  are bounded. From Lemma 2,  $\{\theta_{LB,z}\}$  is a monotonically increasing sequence. By the monotone convergence theorem,  $\{\theta_{LB,z}\}$  converges to some accumulation point  $\hat{\theta}_{LB}$ . By the Bolzano-Weierstrass theorem, there is a subsequence  $\{\rho_{UB,i}\}$  that converges to an accumulation point  $\hat{\rho}_{UB}$ . Every subsequence of a convergent sequence is also convergent, so we have  $\lim_{i \rightarrow \infty} \theta_{LB,i} = \hat{\theta}_{LB}$ .

Let  $\tilde{\mathbf{X}}_1 := \{\mathbf{x}_1 \in \mathbb{R}^{n_x} : M_k(\mathbf{x}_1) \succeq 0; M_{k-d_{g_{j,1}}}(g_{j,1}\mathbf{x}_1) \succeq 0, j = 1, \dots, N_{g_x}\}$  be the relaxed state space, where  $\mathbf{x}_1$  is defined in the same manner as  $\mathbf{q}_1$  but without the epigraph variable  $y$ . For any state moment vector  $\mathbf{x}_1 \in \tilde{\mathbf{X}}_1$ , a sequence  $\{y_{\mathbf{x}_1,z}^*\}$  can be constructed by solving the following optimization problem at each iteration  $z$ :

$$y_{\mathbf{x}_1,z}^* := \min_{\mathbf{q}_1} L_{\mathbf{q}_1}(y) \quad (27a)$$

$$\text{s.t. } \mathbf{q}_1 \in \tilde{\mathbf{Y}}_{1,z} \quad (27b)$$

$$q_1^{\alpha 0} = x_1^\alpha, \quad \alpha \in \mathbb{N}^{n_x}, \quad \sum_{i=1}^{n_x} \alpha_i \leq \lfloor 2k/\kappa_1 \rfloor. \quad (27c)$$

In words,  $y_{\mathbf{x}_1,z}^*$  is the relaxed epigraph value evaluated for the state moments  $x_1^\alpha$  with respect to the relaxed epigraph set  $\tilde{\mathbf{Y}}_{1,z}$ . For each  $\mathbf{x}_1$  in  $\tilde{\mathbf{X}}_1$ , the sequence  $\{y_{\mathbf{x}_1,z}^*\}$  is monotonically increasing (see Lemma 2) and bounded, and thus by the monotone convergence theorem, the limit  $\{y_{\mathbf{x}_1,z}^*\} \rightarrow y_{\mathbf{x}_1,\infty}^*$  always exists. At no iteration  $z$  of the algorithm can the backward recursion generate another  $\hat{\mathbf{V}}_{1,z}(x)$  such that  $\langle \hat{\mathbf{V}}_{1,z}, \mathbf{q}_{1,z} \rangle > y_{\mathbf{x}_1,\infty}^*$ , where we choose  $\mathbf{x}_1$  to have the same state moments as  $\mathbf{q}_{1,z}$ . This implies that

$$\lim_{i \rightarrow \infty} L_{\hat{\mathbf{m}}_{0,i}}(l_0) + L_{\hat{\mathbf{q}}_{1,i}}(y) = \lim_{i \rightarrow \infty} L_{\hat{\mathbf{m}}_{0,i}}(l_0) + y_{\hat{\mathbf{x}}_{1,i}}^* \geq \lim_{i \rightarrow \infty} L_{\hat{\mathbf{m}}_{0,i}}(l_0) + \langle \hat{\mathbf{V}}_{1,i}, \hat{\mathbf{q}}_{1,i} \rangle = \hat{\rho}_{UB}. \quad (28)$$

As long as  $\theta_{LB,i} \leq \rho_{UB,i} - \epsilon$ , that is, the termination criterion has not yet been satisfied, relation (28) implies that the subsequence  $\{\rho_{UB,i}\}$  must also converge to  $\hat{\theta}_{LB}$ . Thus, by virtue of Lemmas 3 and 4, we obtain  $\rho_k^* \leq \hat{\rho}_{UB} = \hat{\theta}_{LB} \leq \rho^*$ .

Given  $\epsilon/2 > 0$ , by the definition of a convergent sequence, there exists  $Z \in \mathbb{N}^+$  and  $I \in \mathbb{N}^+$  such that  $|\theta_{LB,z} - \hat{\theta}_{LB}| < \epsilon/2$  if  $z > Z$  and  $|\rho_{UB,i} - \hat{\theta}_{LB}| < \epsilon/2$  if  $i > I$ . Thus there exists  $J \in \mathbb{N}^+$  such that  $|\rho_{UB,z} - \theta_{LB,z}| \leq |\rho_{UB,z} - \hat{\theta}_{LB}| + |\theta_{LB,z} - \hat{\theta}_{LB}| < \epsilon$  if  $z > J$ .  $\square$

Based on Lemma 1, we can state that the higher the relaxation degree  $2k$ , the closer the undecomposed moment relaxation (26) and therefore  $\rho_k^*$  to the true optimal value  $\rho^*$  of the original GMP (7) with  $T = 2$ , since problem (26) becomes an ever tighter relaxation of (7).

The extension of the convergence properties to the case of multiple stages can be inferred by backward induction. If we add one new stage before the two-stage problem, the original two-stage problem (26) can be seen as the nested second stage of a new upper-level two-stage problem. The nested second stage converges according to Theorem 2 for given initial moments generated by the first stage. We can then apply the same arguments used for the nested problem to show the convergence of the new upper-level two-stage problem.

## 5. Numerical results

We evaluate the algorithm using a real-world long-term borehole storage problem. The Moment DDP approach developed in Section 3 is compared with the DP approach using discretization of the state/action space for the case of a small storage system in Section 5.1. The convergence of the algorithm for a larger problem with multiple storage systems is then shown in Section 5.2.

### 5.1. Single storage system

We consider the system pictured in Fig. 2, similar to the setup in De Ridder et al. (2011), comprising a borehole, a heat pump (HP), a chiller and a boiler. The objective is to satisfy the heating and cooling demand, which vary by time of year, at minimum annual cost. Heating can be supplied either by the boiler or by the HP that draws energy from the borehole. The efficiency of the HP depends on the outlet temperature of the borehole. The cooling demand can be satisfied by either running the chiller or by charging the borehole through a heat-exchanger.

This system is sufficiently small for the DP approach using discretization to be tractable. We evaluate the quality of the approximate value function generated by the Moment DDP approach, as well as the quality of the solution when the approximate value functions are used in a single-stage optimal control problem in comparison with the discretized DP solution. We assume the heating and cooling demand to be given and use measurements from the Empa Campus in Dübendorf Switzerland (Fig. 8 in the Appendix) scaled for a single storage application. The characteristics of the ground borehole are derived from a thermal response test conducted on the Empa campus. The long-term Energy Storage Management Problem (ESMP) over the horizon of one year is formulated

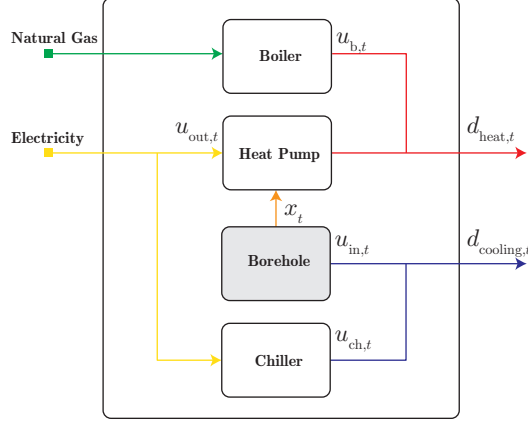


Figure 2: Schematic of the energy system with borehole storage

as follows:

$$\min_{\{x_t\}_{t=1}^T, \{u_{in,t}, u_{out,t}, u_{b,t}, u_{ch,t}\}_{t=0}^{T-1}} \sum_{t=0}^{T-1} c_e(u_{out,t} + u_{ch,t}) + c_g u_{b,t} \quad (29a)$$

$$\text{s.t. } x_{t+1} = x_t + \Delta t \frac{1}{mc} (\lambda(x_t - T_\infty) - a(x_t)u_{out,t} + u_{in,t}), \quad t = 0, \dots, T-1, \quad (29b)$$

$$a(x_t)u_{out,t} + a_b u_{b,t} = d_{heat,t}, \quad t = 0, \dots, T-1, \quad (29c)$$

$$u_{in,t} + a_{ch} u_{ch,t} = d_{cooling,t}, \quad t = 0, \dots, T-1, \quad (29d)$$

$$\underline{T} \leq x_t \leq \bar{T}, \quad t = 1, \dots, T, \quad (29e)$$

$$0 \leq u_{out,t} \leq \bar{u}_{out}; \quad 0 \leq u_{in,t} \leq \bar{u}_{in}, \quad 0 \leq u_{b,t} \leq \bar{u}_b; \quad 0 \leq u_{ch,t} \leq \bar{u}_{ch}, \quad t = 0, \dots, T-1, \quad (29f)$$

where  $x_t$  is the ground temperature,  $u_{in,t}$  the storage charge,  $u_{out,t}$  the HP power when drawing energy from the ground,  $u_{ch,t}$  the chiller power and  $u_{b,t}$  the boiler power. The heating and cooling demands are denoted as  $d_{heat,t}$  and  $d_{cooling,t}$ . The power rating limits are denoted by  $\bar{u}_{out}$ ,  $\bar{u}_{in}$ ,  $\bar{u}_{ch}$  and  $\bar{u}_b$ . The temperature of the borehole  $x_t$  is specified to remain within  $[\underline{T}, \bar{T}]$ .  $T_\infty$  denotes the boundary ground temperature,  $\lambda$  the thermal conductivity and  $mc$  the thermal inertia of the ground. If ground temperatures are not available for measurement, the model provided in Atam et al. (2015) can be used instead. We set  $T = 12$  to obtain monthly value functions, leading to  $\Delta t = 730$  hours for (29b). A linear function  $a(x_t)$  was fitted to the measurements of the coefficient of performance (COP) of the HP in the Energy Hub of the NEST building on the Empa Campus (see Fig. 9 in the Appendix ). The third column of Table 1 in the Appendix summarizes all the numerical energy system data for (29). Due to the temperature-dependent COP, the storage problem (29) is non-convex. After eliminating decision variables  $u_{b,t}$  and  $u_{ch,t}$  using the equality constraints (29c) and (29d), the problem has one state  $x_t$  and two control input decision variables  $u_{in,t}$  and  $u_{out,t}$ .

The following value function approximations are considered to solve (29):

- Discretized dynamic programming with 41 state grid points on  $[\underline{T}, \overline{T}]$  and 1001 grid points per control input on  $[0, \bar{u}]$ .
- Moment DDP with relaxation degree  $2k = 2$ ; this restricts the value function approximation to affine functions. (Recall that the maximum degree of the polynomial approximation of the value function is constrained by  $\deg(V_{t,z})\kappa_t \leq 2k$ , where in this case the highest polynomial degree found in the dynamics is  $\kappa_t = 2$ .)
- Moment DDP with relaxation degree  $2k = 4$ ; this permits quadratic value function approximations, however in this case we add constraints to restrict all quadratic terms to zero. As a result, only affine function approximations are used.<sup>9</sup>
- Moment DDP with relaxation degree  $2k = 4$ ; using the full quadratic value function approximations permitted by this relaxation degree.

The Moment DDP approach is implemented using YALMIP (Lofberg, 2004) and solved with MOSEK<sup>TM</sup>. The discretized DP problem is implemented and solved using the *dpm* toolbox of Sundström & Guzzella (2009) and MATLAB<sup>TM</sup>. The problem data are scaled to be contained in the unit box to improve the numerical performance of the Moment DDP approach.

First, we compare the accuracy of different value function bases for a uniform initial state distribution. In Fig. 3, the approximate value functions are shown together with the reference computed by discretized DP. The kinks in the DP value functions for the months of May to August are caused by the additional cost incurred by using the chiller if the storage temperature is too high for cooling. The kinks in March and April are due to two different operating modes: using the HP to provide heat or both, the HP and the boiler. Affine and quadratic approximate value functions generated by relaxation  $2k = 4$  are a close fit for most months. For the months May to September, the lower sections of the approximate value functions are less accurate. Whereas the slopes of the approximate functions are very close to discretized DP reference, the kink positions are not. However, as subsequent results on the performance of the resulting control policy demonstrate, using the borehole to provide cooling is still optimal. There is a considerable difference between the discretized DP and the piecewise affine value function generated by the relaxation of order  $2k = 2$ .

The convergence of the lower bound  $\rho_{LB}$  and the upper bound  $\rho_{UB}$  of the Moment DDP algorithm for different polynomial basis functions is shown in Fig. 4. Affine basis functions make the Moment DDP algorithm converge faster than quadratic basis functions. The total solver times

---

<sup>9</sup>For consistency, the primal problem over moments also has to be modified (relaxed) by removing some linear equality constraints on higher-order moments arising from the dynamics (17b). For brevity we do not detail this procedure here.

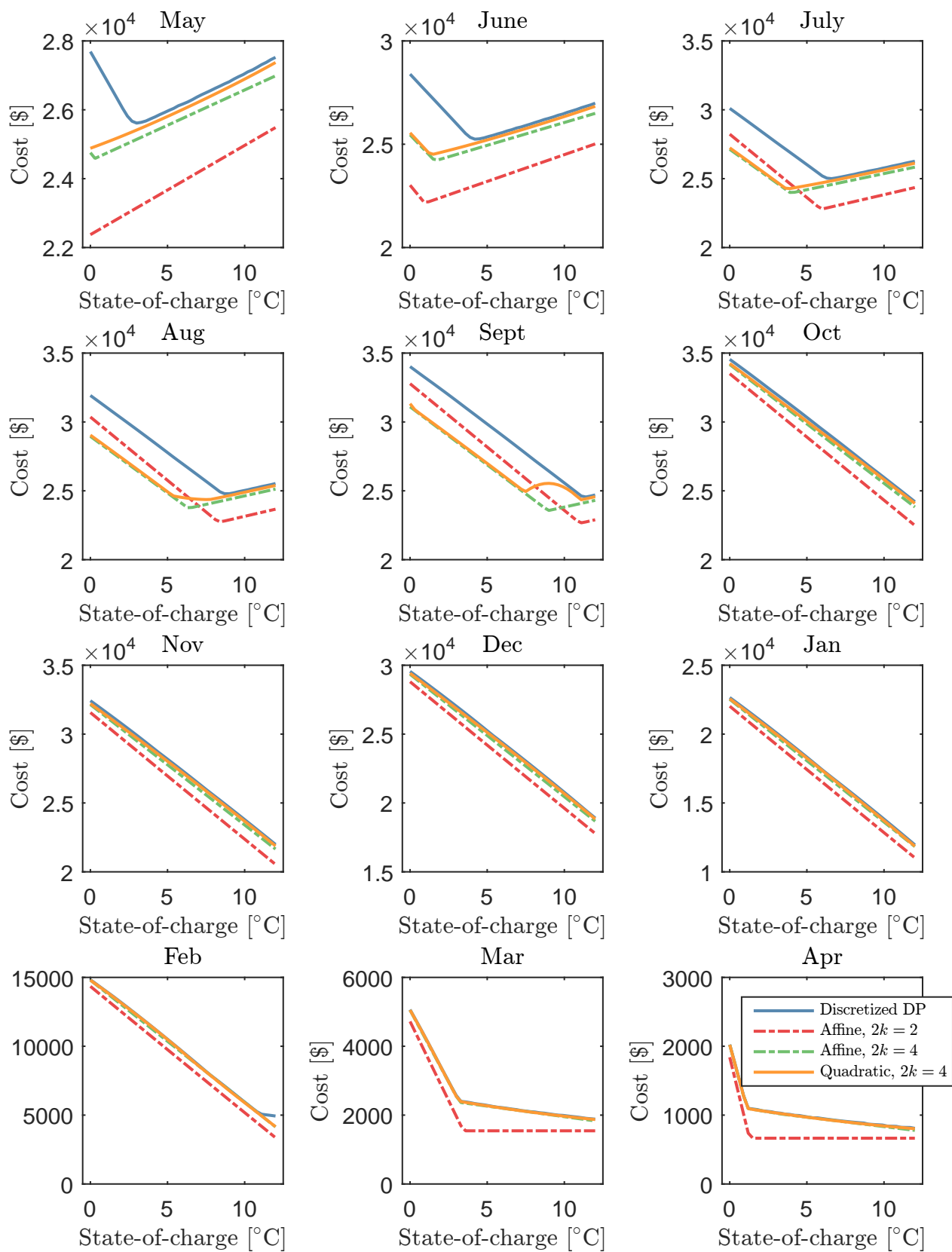


Figure 3: Value function approximations using different basis functions in comparison to discretized DP.



for a predefined convergence tolerance are reported in Table 2 in the Appendix . Note that as in conventional DDP, problems (11) and (17) increase slightly in size at every iteration as we add additional under-approximating value functions.

Finally, instead of (29), we solve a sequence of single-stage problems augmented with approximate value functions obtained by the Moment DDP approach. For each month  $t \in \{1, \dots, 12\}$ , we solve:

$$\begin{aligned} \min_{x_{t+1}, u_{in,t}, u_{out,t}, u_b,t, u_{ch,t}} \quad & c_e(u_{out,t} + u_{ch,t}) + c_g u_b,t \\ & + \max\{V_{t+1,0}(x_{t+1}), \dots, V_{t+1,z}(x_{t+1})\} \end{aligned} \quad (30a)$$

$$\text{s.t. } x_{t+1} = x_t + \Delta t \frac{1}{mC} (\lambda(x_t - T_\infty) - a(x_t)u_{out,t} + u_{in,t}) \quad (30b)$$

$$a(x_t)u_{out,t} + a_b u_b,t = d_{\text{heat},t}, \quad (30c)$$

$$u_{in,t} + a_{ch} u_{ch,t} = d_{\text{cooling},t}, \quad (30d)$$

$$\underline{T} \leq x_{t+1} \leq \bar{T}, \quad (30e)$$

$$0 \leq u_{out,t} \leq \bar{u}_{out}; \quad (30f)$$

$$0 \leq u_{in,t} \leq \bar{u}_{in}; \quad 0 \leq u_b,t \leq \bar{u}_b; \quad 0 \leq u_{ch,t} \leq \bar{u}_{ch} \quad (30g)$$

The start of the storage cycle is assumed to be the beginning of May because the cooling overcomes the heating demand during this period (see Fig. 8). In Fig. 6, we show the total cost of operating the system over the full horizon for a uniformly distributed number of initial states when each month is solved as a single-stage problem (30). We use the generic nonlinear solver IPOPT (Wachter & Biegler, 2006) to compute a locally-optimal solution. The affine value functions perform almost as well as the forward simulation of discretized DP. The quadratic value functions lead to sub-optimal results with a local optimization algorithm for some initial states. This might be due to the non-convexity of the approximate value function in September (see Fig. 3).

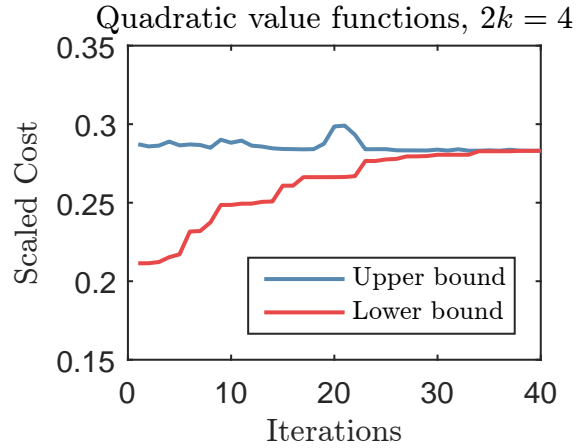
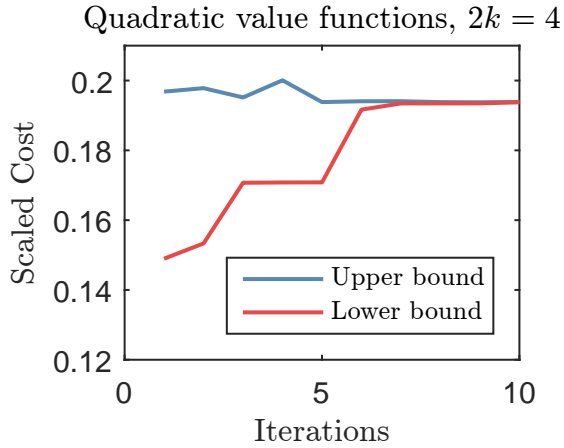
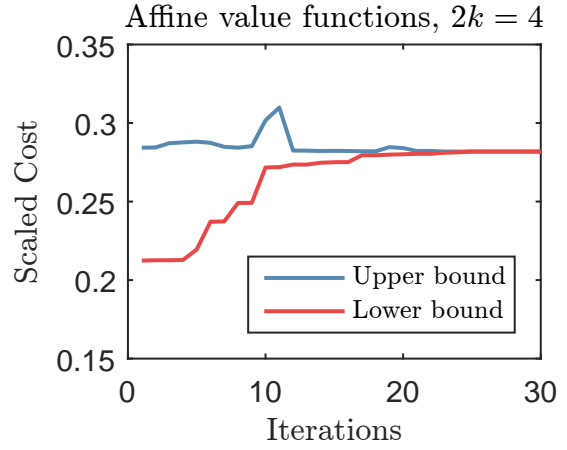
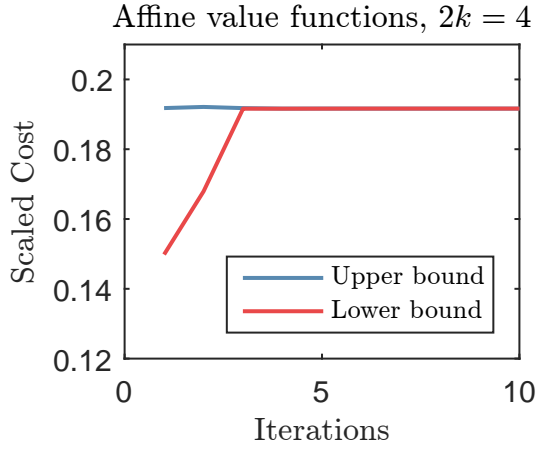
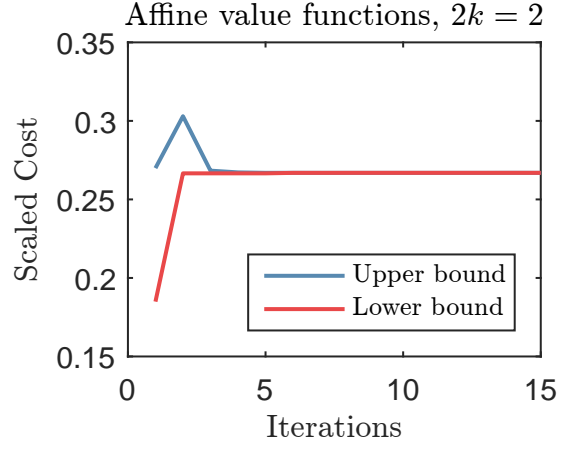
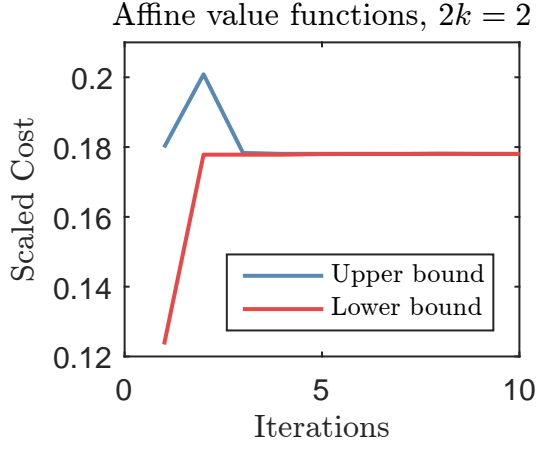


Figure 4: Single storage system: Convergence of the lower bound  $\rho_{LB}$  and the upper bound  $\rho_{UB}$  of the Moment DDP algorithm for different basis functions

Figure 5: Multiple storage systems: Convergence of the lower bound  $\rho_{LB}$  and the upper bound  $\rho_{UB}$  of the Moment DDP algorithm for different basis functions

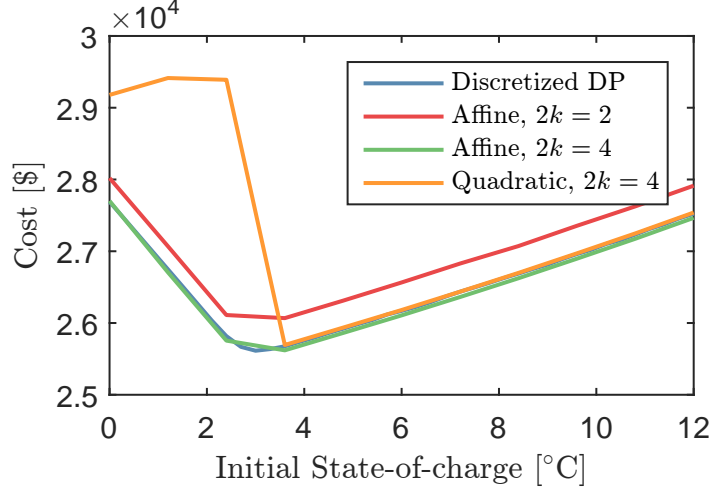


Figure 6: Cost over the full year starting in May for a sequence of single-stage problems with different approximate functions in comparison to discretized DP

### 5.2. Multiple storage systems

We now evaluate the convergence of the Moment DDP approach for a higher dimensional problem, namely an ESMP with three different storage systems:

$$\min_{\{\{x_{i,t}\}_{t=1}^T, \{u_{in,i,t}, u_{out,i,t}\}_{t=0}^{T-1}\}_{i=1}^3, \{u_{b,t}, u_{ch,t}\}_{t=0}^{T-1}} \sum_{t=0}^{T-1} \left( c_e u_{ch,t} + c_g u_{b,t} + \sum_{i=1}^3 c_e u_{out,i,t} \right) \quad (31a)$$

$$\text{s.t. } x_{i,t+1} = x_{i,t} + \Delta t \frac{1}{mc} (\lambda_i (x_{i,t} - T_\infty) + u_{in,i,t} - a(x_{i,t}) u_{out,i,t}),$$

$$t = 0, \dots, T-1, i = 1, 2, 3, \quad (31b)$$

$$\sum_{i=1}^3 a(x_{i,t}) u_{out,i,t} + a_b u_{b,t} = d_{heat,t}, \quad t = 0, \dots, T-1, \quad (31c)$$

$$\sum_{i=1}^3 u_{in,i,t} + a_{ch} u_{ch,t} = d_{cooling,t}, \quad t = 0, \dots, T-1, \quad (31d)$$

$$\underline{T} \leq x_{i,t} \leq \bar{T}, \quad t = 1, \dots, T, i = 1, 2, 3, \quad (31e)$$

$$0 \leq u_{out,i,t} \leq \bar{u}_{out}; \quad 0 \leq u_{in,i,t} \leq \bar{u}_{in}, \quad t = 0, \dots, T-1, i = 1, 2, 3, \quad (31f)$$

$$0 \leq u_{b,t} \leq \bar{u}_b; \quad 0 \leq u_{ch,t} \leq \bar{u}_{ch}, \quad t = 0, \dots, T-1, \quad (31g)$$

With two additional boreholes, the discretized DP approach memory requirements become excessive, since a grid must be spanned over a 9-dimensional decision space after elimination of the boiler and chiller variables using (31c) and (31d). In addition to the energy system data of the fourth column of Table 1, we use the heating and cooling demand of the single storage example of the previous section multiplied by a factor 3 as input data. The convergence of affine and quadratic approximate value functions for a uniform initial state distribution is shown in Fig. 5. All methods converge in a reasonable number of iterations. Table 2 in the Appendix reports the total solver

times for a predefined tolerance.

## 6. Conclusion and Future Work

This paper presented a novel value function approximation scheme for nonlinear multi-stage problems that leverages sum-of-squares techniques within a DDP framework. The scheme is based on a finite-horizon GMP for discrete-time dynamical systems. The primal, a moment problem, and the dual, an SOS program, are used iteratively to refine the statistics of the forward state trajectory and the approximate value functions respectively. Whereas DDP returns value functions that apply locally around trajectories emanating from a single initial state, and generally only for linear system dynamics and cost, the Moment DDP approach returns approximate value functions for a distribution of initial states, and moreover achieves this for systems with polynomial dynamics, costs, and constraints. Depending on the degree of polynomials used, the optimal policy obtained by short-term problems augmented with approximate value functions returned by the Moment DDP approach can be almost as cost-effective as that obtained by discretized DP. We also demonstrated convergence of the Moment DDP approach for a case that is computationally too demanding for discretized DP.

The computational complexity of the Moment DDP approach could be reduced by exploiting any sparsity present in the problem data in (2) (Waki et al., 2006). This would draw on the experience of Molzahn & Hiskens (2015) and Ghaddar et al. (2016), who successfully exploited the sparse structure of electrical networks to obtain global solutions to the nonlinear optimal power flow problem using moment relaxations. Alternative positivity certificates, such as the one proposed in Ahmadi & Majumdar (2014), also offer the possibility of reduced computational complexity.

## Acknowledgments

We would like to thank Viktor Dorer, Roy Smith and Jan Carmeliet for their valuable help and support. We are also grateful to Xinyue Li for her work on the heat pump characterization and to Paul Beuchat, Georgios Darivianakis, Benjamin Flamm, Mohammad Khosravi, and Annika Eichler for fruitful discussions. This research project is financially supported by the Swiss Innovation Agency Innosuisse and by NanoTera.ch under the project HeatReserves, and is part of the Swiss Competence Center for Energy Research SCCER FEEB&D.

## References

- Abgottspon, H. (2015). *Hydro power planning: Multi-horizon modeling and its applications*. Ph.D. thesis ETH Zürich.
- Ahmadi, A. A., & Majumdar, A. (2014). DSOS and SDSOS optimization: LP and SOCP-based alternatives to sum of squares optimization. *2014 48th Annual Conference on Information Sciences and Systems, CISS 2014*, (pp. 2–6).

- Anderson, E. J., & Nash, P. (1987). *Linear programming in infinite-dimensional spaces : theory and applications*. Wiley.
- Atam, E., Patteeuw, D., Antonov, S. P., & Helsen, L. (2015). Optimal Control Approaches for Analysis of Energy Use Minimization of Hybrid Ground-Coupled Heat Pump Systems. *IEEE Transactions on Control Systems Technology*, 24.
- Bertsekas, D. P. (1995). *Dynamic programming and optimal control* volume 1. Athena scientific Belmont, MA.
- Beuchat, P. N., Warrington, J., & Lygeros, J. (2017). Point-wise Maximum Approach to Approximate Dynamic Programming. In *Proceedings of the IEEE Conference on Decision and Control*.
- Cerisola, S., Latorre, J. M., & Ramos, A. (2012). Stochastic dual dynamic programming applied to nonconvex hydrothermal models. *European Journal of Operational Research*, 218, 687–697.
- Darivianakis, G., Eichler, A., Smith, R. S., & Lygeros, J. (2017). A Data-Driven Stochastic Optimization Approach to the Seasonal Storage Energy Management. *IEEE Control Systems Letters*, 1, 394–399.
- De Ridder, F., Diehl, M., Mulder, G., Desmedt, J., & Van Bael, J. (2011). An optimal control algorithm for borehole thermal energy storage systems. *Energy and Buildings*, 43, 2918–2925.
- Ghaddar, B., Marecek, J., & Mevissen, M. (2016). Optimal Power Flow as a Polynomial Optimization Problem. *IEEE Transactions on Power Systems*, 31, 539–546.
- Girardeau, P., Leclere, V., & Philpott, A. B. (2015). On the Convergence of Decomposition Methods for Multistage Stochastic Convex Programs. *Mathematics of Operations Research*, 40, 130–145.
- Guigues, V. (2018). Inexact cuts in Deterministic and Stochastic Dual Dynamic Programming applied to linear optimization problems. *arXiv:1707.00812*, .
- Guigues, V., & Römisich, W. (2012). Sampling-Based Decomposition Methods for Multistage Stochastic Programs Based on Extended Polyhedral Risk Measures. *SIAM Journal on Optimization*, 22, 286–312.
- Hernández-Lerma, O. (1989). *Adaptive Markov Control Processes* volume 79 of *Applied Mathematical Sciences*. New York, NY: Springer New York.
- Hernández-Lerma, O., & Hernández-Hernández, D. (1994). Discounted Cost Markov Decision Processes on Borel Spaces: The Linear Programming Formulation. *Journal of Mathematical Analysis and Applications*, 183, 335–351.
- Hernández-Lerma, O., & Lasserre, J. B. (2012). *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer.
- Jiang, X. S., Jing, Z. X., Li, Y. Z., Wu, Q. H., & Tang, W. H. (2014). Modelling and operation optimization of an integrated energy based direct district water-heating system. *Energy*, 64, 375–388.
- Kamoutsi, A., Sutter, T., Esfahani, P. M., & Lygeros, J. (2017). On Infinite Linear Programming and the Moment Approach to Deterministic Infinite Horizon Discounted Optimal Control Problems. *arXiv:1703.09005*, .
- Korda, M., Henrion, D., & Jones, C. N. (2017). Convergence rates of moment-sum-of-squares hierarchies for optimal control problems. *Systems and Control Letters*, 100, 1–5.
- Lasota, A., & Mackey, M. C. (1994). *Chaos, Fractals, and Noise* volume 97 of *Applied Mathematical Sciences*. New York, NY: Springer New York.
- Lasserre, J. B. (2014). *Moments, positive polynomials and their applications*.. Series on Optimization and Its Applications. Imperial College Press.
- Lasserre, J. B., Henrion, D., Prieur, C., & Trélat, E. (2008). Nonlinear Optimal Control via Occupation Measures and LMI-Relaxations. *SIAM Journal on Control and Optimization*, 47, 1643–1666.

- Lasserre, J. B., & Thanh, T. P. (2013). Convex underestimators of polynomials. *Journal of Global Optimization*, *56*, 1–25.
- Lofberg, J. (2004). YALMIP : a toolbox for modeling and optimization in MATLAB. In *2004 IEEE International Conference on Computer Aided Control Systems Design* (pp. 284–289).
- Molzahn, D. K., & Hiskens, I. A. (2015). Sparsity-Exploiting Moment-Based Relaxations of the Optimal Power Flow Problem. *IEEE Transactions on Power Systems*, *30*, 3168–3180.
- O’Donoghue, B., Wang, Y., & Boyd, S. (2011). Min-max approximate dynamic programming. In *2011 IEEE International Symposium on Computer-Aided Control System Design (CACSD)* (pp. 424–431).
- Pereira, M., & Pinto, L. (1991). Multi-stage stochastic optimization applied to energy planning. *Mathematical Programming*, *52*, 359–375.
- Philpott, A., & Guan, Z. (2008). On the convergence of stochastic dual dynamic programming and related methods. *Operations Research Letters*, *36*, 450–455.
- Powell, W. B. (2011). *Approximate dynamic programming : solving the curses of dimensionality*. Wiley.
- Putinar, M., & Vasilescu, F.-H. (1999). Positive polynomials on semi-algebraic sets. *Comptes Rendus de l’Académie des Sciences - Series I - Mathematics*, *328*, 585–589.
- Savorgnan, C., Lasserre, J. B., & Diehl, M. (2009). Discrete-time stochastic optimal control via occupation measures and moment relaxations. *Proceedings of the IEEE Conference on Decision and Control*, (pp. 519–524).
- Summers, T. H., Kariotoglou, N., Kamgarpour, M., Summers, S., & Lygeros, J. (2012). Approximate Dynamic Programming via Sum of Squares Programming. In *European Control Conference* (pp. 191–197).
- Sundström, O., & Guzzella, L. (2009). A generic dynamic programming Matlab function. *Proceedings of the IEEE International Conference on Control Applications*, (pp. 1625–1630).
- Taylor, J. A. (2015). *Convex Optimization of Power Systems*. Cambridge University Press.
- Wachter, A., & Biegler, L. T. (2006). On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Mathematical Programming*, *106*, 25–57.
- Waki, H., Kim, S., Kojima, M., & Muramatsu, M. (2006). Sums of Squares and Semidefinite Program Relaxations for Polynomial Optimization Problems with Structured Sparsity. *SIAM Journal on Optimization*, *17*, 218–242.
- Wang, Y., O’Donoghue, B., & Boyd, S. (2014). Approximate Dynamic Programming via Iterated Bellman Inequalities. *International Journal of Robust and Nonlinear Control*, *25*, 1472–1496.
- Zakeri, G., Philpott, A. B., & Ryan, D. M. (2000). Inexact Cuts in Benders Decomposition. *SIAM Journal on Optimization*, *10*, 643–657.
- Zou, J., Ahmed, S., & Sun, X. A. (2018). Stochastic dual dynamic integer programming. *Mathematical Programming*, (pp. 1–42).

## Appendix

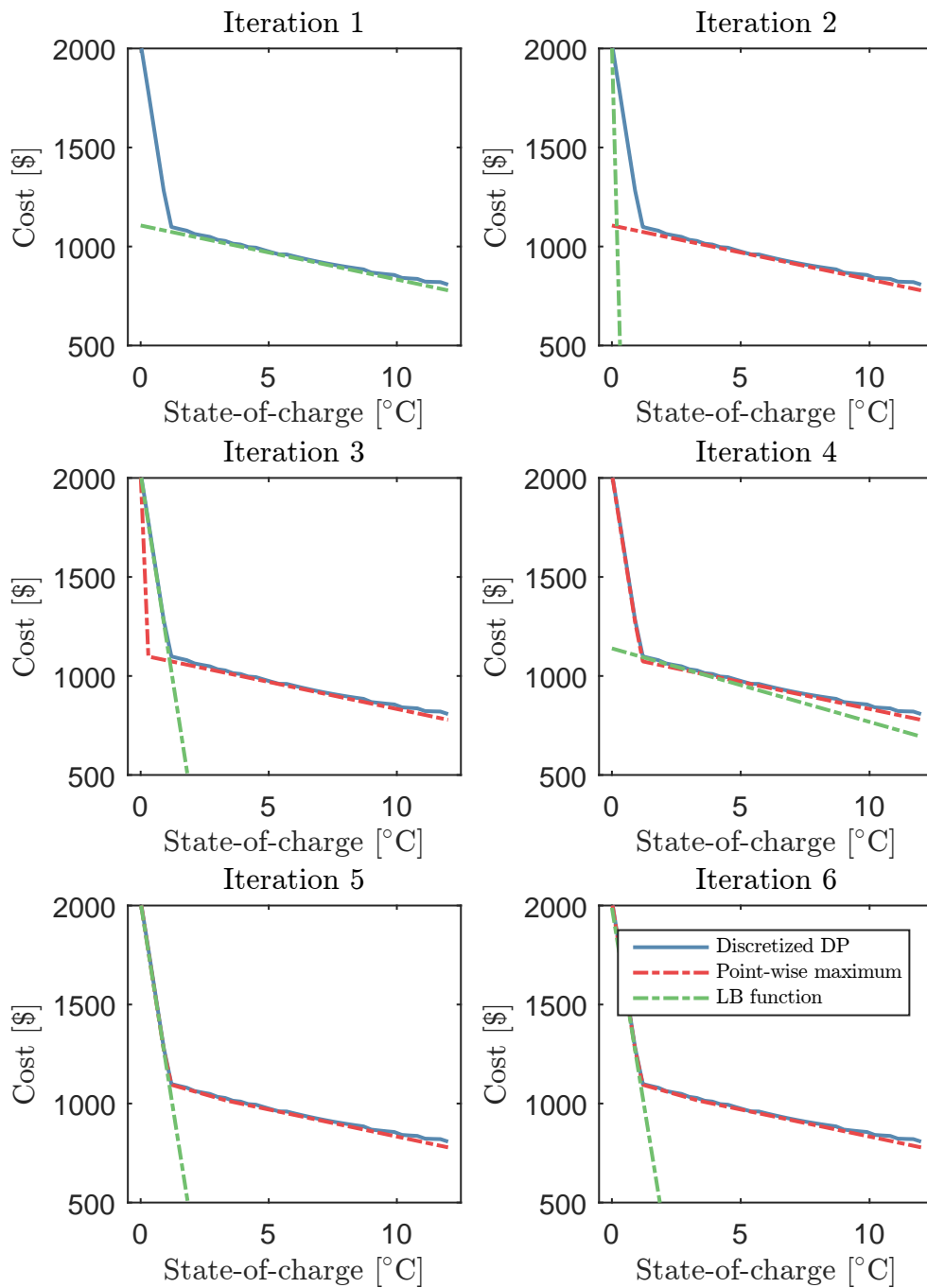


Figure 7: Single storage example of Section 5: Cost of stored energy in the beginning of April ( $t = 11$ ) approximated using affine basis functions and  $2k = 4$ , shown for six DDP iterations. New lower-bounding functions (LB function) are shown in green. The point-wise maximum of all previous lower-bounding functions is shown in red.

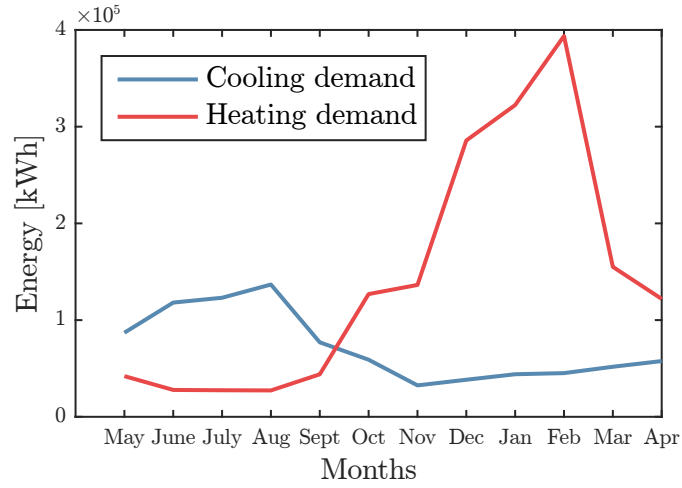


Figure 8: Heating and cooling demand  $d_{\text{heat},t}$  and  $d_{\text{cooling},t}$  of the single storage application over a year

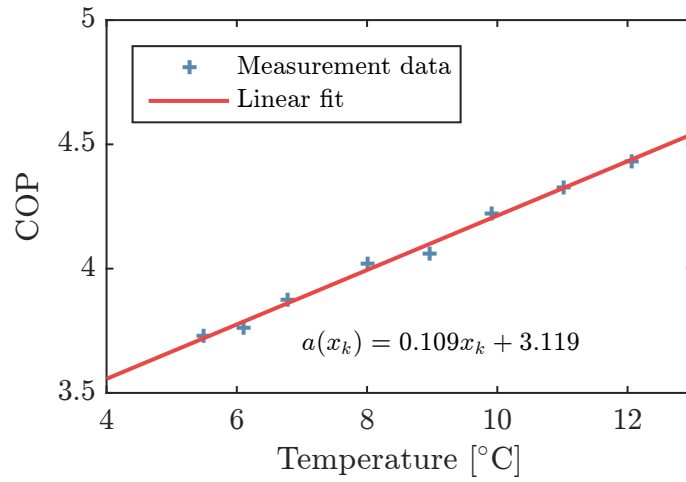


Figure 9: Fitting of the inlet temperature-dependent COP  $a(x_t)$  of the HP



Table 1: Energy system data

	Parameter	Single storage	Multiple storage
<b>Grid Feeders</b>			
Power	Cost $c_e$ :	0.096\$/kWh	0.096\$/kWh
Gas	Cost $c_g$ :	0.063\$/kWh	0.063\$/kWh
<b>Conversion</b>			
HPs	COP $a(x_t)$ :	see Fig. 9	see Fig. 9
	Capacity $\bar{u}_{\text{out}}$ :	60kW	60kW
Boiler	Efficiency $a_b$ :	0.7	0.7
	Capacity $\bar{u}_b$ :	285 kW	855 kW
Chiller	COP $a_{\text{ch}}$ :	5	5
	Capacity $\bar{u}_{\text{ch}}$ :	150kW	450kW
<b>Storage</b>			
Boreholes	Conductivity $\lambda$ :	0.621kW/°C	0.621kW/°C±10%
	Inertia $mc$ :	14805kWh/°C	14805kWh/°C
	Capacity $\bar{u}_{\text{in}}$ :	100kW	100kW
	Ground $T_\infty$ :	12°C	12°C
	Range $[\underline{T}, \bar{T}]$ :	[0,12]°C	[0,12]°C

Table 2: Accumulated MOSEK<sup>TM</sup> solver time over all iterations of the Moment DDP approach obtained on a PC with an Intel-i5 2.2GHz CPU with 8GB RAM for a tolerance of  $\epsilon = 10^{-4}$  (after scaling the problem data to the unit box)

Basis functions/Relaxation	Single storage	Multiple storage
Affine value functions, $2k = 2$	4.77s	6.39s
Affine value functions, $2k = 4$	5.77s	18.65min
Quadratic value functions, $2k = 4$	25.23s	28.24min