Revisiting Approximate Linear Programming Using a Saddle Point Approach

Qihang Lin
Tippie College of Business, The University of Iowa, 21 East Market Street, Iowa City, IA 52242, USA, qihang-lin@uiowa.edu

Selvaprabu Nadarajah
College of Business Administration, University of Illinois at Chicago, 601 South Morgan Street, Chicago, Illinois, 60607, USA, selvan@uic.edu

Negar Soheili
College of Business Administration, University of Illinois at Chicago, 601 South Morgan Street, Chicago, Illinois, 60607, USA, nazad@uic.edu

Approximate linear programs (ALPs) are well-known models for computing value function approximations (VFAs) of intractable Markov decision processes (MDPs) arising in applications. VFAs from ALPs have desirable theoretical properties, define an operating policy, and provide a lower bound on the optimal policy cost, which can be used to assess the suboptimality of heuristic policies. However, solving ALPs near-optimally remains challenging, for example, when approximating MDPs with nonlinear cost functions and transition dynamics or when rich basis functions are required to obtain a good VFA. We address this tension between theory and solvability by proposing a convex saddle-point reformulation of an ALP that includes as primal and dual variables, respectively, a vector of basis function weights and a constraint violation density function over the state-action space. To solve this reformulation, we develop a proximal stochastic mirror descent (PSMD) method. We establish that PSMD returns a near-optimal ALP solution and a lower bound on the optimal policy cost in a finite number of iterations with high probability. We numerically compare PSMD with several benchmarks on inventory control and energy storage applications. We find that the PSMD lower bound is tighter than a perfect information bound. In contrast, the constraint sampling approach to solve ALPs may not provide a lower bound and applying row generation to tackle ALPs is not computationally viable. PSMD policies outperform problem-specific heuristics and are comparable or better than the policies obtained using constraint sampling. Overall, our ALP reformulation and solution approach broadens the applicability of approximate linear programming.

History: Initial version: April 2017; Current version: June 2018.

1. Introduction

Business problems arising in Operations Research, Operations Management, and Financial Engineering, among other areas, involve sequential decision-making under uncertainty. Markov decision process (MDP; Puterman 1994) models of these problems are common but generally intractable

to solve exactly due to the well-known curses of dimensionality (Powell 2011, pages 3 and 112). Approximate dynamic programming (ADP) provides techniques to heuristically solve such MDPs. A common ADP strategy constructs a low-dimensional approximation of the MDP value function as a linear combination of a manageable number of basis functions and uses it to obtain a heuristic policy (Bertsekas 2007, Powell 2011). It is known that value function approximations (VFAs) closely approximating the MDP value function provide policies with small suboptimality gaps.

We focus on the linear programming approach to ADP, which solves the so called approximate linear program (ALP) to compute a VFA (Schweitzer and Seidmann 1985, de Farias and Van Roy 2003). This approach has been successfully applied in a number of domains including economics (Trick and Zin 1997), inventory routing and control (Adelman 2004, Adelman and Klabjan 2012, Adelman and Barz 2013), revenue management (Adelman 2007, Zhang and Adelman 2009, Adelman and Mersereau 2013), queuing (Farias and Van Roy 2007), and health care operations (Restrepo 2008, ch. 4, Patrick et al. 2008). ALPs have attractive theoretical properties such as a guarantee on the error between its VFA and the best possible VFA given a set of basis functions. This error is zero when the basis functions span the exact value function, that is, the ALP VFA coincides with the exact value function. Moreover, the ALP VFA also provides a lower bound on the optimal policy cost, which is useful in assessing the optimality gap of heuristic policies.

ALPs have a manageable number of variables (i.e., VFA weights) because they are derived by applying a VFA on the variable space of the well-known exact linear programming representation of an MDP (Puterman 1994, §6.9). However, the number of constraints in these models remains large (possibly infinite) — one for each state and action — and each constraint may embed hard to compute expectations. Compact reformulations and row generation strategies have been used to solve ALPs near-optimally when expectations appearing in its constraints can be evaluated precisely, the basis functions are simple, for example, affine or piecewise linear functions of the MDP state, and the underlying MDP cost function and transition dynamics have special structure (e.g., Adelman 2007, Vossen and Zhang 2015). Several applications do not satisfy one or more of these requirements and thus the near optimal solution of ALPs remains challenging (see Restrepo 2008, ch. 4, Adelman and Klabjan 2012, and §§7.2−7.4 for examples). Constraint sampling is a heuristic strategy for solving ALPs that is more broadly applicable. This approach constructs a version of the ALP containing only a subset of its constraints obtained via sampling (de Farias and Van Roy 2004, Farias and Van Roy 2006, Restrepo 2008, ch. 4, and Sun et al. 2016). However, the optimal objective of a sampled ALP may not provide a lower bound on the optimal policy cost. Guarantees on the error between the sampled and exact versions of an ALP exist under an idealized sampling distribution (de Farias and Van Roy 2004), while in practice it is known that this error is sensitive to the choice of such distribution.

Overall, the ALP VFA has desirable theoretical properties, but solving an ALP near-optimally is challenging under nonlinear MDP cost functions or transition dynamics and/or when using rich basis functions. In this paper, we attempt to address this tension between the theory and solvability of ALPs in the context of infinite-horizon discounted cost MDPs with compact state and action spaces by developing a novel ALP reformulation and solution approach.

Our reformulation casts the search for a near-optimal ALP VFA as solving a primal-dual problem following two steps. The first step develops a saddle-point formulation that endogenizes the problem of computing the most violated ALP constraint. This saddle-point problem involves a linear primal optimization over basis function weights and a potentially non-convex dual optimization over the state-action space, where the non-convexity could stem from the structure of the immediate cost, transition, or basis functions. The second step of our reformulation eliminates such non-convexity in the dual optimization by lifting its variable space to the infinite dimensional space of continuous state-action density functions and adding a regularization term. In other words, this second step modifies the dual problem to avoid solving a potentially non-convex (finite-dimensional) optimization problem for finding the most violated constraint, and instead, tackles an infinite dimensional convex optimization problem for computing a constraint violation density function.

Our solution approach, dubbed proximal stochastic mirror-descent (PSMD), is a primal-dual stochastic gradient method for solving the aforementioned regularized saddle-point ALP reformulation. PSMD embeds a nontraditional scheme to automatically update the weight of the regularization constant in our saddle-point formulation as its iterations progress. Its primal gradient update works with the basis function weights and is standard (Nemirovski et al. 2009). In contrast, PSMD's dual gradient update tackles a continuous state-action density function, which is challenging, and is based on a novel stochastic functional gradient. We establish that PSMD finds a near-optimal ALP solution in a finite number of iterations with high probability. The implementation of PSMD is not tied to this theoretical iteration complexity but can be stopped based on a run-time limit or a bound on a computable primal-dual gap. Under both stopping criteria, PSMD returns basis function weights and a lower bound on the optimal policy cost. Although the PSMD updates involve solving simple optimization problems, it does have to contend with approximating high dimensional expectations over the state and state-action spaces, for which we use Markov chain Monte Carlo techniques (Robert and Casella 2004) such as Metropolis-Hastings.

We numerically illustrate the behavior of the PSMD lower and upper bounds on an inventory control problem with partial backlogging and no lead time, also highlighting that PSMD learns regions of high ALP constraint violations via its dual updates. Then, we compare PSMD with the constraint sampling ALP approach (ALP-CS) and other benchmark bounds and policies for solving a perishable version of this problem with lead time and a second application involving the

management of end-user energy storage. We do not employ row generation in these applications as it requires solving non-convex optimization problems to generate new constraints or computing expectations with no closed form appearing in the ALP constraints. We find that the PSMD lower bound is stronger than the perfect information bound on both our applications, while ALP-CS may not provide a valid lower bound. On perishable inventory control instances, PSMD policies outperform ones from a tractable special case of our problem, a look-ahead heuristic, and ALP-CS. On consumer energy storage instances, the PSMD policies dominate a look-ahead heuristic and are comparable to the ALP-CS policies. Our results indicate that PSMD is an effective method for solving ALPs, and its policies and lower bounds are of high quality.

The rest of this paper is organized as follows. We review related literature and elaborate on the novelty of our research in §2. We provide background on MDPs and ALPs in §3. We describe the development of our ALP reformulation and PSMD in §4 and §5, respectively, and we discuss PSMD implementation details in §6. We present our numerical study in §7 and conclude in §8. An electronic companion contains proofs and additional results.

2. Literature review and novelty

Our research adds to the ALP literature by extending the class of ALPs that can be solved near-optimally. The perspective of viewing an ALP as a saddle-point problem is new. PSMD is an alternative to row generation and constraint sampling for solving ALPs that addresses some of their known difficulties discussed earlier, exhibits promising numerical performance on important applications, and offers theoretical guarantees. The use of regularization in ALPs has been previously considered by Petrik et al. (2010) and Bhat et al. (2012) for basis function selection. We use regularization in a different manner. In particular, we apply regularization in the dual optimization involving a sampling distribution to ensure that it is well defined. Our approximation guarantees supplement the theory on ALPs found in de Farias and Van Roy (2003), where these authors bound the error between the exact ALP VFA and the best possible VFA given a fixed set of basis functions. Instead, we provide a bound on the error between the computed ALP VFA and the exact ALP VFA. de Farias and Van Roy (2004) derive guarantees of this type for constraint sampling but do not provide a lower bound on the optimal policy cost and assume the exact computation of expectations. PSMD relaxes the latter assumption and extends the applicability of ALP-based lower bounds beyond situations where row generation is viable.

Convergent primal-dual methods based on linear programming have been considered in the literature (e.g., Hernández-Lerma and Lasserre 1996, Klabjan and Adelman 2007, and references therein). Klabjan and Adelman (2007) develop theory for such an approach to solve semi-Markov decision processes in an average cost setting. Here, VFAs are dynamically generated using ridge

basis functions. Adelman and Klabjan (2012) implement this scheme for an inventory control application using row generation which requires solving nonlinear mixed integer programs. Our approach cannot dynamically generate basis functions but it avoids solving non-convex optimization problems by shifting the burden instead to computing expectations, and relaxes the need to exactly evaluate the expectations appearing in the ALP constraints.

Adelman (2003, 2004), Topaloglu and Kunnumkal (2006), Klabjan and Adelman (2007), and Sun et al. (2016) consider ALPs for solving inventory control applications. Sun et al. (2016) apply ALPs in a perishable inventory control setting but do not provide a lower bound on the optimal policy cost as they use a constraint sampling solution approach. Our numerical study adds to this line of work by considering partial backlogging and providing lower bounds. In a finite horizon setting, Nadarajah and Secomandi (2017) and Nadarajah et al. (2015) heuristically solve ALPs with polynomial basis functions and exactly solve relaxations of ALPs with piecewise constant basis functions, respectively, in the context of merchant energy storage (see Secomandi and Seppi 2014 and references therein). Neither paper provides a mechanism to compute ALP based lower bounds. Instead, we use an ALP to manage consumer energy storage in an infinite horizon setting (Grillo et al. 2012, van de Ven et al. 2013, Erseghe et al. 2014) and provide lower bounds.

Our work identifies approximate linear programming as an application domain for first-order methods in stochastic optimization. We do not attempt to review this extant literature but refer the reader to Juditsky and Nemirovski (2011a,b), Bubeck (2015), and Duchi (2016) for excellent treatments of this topic. PSMD builds on the mirror-descent method in Nemirovski et al. (2009) and primal-dual methods for solving saddle-point problems (e.g., Chen et al. 2014b and references therein). Specifically, we develop a new functional stochastic gradient based dual update to tackle the infinite-dimensional nature of the dual decision variable in our ALP saddle-point reformulation. This functional gradient is available in closed form because we use a Kullback-Liebler regularization term in the saddle-point problem as well as our update of the endogenized state-action distribution. Such a closed-form expression is not available when using a Euclidean regularizer as done in Chen et al. (2014b). In addition, PSMD features a mechanism to automatically update the regularization weight in the saddle-point formulation as iterations progress. These differences also entailed new elements in our analysis to establish theoretical guarantees for PSMD.

3. Background material

We describe an infinite horizon MDP model in §3.1. In §3.2, we discuss the approximate linear programming approach to compute a VFA and feasible policies for this MDP.

3.1. MDP model

We consider decision-making applications that give rise to infinite horizon MDPs (Puterman 1994, Hernández-Lerma and Lasserre 1996, Bertsekas 2007). We denote the state and action spaces by $\mathcal{S}$ and $\mathcal{A}$, respectively, and assume that both of these sets are continuous, convex, and compact. Examples of applications that fall into this class of MDPs include inventory control, energy storage, and queuing control (de Farias and Van Roy 2004, Zipkin 2008b, van de Ven et al. 2013, Chen et al. 2014a, Erseghe et al. 2014). We also assume these sets are full-dimensional to avoid technical complications. Although we assume continuous state and action spaces throughout, low-dimensional discrete states can easily be incorporated into our approach.

At each time step, the process underlying the MDP transitions from a given state s to a state s' in response to an action a chosen by the decision maker. An action a executed at state s incurs an immediate cost $c(s,a)$, which is a continuous function over the state and action space. This action also results in a random transition to states in set $\mathcal{S}$ according to a probability distribution $p(\cdot|s,a)$. A (stationary) policy $\pi : \mathcal{S} \to \mathcal{A}$ assigns an action from set $\mathcal{A}$ to each state in set $\mathcal{S}$. The expected discounted cost of using policy π over an infinite planning horizon is

$$V^{\pi}(s) := \mathbb{E}_s^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t c(s_t, \pi(s_t))\right],$$

where $\gamma \in (0,1)$ is the discount factor, s_t is the state reached at step t by executing policy π starting from state s_0 equal to state s, and $\mathbb{E}_s^{\pi}$ is expectation defined by the policy π starting from state s and the transition probability distribution functions. An MDP minimizes $V^{\pi}(s)$ over all feasible policies, denoted by set Π, to identify an optimal policy as follows:

$$V^*(s) := \inf_{\pi \in \Pi} V^{\pi}(s), \quad \forall s \in \mathcal{S}. \tag{1}$$

The term $V^*(s)$ is the minimum total discounted expected cost of operating over an infinite horizon starting from state s. We make the following assumption in the rest of the paper to ensure that the infimum in (1) can be replaced by a minimum (Hernández-Lerma and Lasserre 1996, ch. 3).

Assumption 1. *The function $c(\cdot,\cdot)$ is Lipschitz continuous on $\mathcal{S} \times \mathcal{A}$. Moreover, the function $\int_{\mathcal{S}} f(s')p(ds'|\cdot,\cdot)$ is Lipschitz continuous on $\mathcal{S} \times \mathcal{A}$ for every measurable bounded function $f(\cdot)$ on $\mathcal{S}$.*

An optimal policy π^* to (1) solves the stochastic dynamic program (SDP)

$$V^*(s) = \min_{a \in \mathcal{A}} \left[c(s,a) + \gamma \mathbb{E}_p\left[V^*(s')|s,a\right]\right], \quad \forall s \in \mathcal{S}, \tag{2}$$

where $\mathbb{E}_p[V^*(s')|s,a] := \int_{\mathcal{S}} V^*(s')p(s'|s,a)ds'$ denotes expectation defined by $p(\cdot|s,a)$. In theory, this SDP can be cast as an infinite linear program (Hernández-Lerma and Lasserre 1996, ch. 6)

$$\begin{aligned} \max_{V} \ & \mathbb{E}_q\left[V(s)\right] \\ \text{s.t.} \ & V(s) - \gamma \mathbb{E}_p\left[V(s')|s,a\right] \le c(s,a), \quad \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \end{aligned} \tag{3}$$

Here $q(\cdot)$ is a continuous probability density function on $\mathcal{S}$, $\mathbb{E}_q[V(s)] := \int_{\mathcal{S}} V(s)q(s)ds$ the expectation defined by this density function, and $V(s)$ one of the infinitely many variables of the linear program (3) corresponding to state s, which serves as a surrogate to the MDP value function. The objective function of this linear program maximizes the expectation of $V(\cdot)$ with respect to $q(\cdot)$. The constraints are obtained by replacing the minimization over actions in (2) by a set of inequalities.

3.2. Overview of ALPs and their solution

Solving linear program (3) is typically intractable because (i) it has infinitely many variables and constraints, and (ii) each constraint may include a high-dimensional expectation. A well-known approach to deal with the infinitely many variables is to approximate the function $V(\cdot)$ in (3) by a linear combination of B "basis" functions $\phi_b : \mathcal{S} \mapsto \mathbb{R}$, $b = 1, \ldots, B$, that is, $V(s) \approx \tau + \sum_{b=1}^{B} \theta_b \phi_b(s)$, where $\theta := (\theta_1, \ldots, \theta_B) \in \mathbb{R}^B$ are basis function weights and $\tau \in \mathbb{R}$ defines a constant. We make the benign Assumption 2.

ASSUMPTION 2. *The basis functions $\phi_b(\cdot)$, $b = 1, \ldots, B$, are Lipschitz continuous over $\mathcal{S}$.*

Employing the above value function approximation on (3) yields the following ALP (Schweitzer and Seidmann 1985, de Farias and Van Roy 2003):

$$\begin{aligned} F := \max_{(\tau,\theta)\in\mathbb{R}^{B+1}} \ & \tau + \sum_{b=1}^{B} \theta_b \mathbb{E}_q[\phi_b(s)] \\ \text{s.t.} \quad & (1-\gamma)\tau + \sum_{b=1}^{B} \theta_b \left(\phi_b(s) - \gamma \mathbb{E}_p[\phi_b(s')|s,a]\right) \le c(s,a), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A}. \end{aligned} \tag{4}$$

The vector variable θ represents the ALP basis function weights while the scalar variable τ can be interpreted as a constant shift ensuring the feasibility of θ in the constraints of (4).

An optimal ALP solution (τ^*, θ^*) defines a VFA $\tau^* + \sum_{b=1}^{B} \theta_b^* \phi_b(\cdot)$, which is known to be a lower bound on the MDP value function $V^*(\cdot)$ due to the nature of the ALP constraints (de Farias and Van Roy 2003). Thus, the ALP optimal objective function value F is a lower bound on the optimal policy cost and useful for assessing the optimality gaps of feasible policies from ALPs and other heuristics, which provide upper bounds on the optimal policy cost. The ALP policy decision at stage s can be computed by solving the following problem obtained by replacing the value function in the right-hand side of SDP (2) by the ALP VFA:

$$\min_{a\in\mathcal{A}} \left[c(s,a) + \gamma \sum_{b=1}^{B} \theta_b \mathbb{E}_p[\phi_b(s')|s,a] \right]. \tag{5}$$

We omit the term $\gamma\tau$ in (5) as it is a constant and does not affect the optimal action, that is, the ALP policy depends only on the basis function weights θ.

The number of ALP variables are manageable but solving the ALP directly is challenging since it has infinitely many constraints and contains expectations that are potentially difficult to compute. The expectation appearing in the ALP objective needs to be computed only once and is with respect to a probability density function $q(\cdot)$ of our choosing. Therefore, it is typically manageable. However, the expectation appearing in the ALP constraints needs to be computed for every state and action, and is with respect to the MDP probability transition function $p(\cdot|s,a)$. Row generation (Trick and Zin 1997, Adelman 2004, 2007) and constraint sampling (de Farias and Van Roy 2004, Farias and Van Roy 2006, Restrepo 2008, ch. 4, and Sun et al. 2016) are common solution strategies to address one or both of these challenges.

Row generation involves solving a relaxation of the ALP containing only a subset of its constraints. Given an optimal solution of this relaxation, we identify the most violated ALP constraint, if any, by solving the following separation problem: $\min_{(s,a)\in\mathcal{S}\times\mathcal{A}}[c(s,a) - \sum_{b=1}^{B}\theta_b(\phi_b(s) - \gamma\mathbb{E}_p[\phi_b(s')|s,a])]$. If no violated constraint is identified, then the incumbent solution is optimal to the ALP and we terminate. If a violated constraint exists, it will cut off the incumbent solution. We add this constraint to strengthen the relaxation of the ALP and repeat the procedure until termination. Row generation is guaranteed to converge asymptotically (Adelman 2007) and solve the exact ALP model, thus providing an optimal ALP solution and a lower bound on the optimal policy cost. However, its computational feasibility depends on the separation problem, which may be non-convex due to the structure of the cost function, nature of state transitions, and/or the form of basis functions. Under these conditions, expectations present in the objective function of the separation problem may not be available in closed form. Indeed, one could replace these expectations by their sample average approximations and solve the separation problem heuristically, while forgoing the termination guarantee and the ALP lower bound.

The constraint sampling technique for tackling an ALP pre-samples state-action pairs by simulating a baseline heuristic control policy and solves a version of the ALP formulated on this sampled state-action space. de Farias and Van Roy (2004) establish a sample complexity for this approach under an idealized sampling distribution – loosely speaking one that depends on the unknown optimal policy. However, the quality of the computed solution depends on the actual sampling distribution employed. Given a sampling distribution, the constraint sampling approach is easy to implement but does not provide a mechanism to obtain a lower bound on the optimal policy cost.

4. Saddle point ALP reformulation

We present a primal-dual reformulation of ALPs in this section. The first step of our reformulation exploits the structure of the ALP constraints to obtain an equivalent non-convex saddle-point model. To alleviate non-convexities, we subsequently lift this saddle-point problem to a higher

dimensional variable space and add a regularization term. The result is an infinite dimensional convex saddle-point model that we leverage in §5 to find near optimal ALP basis function weights and a lower bound on the optimal policy cost.

The ALP constraints in (4) can be replaced by a single constraint that places an upper bound on τ, that is

$$\tau \leq \bar{\tau}(\theta) := \frac{1}{1-\gamma} \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left\{ c(s,a) - \sum_{b=1}^{B} \theta_b(\phi_b(s) - \gamma \mathbb{E}_p\left[\phi_b(s')|s,a\right]) \right\}. \tag{6}$$

Since the ALP objective function maximizes τ, the inequality (6) must hold as an equality in an optimal solution. The optimization in the definition of $\bar{\tau}(\theta)$ is equivalent to the one encountered in row generation for computing the most violated ALP constraint (see §3.2). Specifically, if the term $\bar{\tau}(\theta)$ is strictly negative, its absolute value equals the largest ALP constraint violation of the solution $(0,\theta)$ scaled by $1/(1-\gamma)$. Otherwise this absolute value equals the smallest scaled ALP constraint slack of the solution $(0,\theta)$. Therefore, setting τ equal to the right-hand side value of (6) amounts to moving from $(0,\theta)$ to $(\bar{\tau}(\theta),\theta)$, where the latter solution lies on the boundary of the ALP feasible set. Indeed, if the ALP basis functions span the MDP value function, then there exists a weight vector θ for which $\bar{\tau}(\theta)$ is zero.

Based on the above observation, we can substitute for τ in the ALP objective in (4) with $\bar{\tau}(\theta)$ to obtain the following saddle-point reformulation:

$$F = \max_{\theta\in\Theta} \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a), \tag{7}$$

$$\text{where} \quad f(\theta,s,a) := \frac{1}{1-\gamma} \left\{ c(s,a) - \sum_{b=1}^{B} \theta_b(\phi_b(s) - \gamma \mathbb{E}_p\left[\phi_b(s')|s,a\right]) \right\} + \sum_{b=1}^{B} \theta_b \mathbb{E}_q\left[\phi_b(s)\right],$$

and $\Theta \subseteq \mathbb{R}^B$ is a compact set containing the optimal ALP solution in its interior. A compact domain Θ is commonly assumed in theoretical analyses found in the approximate dynamic programming and stochastic programming literature (e.g., Shapiro et al. 2009, Birge and Louveaux 2011, Desai et al. 2012). The primal (outer) optimization in (7) chooses the VFA weight vector θ and the saddle-point objective $f(\cdot,s,a)$ is linear in these weights. The dual (inner) optimization over the state-action pairs finds the most violated ALP constraint for a given weight vector θ, which is a potentially non-convex problem. (The last term defining $f(\theta,s,a)$ is independent of the state-action pair.)

Next, we establish in Proposition 1 that the non-convex finite-dimensional dual minimization in (7) can be replaced by a linear and infinite dimensional infimum over all continuous probability density functions specified over the state-action set. The linear space of probability density functions defined on $\mathcal{S}\times\mathcal{A}$ are denoted by $\mathcal{Y}$. Formally, $\mathcal{Y} := \{y : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}_{++} | \int_{\mathcal{S}\times\mathcal{A}} y(s,a)d(s,a) = 1\}$. We denote by $\mathbb{E}_y[f(\theta,s,a)]$ expectation $\int_{\mathcal{S}\times\mathcal{A}} f(\theta,s,a)y(s,a)d(s,a)$, which is well defined because $\mathcal{S}\times\mathcal{A}$ is compact and assumptions 1 and 2 are true.

PROPOSITION 1. *It holds that*

$$F = \max_{\theta\in\Theta} \inf_{y\in\mathcal{Y}} \mathbb{E}_y[f(\theta,s,a)]. \tag{8}$$

For a given θ, by virtue of optimality, the distribution y in the inner optimization of (8) must be chosen such that the set of state-action pairs corresponding to the largest constraint violation have positive density. This set could contain only a finite number of state-action pairs, a scenario captured only if y represents a discrete distribution. Such a discrete distribution does not belong to the set $\mathcal{Y}$. We use an infimum in (8) for this reason.

To replace the infimum in (8) by a minimum, we add a Kullback-Liebler (KL) regularization term $D(y,p_u) := \mathbb{E}_y\left[\log \frac{y(s,a)}{\bar{p}}\right]$ to the objective in (8), where y is a continuous probability density function and p_u is a uniform probability density function on $\mathcal{S}\times\mathcal{A}$ with constant density value $\bar{p} = (\int_{\mathcal{S}\times\mathcal{A}} 1 d(s,a))^{-1}$, that is, $p_u(s,a) = \bar{p}$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. It is well known that the KL divergence term $D(\cdot,\cdot)$ measures the difference between continuous probability distributions (Bishop 2006, ch. 1). The resulting regularized saddle-point problem is

$$F(\lambda) := \max_{\theta\in\Theta} \min_{y\in\mathcal{Y}} \left[\mathbb{E}_y\left[f(\theta,s,a)\right] + \lambda D(y,p_u)\right], \tag{9}$$

where $\lambda\in(0,1]$ is the regularization parameter. The term $\mathbb{E}_y[f(\theta,s,a)]$ is linear in both y and θ and the KL divergence term $\lambda D(\cdot,\cdot)$ is convex in y. Therefore, (9) is a convex saddle-point problem.

Proposition 2 confirms that a dual minimizer of (9) lies in $\mathcal{Y}$ for any $\theta\in\Theta$, and in addition, relates $F(\lambda)$ to F. Let $\exp(\cdot)$ denote the exponential function.

PROPOSITION 2. *For a given θ and $\lambda\in(0,1]$, a dual minimizer in (9) is*

$$y^*_{\lambda,\theta}(s,a) := \frac{\exp\left(-f(\theta,s,a)/\lambda\right)}{\int_{\mathcal{S}\times\mathcal{A}} \exp\left(-f(\theta,s,a)/\lambda\right) d(s,a)},$$

and $y^*_{\lambda,\theta}\in\mathcal{Y}$. Moreover, $F\le F(\lambda)$ for all $\lambda\in(0,1]$ and for a given $\alpha>0$ there exists a sufficiently small $\lambda\in(0,1]$ such that $F(\lambda)-F\le\alpha$.

Intuitively, adding the KL divergence term in (9) ensures $y^*_{\lambda,\theta}\in\mathcal{Y}$, because the uniform density function p_u belongs to the desired set $\mathcal{Y}$ and the regularization term discourages deviations of y from this uniform density. The inequality $F\le F(\lambda)$ is driven by the non-negativity of $\lambda D(y,p_u)$. The key consequence of Proposition 2 is that the saddle-point problem (9) closely approximates (7) and, thus the ALP, when λ is sufficiently small. Accordingly, this reformulation shifts the difficulty of solving finite dimensional non-convex optimization problems for finding the most violated ALP constraint in (7) to (i) finding a small λ and (ii) solving the infinite dimensional convex optimization problem in (9) involving an additional expectation over the state-action space.

5. PSMD algorithm and guarantees

In this section, we discuss a primal-dual first-order approach for solving the saddle-point problem (9). In §5.1, we present our main algorithm for finding a near-optimal ALP solution under a generic (unspecified) stopping criterion. In §5.2, we discuss specific stopping criteria for this algorithm and a lower bound on the optimal policy cost.

5.1. Algorithm

Mirror descent methods are popular in the literature for finding near-optimal solutions to saddle-point problems. A common template for such methods involves computing gradients of the objective function with respect to primal and dual variables and using them to move towards a new solution. When applying these gradient based moves, feasibility of the updated solution is ensured using projections. The primal optimization problem in the ALP saddle-point formulation (9) can be handled by mirror-descent methods in the literature, but the dual optimization problem in this formulation and its associated high-dimensional expectations are challenging to tackle.

Below, we develop a primal-dual[1] stochastic mirror-descent method for computing an α-optimal ALP solution, that is, a vector (τ, θ) satisfying the constraints in (4) and inequality $F - (\tau + \sum_{b=1}^{B} \theta_b \mathbb{E}_q [\phi_b(s)]) \leq \alpha$. Our method employs sample average approximations to handle high dimensional expectations and features a novel dual update and an automated scheme for choosing the regularization constant λ in (9). We begin by describing this method, which we already labeled as proximal stochastic mirror-descent (PSMD), and then state its correctness.

The steps of PSMD are summarized in Algorithm 1. The inputs to this algorithm are the values of a stopping tolerance TOL (e.g., bound on the number of iterations or CPU time), the initial regularization coefficient $\lambda_0 \in (0, 1]$, and the initial step length $\eta_0 > 0$. Step 1 initializes the iteration counter t to zero, the averaged and time zero primal variables $\bar{\theta}$ and θ_0 to a vector of zeros; the averaged and time zero dual variables $\bar{y}$ and y_0, respectively, to be the uniform density p_u; and the averaged regularization coefficient $\bar{\lambda}$ to λ_0. PSMD executes steps 3 to 7 until a generic stopping criterion $\text{STOP}(t, \bar{\theta}, \bar{y}, \bar{\lambda}, \text{TOL})$ is satisfied. We discuss each of these steps below at iteration t.

- Step 3 executes the primal update, which is similar to the one found in Nemirovski et al. (2009). Given θ_t at iteration t, an idealized update to obtain vector θ_{t+1} is

$$\theta_{t+1} := \arg\min_{\theta \in \Theta} \left[\eta_t \left\langle \theta_t - \theta, \mathbb{E}_{y_t} \left[\nabla_\theta f(\theta_t, s, a) \right] \right\rangle + \frac{1}{2} \|\theta - \theta_t\|_2^2 \right], \tag{13}$$

where η_t represents the step length at iteration t, $\langle \cdot, \cdot \rangle$ the dot product, ∇_θ the gradient operator with respect to θ, and $\| \cdot \|_2$ the 2-norm. The optimization problem (13) can be viewed as a gradient ascent step followed by a projection, which is more apparent when (13) is rewritten in the following equivalent form[2]:

$$\theta_{t+1} := \arg\min_{\theta \in \Theta} \frac{1}{2} \left\| \theta - \left(\theta_t + \eta_t \mathbb{E}_{y_t} \left[\nabla_\theta f(\theta_t, s, a) \right] \right) \right\|_2^2 .$$

Algorithm 1 PSMD

Input: Stopping tolerance TOL, regularization constant $\lambda_0 \in (0,1]$, and step length $\eta_0 > 0$.

1: Set t $= 0$, $\bar{\theta} = \theta_0 = (0,\ldots,0) \in \mathbb{R}^B$, $\bar{y} = y_0 = p_u$, and $\bar{\lambda} = \lambda_0$.

2: **while** STOP$(t, \bar{\theta}, \bar{y}, \bar{\lambda}, \text{TOL}) =$ FALSE **do**

3: Perform primal update: Sample state-action pair $(\hat{s}, \hat{a})$ from the density function y_t, generate next stage state $\check{s}$ from the MDP density function $p(\cdot|\hat{s}, \hat{a})$, and set

$$\theta_{t+1} = \arg\min_{\theta \in \Theta} \left[\eta_t \left\langle \theta_t - \theta, \nabla_\theta \hat{f}(\theta_t, \hat{s}, \hat{a}) \right\rangle + \frac{1}{2} \|\theta - \theta_t\|_2^2 \right], \tag{10}$$

where

$$\hat{f}(\theta, \hat{s}, \hat{a}) := \frac{1}{1-\gamma} \left\{ c(\hat{s}, \hat{a}) + \gamma \sum_{b=1}^{B} \theta_b \phi_b(\check{s}) - \sum_{b=1}^{B} \theta_b \phi_b(\hat{s}) \right\} + \sum_{b=1}^{B} \theta_b \mathbb{E}_q \left[\phi_b(s) \right]. \tag{11}$$

4: Perform dual update:

$$y_{t+1} \propto y_t^{(1+\eta_t \lambda_t)^{-1}} \exp\left(-\eta_t \hat{f}(\theta_t, \cdot, \cdot)/(1+\eta_t \lambda_t) \right). \tag{12}$$

5: Update step length and regularization coefficient:

$$\eta_{t+1} = \eta_0/\sqrt{t+2}; \quad \lambda_{t+1} = \lambda_0/\sqrt{t+2}.$$

6: Compute averaged regularization coefficient and primal and dual solutions:

$$\bar{\lambda} = \frac{1}{\sum_{t'=0}^{t+1} \eta_{t'}} \sum_{t'=0}^{t+1} \eta_{t'} \lambda_{t'}; \quad \bar{\theta} = \frac{1}{\sum_{t'=0}^{t+1} \eta_{t'}} \sum_{t'=0}^{t+1} \eta_{t'} \theta_{t'}; \quad \bar{y} = \frac{1}{\sum_{t'=0}^{t+1} \eta_{t'}} \sum_{t'=0}^{t+1} \eta_{t'} y_{t'}.$$

7: Update iteration counter: $t = t+1$.

8: **end while**

Output: $\bar{\theta}$ and $\bar{y}$.

Here, $\theta_t + \eta_t \mathbb{E}_{y_t} \left[\nabla_\theta f(\theta_t, s, a) \right]$ attempts to improve the objective value by starting at θ_t and moving along the gradient[3] of $\mathbb{E}_y \left[f(\theta, s, a) \right]$, which is $\mathbb{E}_y \left[\nabla_\theta f(\theta, s, a) \right]$. The updated solution θ_{t+1} is the Euclidean projection of $\theta_t + \eta_t \mathbb{E}_{y_t} \left[\nabla_\theta f(\theta_t, s, a) \right]$ onto Θ.

The convex program (13) can be solved after the potentially high dimensional expectations in the term $\mathbb{E}_{y_t} \left[\nabla_\theta f(\theta_t, s, a) \right]$ are handled as follows. First, we approximate the expectation with respect to y_t by an unbiased point estimate $\nabla_\theta f(\theta_t, \hat{s}, \hat{a})$ defined on the sample $(\hat{s}, \hat{a})$ from the density function y_t. Second, we substitute $f(\theta_t, \hat{s}, \hat{a})$ with the unbiased point estimate $\hat{f}(\theta_t, \hat{s}, \hat{a})$ obtained by replacing the expectation $\mathbb{E}_p[\phi_b(s)|\hat{s}, \hat{a}]$ in $f(\theta_t, \hat{s}, \hat{a})$ with the sample average approximation $\phi_b(\check{s})$ that relies on a single state $\check{s}$ generated from the MDP transition density function $p(\cdot|\hat{s}, \hat{a})$. We assume

expectation $\mathbb{E}_q[\phi_b(s)]$ in (11) can be evaluated because the probability distribution q is a user choice and this expectation is not conditioned on the state or action, that is, it needs to be computed only once. The term $\nabla_\theta \hat{f}(\theta_t, \hat{s}, \hat{a})$ constructed in the preceding manner is an unbiased *stochastic gradient* of $\mathbb{E}_{y_t}[f(\theta_t, s, a)]$ with respect to θ. Using it in (13) results in the update (10). The optimization in (10) can be handled by an off-the-shelf convex optimization solver (e.g. CPLEX).

• Step 4 executes the dual update, which is non-traditional and differs from the one provided by Nemirovski et al. (2009). This difference is because the saddle-point objective $\mathbb{E}_y\left[\hat{f}(\theta_t, s, a)\right] + \lambda D(y, p_u)$, with $f(\theta_t, s, a)$ replaced by its unbiased point estimate $\hat{f}(\theta_t, s, a)$, is convex in the dual decision variable y due to the presence of the function $\lambda D(\cdot, p_u)$ and this variable is infinite dimensional. Therefore, unlike the θ-update, using a gradient of the objective with respect to y amounts to employing a linear approximation of the convex function $\lambda D(\cdot, p_u)$ in the y-update, which is weak in general. To resolve this issue, we use $\lambda D(y, p_u)$ directly in the y-update and define the move to y_{t+1} by

$$y_{t+1} \in \arg\min_{y \in \mathcal{Y}} \left[\eta_t \left(\mathbb{E}_{y-y_t}[\hat{f}(\theta_t, s, a)] + \lambda_t D(y, p_u)\right) + D(y, y_t)\right], \tag{14}$$

where $D(y, y_t) := \mathbb{E}_y\left[\log \frac{y(s,a)}{y_t(s,a)}\right]$ is a KL prox-function that is added as part of the dual update. In other words, it plays an analogous role to the Euclidean prox-function in the primal update.

Solving (14) requires handling the potentially high dimensional expectations with respect to y in its objective. These expectations cannot be replaced by sample average approximations apriori because the minimization is with respect to y and such replacement would result in a biased update. Fortunately, we can apply the Karush-Kuhn-Tucker optimality conditions to (14) and obtain $y_{t+1}(s, a)$ in closed form[4]:

$$y_{t+1}(s, a) = \frac{(y_t(s, a))^{(1+\eta_t\lambda_t)^{-1}} \exp\left(-\eta_t \hat{f}(\theta_t, s, a)/(1 + \eta_t\lambda_t)\right)}{\int_{\mathcal{S}\times\mathcal{A}} (y_t(s, a))^{(1+\eta_t\lambda_t)^{-1}} \exp\left(-\eta_t \hat{f}(\theta_t, s, a)/(1 + \eta_t\lambda_t)\right) d(s, a)}. \tag{15}$$

The normalization constant in the denominator includes a high-dimensional integral but its computation can be avoided when sampling from y_{t+1}. For example, methods such as Metropolis-Hastings (see, Robert and Casella 2004, ch. 7) use the "relative density" of different state-action pairs for sampling, which is proportional to the numerator of (15), and not affected by the denominator. Thus, the PSMD y-update becomes (12). This proportionality expression models an unbiased *functional stochastic gradient* of $\mathbb{E}_y[f(\theta_t, s, a)]$ with respect to y.

• Step 5 updates the step length and the regularization coefficient. Both these quantities decrease with the number of iterations.

- Step 6 computes moving averages of regularization coefficients and primal and dual variable values encountered by PSMD. As PSMD iterations progress and the regularization coefficient and step length decrease, the values of the averaged primal and dual variables intuitively become "closer" to being primal and dual optimal in the saddle-point problem (7).

At termination, PSMD returns a VFA weight vector $\bar{\theta}$ and a state-action density function $\bar{y}$. The suboptimality of this solution can be assessed using min-max duality theory (Fan 1953). Define the following primal and dual functions:

$$\mathcal{P}(y) := \max_{\theta \in \Theta} \mathbb{E}_y[f(\theta, s, a)]; \tag{16}$$

$$\mathcal{D}(\theta) := \inf_{y \in \mathcal{Y}} \mathbb{E}_y[f(\theta, s, a)]. \tag{17}$$

It is well known that $\mathcal{D}(\theta) \leq F \leq \mathcal{P}(y)$, that is, $\mathcal{P}(y) - \mathcal{D}(\theta)$ is the duality gap of the saddle-point problem (7). Theorem 1 establishes a high probability big-$\mathcal{O}$ iteration complexity for PSMD to return an α-optimal ALP solution.

THEOREM 1. *Given $\alpha > 0$ and $\delta \in (0,1)$, the PSMD primal-dual pair $(\bar{\theta}, \bar{y})$ satisfies inequality $\mathcal{P}(\bar{y}) - \mathcal{D}(\bar{\theta}) \leq \alpha$ with probability of at least $1-\delta$ in at most*

$$\mathcal{O}\left(\frac{1}{\alpha^2} \log\left(\frac{1}{\alpha}\right) \log\left(\frac{1}{\delta}\right)\right)$$

iterations.

Following standard convention, our big-$\mathcal{O}$ iteration complexity in Theorem 1 highlights the worst-case convergence behavior of PSMD, which occurs when α and/or δ are small, and excludes positive constants and lower order terms (Bach and Shallit 1996). If α and δ are large, the big-$\mathcal{O}$ expression will become small (possibly negative) and other terms in the overall iteration complexity expression will dominate to ensure the total number of iterations is positive (see (EC.25) on page 10 of the Electronic Companion for the complete iteration bound expression).

5.2. Stopping criteria and lower bound

Applying PSMD requires choosing a stopping criterion STOP in Algorithm 1. The statement of first order methods typically assume termination after a number of iterations T, which plays the role of the stopping tolerance TOL, equal to the iteration complexity of the algorithm (e.g., Hazan and Kale 2014). Unfortunately, worst-case iteration bounds are often highly conservative to provide a practical stopping criterion. Therefore, T is set in a heuristic manner or replaced by a total run time threshold during implementation. Under such stopping criteria, PSMD will indeed return basis function weights $\bar{\theta}$ and a state-action density function $\bar{y}$ but it is apriori unclear how a lower bound on the optimal policy cost can be estimated. Furthermore, while terminating PSMD based on

iteration or run time limits may suffice in many practical situations, it would still be useful to have a mechanism to terminate PSMD using a target optimality gap when heuristic criteria fail to work.

To address the above shortcomings, we present an approximation of the exact PSMD primal-dual gap $\mathcal{P}(\bar{y}) - \mathcal{D}(\bar{\theta})$ in Theorem 1. Consider the lower bound $\mathcal{D}(\bar{\theta})$ on F, which is challenging to compute because it involves an infimum over an open set with support over the state-action space. To avoid the infimum, we consider the following regularized version of $\mathcal{D}(\bar{\theta})$:

$$\min_{y \in \mathcal{Y}} E_y[f(\bar{\theta}, s, a)] + \bar{\lambda}\left(D(y, p_u) + \bar{C} + n\log(\bar{\lambda})\right), \tag{18}$$

where $\bar{\lambda}$ is the averaged regularization coefficient computed by PSMD; n is the dimension in the state-action space $\mathcal{S} \times \mathcal{A}$; the constant $\bar{C}$ is defined as

$$\bar{C} := \log(\bar{p}) - \log\left(\frac{\Gamma(n/2+1)}{\pi^{n/2} R^n}\right) - L(R + Q_{\mathcal{S}\times\mathcal{A}});$$

R is the radius of the largest ball that can be included in $\mathcal{S} \times \mathcal{A}$; $Q_{\mathcal{S}\times\mathcal{A}} := \max_{(s,a),(s',a') \in \mathcal{S}\times\mathcal{A}} \|(s,a) - (s',a')\|_2$ is the diameter of the set $\mathcal{S} \times \mathcal{A}$; $\Gamma(z)$ is the standard gamma function $\int_0^\infty x^{z-1}e^{-x}dx$ for $z > 0$; and L is the Lipchitz constant associated with $f(\cdot,\cdot,\cdot)$[5]. The optimization over the state-action space can be side-stepped because (18) has a closed-form optimal solution analogous to the optimization problem (14). Specifically, it can be verified that $y^*_{\bar{\lambda},\bar{\theta}}(s,a)$ defined in Proposition 2 is the optimal solution of (18). The objective function of (18) evaluated at $y^*_{\bar{\lambda},\bar{\theta}}$ becomes

$$\mathbb{E}_{y^*_{\bar{\lambda},\bar{\theta}}}[f(\bar{\theta}, s, a)] + \bar{\lambda} D(y^*_{\bar{\lambda},\bar{\theta}}, p_u) + \bar{\lambda}\bar{C} + n\bar{\lambda}\log(\bar{\lambda}),$$

where $\bar{\lambda} D(y^*_{\bar{\lambda},\bar{\theta}}, p_u)$ is non-negative. Theorem 2 shows that dropping $\bar{\lambda} D(y^*_{\bar{\lambda},\bar{\theta}}, p_u)$ yields a lower bound on the optimal policy cost:

$$l(\bar{\theta}) := \mathbb{E}_{y^*_{\bar{\lambda},\bar{\theta}}}[f(\bar{\theta}, s, a)] + \bar{\lambda}\bar{C} + n\bar{\lambda}\log(\bar{\lambda}).$$

Specifically, (i) $l(\bar{\theta})$ lower bounds $\mathcal{D}(\bar{\theta})$ and (ii) $l(\bar{\theta})$ can be combined with an upper bound $\mathcal{P}(\bar{y})$ on F to deliver performance guarantees analogous to Theorem 1, that is, $l(\bar{\theta})$ is not too conservative.

THEOREM 2. *At any iteration of PSMD, it holds that $l(\bar{\theta}) \le \mathcal{D}(\bar{\theta})$. Moreover, given $\alpha > 0$ and $\delta \in (0,1)$, the PSMD primal-dual pair $(\bar{\theta}, \bar{y})$ satisfies inequality $\mathcal{P}(\bar{y}) - l(\bar{\theta}) \le \alpha$ with probability of at least $1-\delta$ in at most*

$$\mathcal{O}\left(\frac{1}{\alpha^2}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{1}{\delta}\right)\right)$$

iterations. Moreover, $l(\bar{\theta})$ is a lower bound on F such that $F - l(\bar{\theta}) \le \alpha$.

The results in Theorem 2 have important algorithmic implications. Since F is a lower bound on the optimal policy cost, $l(\bar{\theta})$ also defines a lower bound on this cost. Availability of a lower bound suggests a simple PSMD implementation strategy: terminate PSMD based on a run time or iteration limit, simulate the ALP policy associated with the incumbent $\bar{\theta}$ to obtain an upper bound, evaluate the lower bound $l(\bar{\theta})$, and use it to compute an optimality gap. If this optimality gap is small, the solution is acceptable. We found this strategy to work well in our numerical study in §7. Nevertheless, when the optimality gap is large, it is unclear if this gap is due to the ALP providing a poor VFA or PSMD being stopped prematurely, that is, the difference $F - l(\bar{\theta})$ could be large. The latter concern can be eliminated by stopping PSMD when the inequality $\mathcal{P}(\bar{y}) - l(\bar{\theta}) \leq \alpha$ holds for a desired $\alpha > 0$, which is guaranteed to happen by Theorem 2. Moreover, this inequality may be satisfied in a significantly fewer number of iterations than our worst-case iteration bound while still ensuring an α-optimal ALP solution. PSMD can also be modified to terminate by checking a relative optimality gap with respect to an ALP feasible solution. We discuss this PSMD variant in the Electronic Companion EC.2.

6. PSMD implementation

In this section, we discuss several implementation details of PSMD. We begin by providing guidelines for making PSMD design choices, then focus on the execution of Algorithm 1, and finally discuss the evaluation of the bounds $l(\bar{\theta})$ and $\mathcal{P}(\bar{y})$. These guidelines are used in the computational study described in §7.

Specifying a VFA in PSMD requires choosing basis functions and a compact support Θ for basis function weights. The former choice is often based on the structure of the cost and state transition functions. The latter set can be constructed by applying regression on the costs from simulating a baseline policy and defining a large box around the resulting weight vector. The compact set Θ can itself be interpreted as an additional restriction on the VFAs to avoid large basis function weights. Such restrictions are common in the extant regression literature (Tibshirani 1996) and consistent with ALP being a constrained regression (Lemma 1 in de Farias and Van Roy 2003; also see Petrik et al. 2010 for the use of regularization in an ALP).

The inputs to PSMD are a stopping tolerance (TOL), a regularization coefficient (λ_0), and a step length (η_0). Choosing TOL as a run time limit is typically based on the user determining an acceptable time budget. Selecting TOL instead to be an absolute optimality gap α entails an assessment of the scale of the problem. This can be done, for example, by simulating a baseline heuristic such as a myopic policy or a look-ahead policy and choosing α equal to a certain percentage of the expected cost under this policy.

The remaining input parameters λ_0 and η_0 affect empirical performance even though PSMD terminates regardless of these choices. Intuitively, a large value of λ_0 places more weight on the

KL-divergence term in the objective function of (9) and, hence, prevents the dual solution y_t in each iteration moving away from p_u. Consequently, for a very large λ_0, PSMD bound changes across iterations may be too slow. On the other extreme, choosing λ_0 to be very small results in erratic bound changes in each iteration. The behavior of the PSMD bounds with respect to η_0 is the opposite; that is, a small value causes slow bound changes, while a large value results in erratic bound fluctuations. Some tuning of these parameters is needed via a training step, as with most first-order and machine learning methods (see, TensorFlow 2018). For our computational experiments in §7, we create a grid of possible values for λ_0 and η_0 and then run PSMD for a limited number of iterations to evaluate the bound improvement. We select the best parameter values from this training set as λ_0 and η_0 and execute PSMD until termination.

When executing PSMD, sample average approximations of expectations need to be evaluated in steps 3 and 4 of Algorithm 1. These unbiased approximations use a single sample, which suffices to establish our theoretical guarantees, but such point estimates may exhibit high variance and affect empirical performance. Low-variance versions of these sample average approximations can be constructed by using more samples. Specifically, the term $\hat{f}(\theta,\hat{s},\hat{a})$ can be replaced by $\hat{f}^N(\theta,\hat{s},\hat{a})$ based on N samples from $p(\cdot|\hat{s},\hat{a})$ in set $\{s_n, n=1,\ldots,N\}$:

$$\hat{f}^N(\theta,\hat{s},\hat{a}) := \frac{1}{1-\gamma}\left\{c(\hat{s},\hat{a}) + \gamma\sum_{b=1}^{B}\theta_b\left(\frac{1}{N}\sum_{n=1}^{N}\phi_b(s_n)\right) - \sum_{b=1}^{B}\theta_b\phi_b(\hat{s})\right\} + \sum_{b=1}^{B}\theta_b\mathbb{E}_q\left[\phi_b(s)\right].$$

Subsequently, a low-variance stochastic gradient $\nabla_\theta\hat{f}(\theta_t,\hat{s},\hat{a})$ can be obtained by leveraging $\hat{f}^N(\theta_t,s,a)$ and generating H samples $\{(\hat{s}_h,\hat{a}_h), h=1,\ldots,H\}$ from y_t. The resulting stochastic gradient $\frac{1}{H}\sum_{h=1}^{H}\nabla_\theta\hat{f}^N(\theta_t,\hat{s}_h,\hat{a}_h)$ can be used in the primal update (10) in lieu of $\nabla_\theta\hat{f}(\theta_t,\hat{s},\hat{a})$. This gradient relies on samples from both the state-action distribution y_t and the MDP state transition distribution $p(\cdot|\hat{s},\hat{a})$. The distribution y_t has the closed form (15), which, as discussed in §5, requires a technique such as Metropolis Hastings to sample from its unscaled probability density function (its numerator) and avoid evaluating the normalizing constant (its denominator) containing a high-dimensional integral. In contrast, sampling from $p(\cdot|\hat{s},\hat{a})$ is straightforward when its density function takes a well-known form. The values of N and H can be selected experimentally using a training strategy similar to the one discussed for the choices of η_0 and λ_0.

Next, we discuss the evaluation of $l(\bar{\theta})$, which provides a lower bound on F and hence the optimal policy cost. Computing this bound requires evaluating high-dimensional expectations. However, unlike steps 3 and 4 of Algorithm 1, which focus on computing unbiased gradients, we are interested here only in a high-confidence bound estimate. Results from the sample average approximation

literature can be leveraged for this purpose (see, Shapiro et al. 2009, ch. 5). Specifically, the sample average approximation of $l(\bar{\theta})$ is

$$\hat{l}(\bar{\theta}) := \mathbb{E}_{(H',N')}\left[\frac{1}{H'}\sum_{h=1}^{H'} \hat{f}^{N'}(\bar{\theta}, \hat{s}_h, \hat{a}_h)\right] + \bar{\lambda}\bar{C} + n\bar{\lambda}\log(\bar{\lambda}),$$

where N' samples are needed from $p(\cdot|s,a)$ for each of the H' samples generated from the density

$$\hat{y}_{\bar{\lambda},\bar{\theta}}(s,a) \propto \exp\left(-\hat{f}^{N'}(\bar{\theta}, s, a)/\bar{\lambda}\right). \tag{19}$$

The density function $\hat{y}_{\bar{\lambda},\bar{\theta}}$ is an optimal solution to $\min_{y\in\mathcal{Y}}[\mathbb{E}_y[\hat{f}^{N'}(\bar{\theta}, s, a)] + \bar{\lambda}D(y, p_u)]$ and its expression is equivalent to the definition of $y^*_{\bar{\lambda},\bar{\theta}}$ (see Proposition 2) with $f(\theta, s, a)$ replaced by $\hat{f}^{N'}(\theta, s, a)$. The Metropolis Hastings procedure can be used to generate state-action samples from $\hat{y}_{\bar{\lambda},\bar{\theta}}$ when evaluating $\hat{l}(\bar{\theta})$ and the outer expectation $\mathbb{E}_{(H',N')}$ can be viewed as averaging over independent realizations of these samples. Similarly, if $\mathcal{P}(\bar{y})$ needs to be computed, the optimization problem (16) can be replaced by

$$\hat{\mathcal{P}}(\bar{y}) := \mathbb{E}_{(H',N')}\left[\max_{\theta\in\Theta} \frac{1}{H'}\sum_{h=1}^{H'} \hat{f}^{N'}(\theta, \hat{s}_h, \hat{a}_h)\right],$$

where we use H' and N' samples from the densities $\bar{y}$ and $p(\cdot|s,a)$, respectively, to construct the sample average approximation. Proposition 3 establishes that $\hat{l}(\bar{\theta})$ and $\hat{\mathcal{P}}(\bar{y})$ provide valid bounds.

PROPOSITION 3 **(Proposition 5.6 in Shapiro et al. 2009)**. *It holds that*

$$\hat{l}(\bar{\theta}) \le \mathcal{D}(\bar{\theta}) \le \mathcal{P}(\bar{y}) \le \hat{\mathcal{P}}(\bar{y}).$$

Replacing expectation $\mathbb{E}_{(H',N')}$ in the definitions of $\hat{l}(\bar{\theta})$ and $\hat{\mathcal{P}}(\bar{y})$ by a finite number of independent trials yields valid bound estimates. Moreover, these estimates converge to $l(\bar{\theta})$ and $\mathcal{P}(\bar{y})$ when H', N', and the number of independent trials used to approximate $\mathbb{E}_{(H',N')}$ are all increased (see Shapiro et al. 2009).

Finally, the value of the constant $\bar{C}$ needs to be specified to compute $\hat{l}(\bar{\theta})$. The parameters $\bar{p}$, R, and $Q_{\mathcal{S}\times\mathcal{A}}$ appearing in its definition are easy to compute directly if the state-action space has a simple representation, which is the case in our computational study in §7. If not, bounds on these parameters can be computed by enveloping the state-action space by a box and then computing analogous parameters for this bounding box. The Lipschitz constant L used in $\bar{C}$ can be bounded by taking the gradient of the individual terms in $f(\theta, s, a)$ over the state-action space. Since the state-action space is compact, a bound on these gradients can be easily computed.

7. Computational study

In this section, we investigate the numerical performance of PSMD. We describe our PSMD setup in §7.1 and graphically illustrate its behavior on an inventory control problem with a two-dimensional state space in §7.2. We then compare PSMD with constraint sampling (ALP-CS) and other benchmarks in §7.3 and §7.4, respectively, on applications related to (i) perishable inventory control with partial backlogging and lead time and (ii) consumer energy storage management. We do not consider row generation in these applications as the separation problem is non-convex and/or the exact computation of expectations in the ALP constraints is not possible; see §3.2 for a discussion. All methods in this section were implemented in Matlab running on a 64-bit Microsoft Windows 10 machine with a 2.70 Ghz Intel Core i7-6820HQ CPU and 8GB of memory.

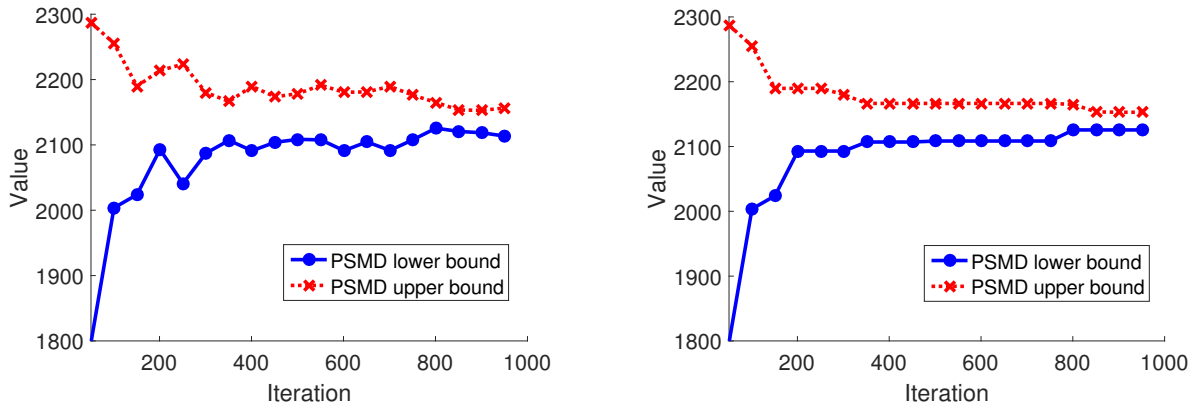
7.1. PSMD setup

We implemented PSMD using the guidelines discussed in §6. We describe here choices in the setup of PSMD that are common across all the applications we consider. We fixed the input parameters η_0 and λ_0 of Algorithm 1 equal to 0.1 and 0.0001, respectively. These choices relied on a training procedure where we observed the reduction in the PSMD lower bound estimate $\hat{l}(\bar{\theta})$ at the 0-th and the 50-th iterations. We selected H' and N' equal to 1,000 and 10,000, respectively, when computing the PSMD lower bound as the resulting standard error was less than 1% of the estimate. We stopped PSMD using a run time limit in order to facilitate the comparison with ALP-CS. We also selected the parameters H and N needed to define low-variance PSMD updates. Our instance-specific choices for the run time limit, H, and N are discussed in §§7.2-7.4. We used CPLEX to solve the convex optimization problem in the θ-update (10) and the Metropolis-Hastings sampling method implemented by the *mhsample* function in Matlab to implement the y update (12). We simulated the ALP policy (5) for given PSMD basis function weights $\bar{\theta}$ over trajectories of uncertainty generated in Monte Carlo simulation. Averaging the discounted total costs on these trajectories provides an upper bound estimate on the optimal policy cost. We chose the number of such trajectories so that the standard error of this upper bound estimate was less than 1% of its mean.

7.2. PSMD Illustration

We consider a single product inventory control system with partially backlogged demand and zero order lead time inspired by Nahmias and Smith (1994), Rabinowitz et al. (1995), and Benjaafar et al. (2010). We use this application to illustrate the behavior of the PSMD lower and upper bounds as well as the dual state-action distribution as functions of iterations.

MDP formulation and instance. The MDP state, denoted by s, represents on-hand inventory with negative values indicating backlogged orders. This state must adhere to an upper bound $u_s > 0$ and a lower bound $l_s < 0$, respectively, due to limited holding space and a maximum limit



(a) Bounds at current iteration.

(b) Best historic bounds.

Figure 1 PSMD lower and upper bounds as functions of iterations on an instance of the inventory control application with partial backlogging and no lead time.

on backlogging. The finite order quantity $a \in [0, \bar{a}]$ defines the MDP action, where $\bar{a}$ is the maximum order size. We assume an order is placed before stochastic demand G is revealed and that demand follows distribution P_G. Given order quantity and demand, the new inventory level is $s' := \min(\max(s+a-G, l_s), u_s)$, where we suppress the dependence of s' on s, a, and G for ease of notation. This transition function accounts for two events: (i) if the demand reduces the inventory level below the backlogging limit (i.e., $s+a-G < l_s$), then the excess demand $(l_s - s - a + G)$ is lost at a cost equal to c_l per unit; and (ii) if the inventory level exceeds the space limit (i.e., $s+a-G > u_s$), the additional units $(s+a-G-u_s)$ are disposed at a cost of c_d per unit. The per unit purchasing, holding, and backlogging costs are c_p, c_h, and c_b, respectively. Using these definitions, the MDP cost function at state s and action a is

$$c(s,a) = c_p a + c_h \mathbb{E}[(s')_+] + c_b \mathbb{E}[(-s')_+] + c_d \mathbb{E}[(s+a-G-u_s)_+] + c_l \mathbb{E}[(l_s - s - a + G)_+],$$

where $(s)_+ = \max\{0, s\}$ and expectations are with respect to P_G. The MDP transition density $p(s'|s,a)$ is determined by P_G and the nonlinear state transition function s'.

We consider an instance where γ equals 0.95; l_s and u_s are chosen as -10 and 10, respectively; P_G is a truncated normal distribution on the interval $[0, 10]$ with mean and standard deviation of 5 and 2, respectively; and c_p, c_d, c_l, c_h, and c_b are equal to 20, 10, 100, 2, and 10, respectively.

Results. We used PSMD to solve an ALP formulated with polynomial basis functions that included the constant 1, the linear term s, and the quadratic term s^2. We chose H and N in PSMD via experimentation to be 10 and 50, respectively. Figure 1 displays the current and historic best PSMD bound estimates as functions of the number of iterations. Figure 1(a) shows an improving trend in both the lower and upper bounds with the number of iterations but these improvements

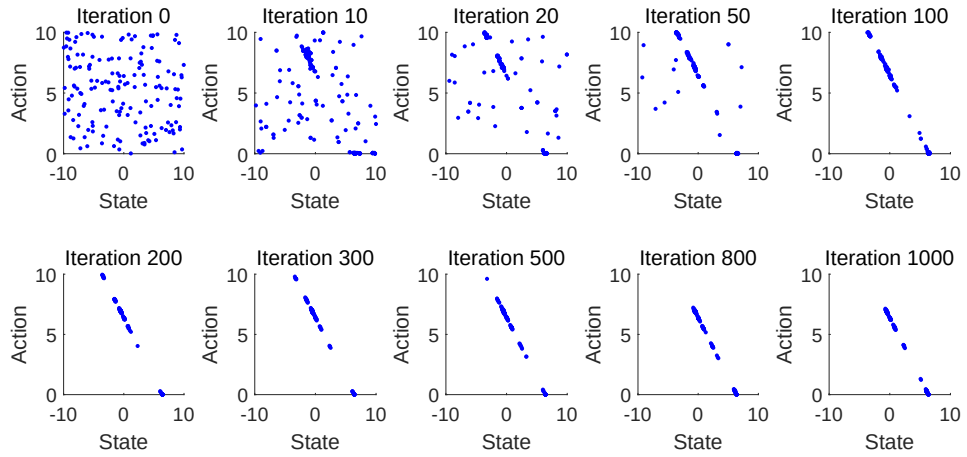


Figure 2 **State-action sampling distribution y_t computed by PSMD on an instance of the inventory control application with partial backlogging and no lead time.**

are not monotone. This non-monotonicity is expected of methods based on (stochastic) gradient updates. In contrast, the best incumbent historic PSMD lower and upper bounds shown in Figure 1(b) are indeed monotonically increasing and decreasing, respectively. The optimality gap computed using the best PSMD upper and lower bounds is 1.51% after 1000 PSMD iterations and 31 seconds.

The quality of the PSMD bounds displayed in Figure 1 depend on its VFA weight vectors θ_t at each iteration, which are updated using information from the state-action distribution y_t. This distribution attempts to learn regions of large ALP constraint violation. Figure 2 plots 100 state-action samples from y_t as a function of t. At iteration 1, PSMD has a uniform (no information) prior y_0 on ALP constraint violations. The subfigures in Figure 2 indicate rapid learning during early iterations, that is, the PSMD updates give rise to violation distributions y_t with density concentrated in smaller regions of the state-action space. Moreover, significant PSMD bound improvements in Figure 1(b) occur during iterations where y_t in Figure 2 changes substantially. This suggests that quick learning of the constraint violation distribution is important to obtain good quality bounds. However, the highly localized nature of y_{1000} indicates that finding a pre-specified sampling density to capture regions of high ALP constraint violation, as is often done in ALP-CS, may be difficult. Thus, a procedure that dynamically learns constraint violations, such as PSMD, is appropriate.

7.3. Perishable inventory control with partial backlogging and lead time

We extend the partially backlogged system described in §7.2 by considering a product with finite lifetime and a positive replenishment lead time (Karaesmen et al. 2011, Chen et al. 2014a, Wang 2014, Sun et al. 2016). This setting adds the feature of partial backlogging to one of the applications studied in Wang (2014) and Sun et al. (2016).

MDP formulation and instances. We assume that the order lead time is J periods and the item life time is I periods from receipt. We denote by q_j, $1 \le j \le J-1$, the orders that will be received j periods from now, and by z_i, $0 \le i \le I-1$, the on-hand inventory with i periods of lifetime remaining. The MDP state is the vector

$$s = (z_0, z_1, \ldots, z_{I-1}, q_1, q_2, \ldots, q_{J-1}) \in \mathbb{R}^{I+J-1}.$$

All on-hand and pipeline inventories are non-negative except for z_0. If $z_0 \ge 0$, then there are no backlogged orders and any remaining expired units will be disposed at the end of the current period at a cost of c_d per unit. If $z_0 < 0$, the amount of backlogged orders is $(G - \sum_{i=0}^{I-1} z_i)_+$, where G represents stochastic demand with distribution P_G. We limit the amount of backlogging by bounding z_0 below by $l_s < 0$. The order quantity a is bounded above by $\bar{a}$ and thus belongs to the interval $[0, \bar{a}]$. This bound on the order quantity implies that $z_i \in [0, \bar{a}]$ for $i = 1, \ldots, I-1$ and $q_j \in [0, \bar{a}]$ for $j = 1, \ldots, J-1$. Assuming demand is realized before an order arrival and is satisfied using a First In First Out (FIFO) rule, the MDP state transitions to

$$s' = (\max\{z_1 - (G - z_0)_+, l_s - \sum_{i=2}^{I-1} z_i\}, z_2, \ldots, z_{I-1}, q_1, q_2, \ldots, q_{J-1}, a).$$

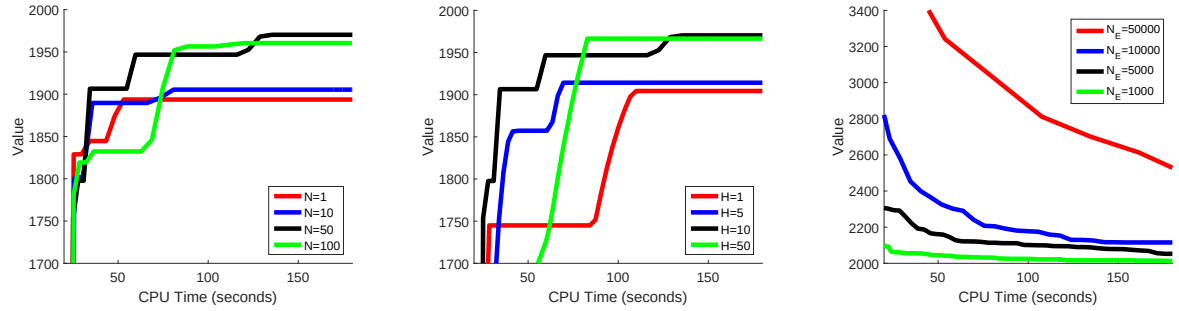
The maximum appearing in the first element of s' has two components. The first component is equal to the inventory with one period of lifetime remaining minus the demand unsatisfied by the expiring inventory. The second component imposes a lower bound l_s on the maximum backlog because $z_1 - (G - z_0)_+ + \sum_{i=2}^{I-1} z_i \ge l_s$ if and only if $z_1 - (G - z_0)_+ \ge l_s - \sum_{i=2}^{I-1} z_i$. Elements 2 through $I+J-2$ in s' correspond to elements 3 through $I+J-1$ in s, and element $I+J-1$ is the order quantity a placed at state s. Using the same notation for unit cost terms as §7.2, we define the MDP cost function

$$c(s,a) = \gamma^J c_p a + \mathbb{E}\left[c_h \left(\sum_{i=1}^{I-1} z_i - (G - z_0)_+ \right)_+ + c_b \left(G - \sum_{i=0}^{I-1} z_i \right)_+ + c_d (z_0 - G)_+ + c_l \left(l_s + G - \sum_{i=0}^{I-1} z_i \right)_+ \right],$$

where the first term represents the purchasing cost paid when receiving the order, the second is the holding cost, the third models the backlog cost, the fourth accounts for the disposal cost, and the fifth is the cost of lost sales from exceeding the backlogging limit of l_s.

To obtain instances for our computational experiments we considered $I = 2$ and $J = 2$. We chose P_G to be a truncated normal in the interval $[0, 10]$ with mean 5 and the standard deviation 2. We fixed c_p and c_l to 20 and 100, respectively. We varied the remaining parameters as indicated in Table 1 and set $\bar{a} = |l_s|$. For policy evaluation, we chose the initial state to be the singleton where each state element equals the mean of the demand distribution.

PSMD implementation. We use PSMD to solve an ALP formulation based on the basis functions 1, z_0, z_1, q_1, and $\{(z_0 - \nu)_+, (z_0 + z_1 - 2\nu)_+, (z_0 + z_1 + q_1 - 3\nu)_+, (2\nu - z_0 - z_1 - q_1)_+, (\nu -$



(a) PSMD lower bound trajectories for different N (H fixed at 10).

(b) PSMD lower bound trajectories for different H (N fixed at 50).

(c) ALP-CS objective function value trajectories for different N_E.

Figure 3 PSMD lower bound and ALP-CS objective function values on a representative perishable inventory control instance.

Table 1 PIB and ALP-CS objective function values as percentages of the PSMD lower bound on the perishable inventory control instances.

Instance parameters						ALP-CS Objective		
c_h	c_d	c_b	γ	l_s	PIB	Minimum	Average	Maximum
2	5	10	0.95	-10	92.8	100.2	100.7	101.2
2	5	10	0.99	-10	84.0	100.9	101.1	101.5
2	5	10	0.95	-50	93.0	103.6	107.2	111.4
2	5	10	0.99	-50	87.4	105.6	109.1	111.3
5	10	8	0.95	-10	90.4	100.4	101.2	101.8
5	10	8	0.99	-10	84.4	102.1	105.0	109.7
5	10	8	0.95	-50	94.7	103.7	109.6	116.7
5	10	8	0.99	-50	88.5	102.4	106.7	111.4
2	10	10	0.95	-10	96.4	100.4	100.8	101.5
2	10	10	0.99	-10	87.8	102.4	102.9	104.1
2	10	10	0.95	-30	90.1	99.0	100.1	101.4
2	10	10	0.99	-30	90.2	100.9	107.7	104.2
			Average		89.9	101.8	104.3	106.3

$z_1 - q_1)_+ | \nu \in \{\mathbb{E}[G], G^{0.25}, G^{0.5}\}\}$, where $G^{0.25}$ and $G^{0.5}$ are the 25-th and 50-th quartiles of the demand distribution. Our use of hinge-type basis functions is motivated by the structure of the cost function. We selected the PSMD initial distribution q to be consistent with the initial state used for policy evaluation. The effect of changing N and H in PSMD is illustrated in Figures 3(a) and 3(b), respectively, using a representative instance corresponding to row 3 of Table 1. The number of iterations and computation time needed to achieve a good lower bound does indeed depend on these parameters. Choosing H and N equal to 10 and 50, respectively, gives a good balance between run time and lower bound quality. In addition, the PSMD lower bound stabilizes within 3 minutes. Since this property was true across all instances, we set the PSMD run time limit to 3 minutes.

Benchmarks. We next discuss three heuristic policies and a lower bound to assess the performance of PSMD. The benchmark heuristic policies rely on a tractable simplification of the

problem, a look-ahead heuristic, and the ALP-CS VFA, respectively. The lower bound benchmark is based on perfect information (Brown et al. 2010).

Our first benchmark policy assumes non-perishability (i.e., $I = \infty$) and allows unrestricted order quantities and full backlogging. Under the first assumption, the MDP state simplifies to $s = (z, q_1, q_2, \ldots, q_{J-1})$. We model the second and third assumptions by relaxing the constraints $a \le \bar{a}$ and $z \ge l_s$, respectively, as opposed to choosing $\bar{a} = -l_s = \infty$. In other words, the parameters $\bar{a}$ and l_s are finite and consistent with their values in the original problem. When taking action a, the MDP state transitions to $s' = (z - G + q_1, q_2, \ldots, q_{J-1}, a)$ and the corresponding cost is

$$c_{\mathrm{FB}}(s,a) = \gamma^J c_p a + \mathbb{E}\big[c_h\,(z-G)_+ + c_b \min\{(G-z)_+, -l_s\} + c_l\,(G-z+l_s)_+\big],$$

where we ensure that the backlogging cost in $c_{\mathrm{FB}}(s,a)$ is consistent with $c(s,a)$ in the original problem, that is, the unit backlogging cost is c_b for each backlogged unit up to $|l_s|$ and is c_l for each unit of backlog greater than $|l_s|$. The optimal ordering policy is characterized in Proposition 4.

Proposition 4. *Assuming non-perishability, unrestricted order quantities, and full backlogging, the optimal order quantity at state s is*

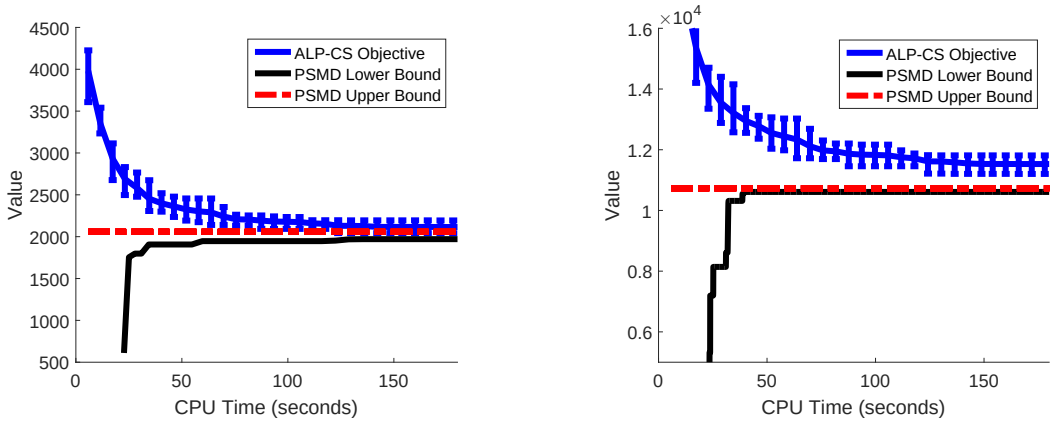
$$a_{\mathrm{FB}}(s) := \arg\min_{a\ge 0} \Big\{(1-\gamma)\gamma^J c_p a + \gamma^J \mathbb{E}\Big[c_h\,(Z+a-G(J))_+ + c_b \min\{(G(J)-Z-a)_+, -l_s\} + c_l\,(G(J)-Z-a+l_s)_+\Big]\Big\},$$

where $Z = z + \sum_{j=1}^{J-1} q_j$ is the cumulative on-hand inventory, $G(J) = \sum_{j=0}^{J} G_j$ is the cumulative demand over $J+1$ periods, and G_j is a realization of random demand G at a stage that occurs j periods in the future.

Note that $a_{FB}(s)$ in Proposition 4 is computed assuming unconstrained order quantities and may thus be infeasible to our original problem, where $a \le \bar{a}$. Therefore, we use the feasible order quantity $\min\{a_{FB}(s), \bar{a}\}$ instead to define a benchmark policy, which we label as FB.

The second benchmark relies on a look-ahead model that solves a one-period version of the MDP formulation, which terminates when the order placed at the first period arrives. Such look-ahead heuristics (LAH) perform well in inventory control applications considered by Zipkin (2008a) and Sun et al. (2016). Brown and Smith (2014) show that the performance of these policies can be improved by tuning the terminal stage cost. We thus implemented the version of LAH described in §5.3 of Brown and Smith (2014), that is, we added a linear terminal cost $c_s(\sum_{i=0}^{I-1} z_i + \sum_{j=1}^{J-1} q_j)$ where c_s is an unknown constant. We searched over values of c_s belonging to set $\{-5, -3, -1, 0, 1, 3, 5\}$ to obtain the LAH policy with the smallest upper bound estimate.

We also employed constraint sampling (ALP-CS) to tackle the same ALP solved by PSMD and used the resulting VFA weights in (5) to obtain a benchmark ordering policy. We considered different sampled versions of the ALP by varying (i) the initial distribution q and (ii) the procedure used



(a) Instance defined in row three of Table 1.

(b) Instance defined in row four of Table 1.

Figure 4 Comparison of the PSMD upper bound against the PSMD lower bound and the ALP-CS objective function value on two representative perishable inventory control instances.

to sample the state-action pairs needed to define the ALP constraints. For the initial state distribution q, we employed a degenerate distribution with density only at the initial state used for policy evaluation, a uniform distribution, and the state-visit distribution under the LAH policy already discussed. For defining the ALP constraints, we considered uniform sampling and simulated the LAH policy to generate the state-action pairs. Combining our choices for q and constraint-sampling distributions resulted in six different sampled ALPs. In each ALP-CS implementation, we also had to replace the expectations appearing in the ALP constraints by sample average approximations with N_E samples. We found that using an initial distribution that is consistent with the initial state and employing uniform constraint sampling gave the best ALP-CS policies on our instances. We thus use this combination in our ALP-CS implementation. Figure 3(c) displays the ALP-CS optimal objective as a function of time for different N_E values. Here CPU time is a surrogate for the number of ALP constraints because new sampled ALP constraints are added as time progresses. Larger values of N_E potentially lead to better lower bounds at the cost of extra computational time. For N_E equal to 50,000 the ALP-CS optimal objective value does not stabilize within three minutes and we verified that its stable value is not significantly different from the one for N_E equal to 10,000. We thus chose N_E as 10,000 for our ALP-CS implementation. Moreover, for each ALP-CS model we sampled constraints for three minutes to facilitate a fair comparison with PSMD.

We simulated all policies starting from the same initial state to estimate their values. Computing the actions prescribed by the PSMD policy and the benchmark policies required optimizing over order quantities. We solved this one-dimensional optimization problem by discretizing the action interval $[0, \bar{a}] \equiv [0, |l_s|]$ into 100 equally spaced points for all policies.

Finally, we considered a perfect information (lower) bound (PIB) to benchmark the PSMD lower bound (Brown et al. 2010). PIB is estimated by solving deterministic versions of our MDP

along sample paths of uncertainty generated using Monte Carlo simulation and averaging the resulting discounted total costs across these sample paths. We chose the number of stages on each sample path to be large enough (≈ 1000) so that the discounted cost is not affected by this horizon truncation. This led to accurate bound estimates with small standard errors. Please see Brown and Haugh (2017) for a detailed discussion and a potentially faster implementation for computing PIB. We also tracked the ALP-CS optimal objective function value to check if it provides a lower bound in our computations, even though this is not theoretically guaranteed.

Results. Table 1 reports the PIB and the ALP-CS optimal objective function values as percentages of the PSMD lower bound on twelve instances. We constructed five ALP-CS models using different random seeds to understand the impact of constraint sampling variability on its objective function value and report the minimum, average, and maximum of this value. The PSMD lower bound is 10% stronger on average than PIB. The ALP-CS optimal objective function value varies significantly across trials and can be 11% larger than the PSMD lower bound. Figure 4 compares the PSMD lower bound and the average ALP optimal objective function value against the PSMD policy value on the instances corresponding to rows three and four in Table 1 (the variability in the ALP-CS objective is represented by error bars). On the former instance (left subfigure), the mean ALP-CS optimal objective function value converges to the policy upper bound (red line) although this value is above the upper bound on some trials. On the latter instance (right subfigure), the ALP-CS objective function values are above the policy value by a significant amount in all independent trials. These findings indicate that the ALP-CS optimal objective function value may not provide a valid lower bound. In contrast, the PSMD lower bound is indeed valid, as expected from the theory in §5.2, and quickly stabilizes on both instances.

Table 2 reports the optimality gaps of the LAH, FB, ALP-CS, and PSMD policies with respect to the PSMD lower bound on the same instances of Table 1. We define the optimality gap of a policy as 100(1 - [PSMD lower bound/policy value]). The LAH optimality gaps range between 2.40% and 20.58% and average to 11.40%. The analogous range and average for FB are 4.89%−16.87% and 10.64%, respectively. Our results show that the FB policies outperform the LAH policies on average. ALP-CS policies exhibit an average optimality gap of 7.10% with this gap spread between 4.16% and 14.16% across instances. Policies from LAH and FB are clearly dominated by the ones from ALP-CS. On the other hand, the PSMD optimality gaps vary between 2.17% and 11.73%, and their average is 5.93%. Specifically, PSMD delivers policies that improve on the ones from ALP-CS on most instances. In addition, the small average PSMD optimality gap shows that its lower bound is also of high quality. The average CPU time to estimate the LAH and FB policies values were five minutes and three minutes, respectively. The analogous time for both the ALP-CS and PSMD policies was four minutes on average.

Table 2 **Optimality gaps of the LAH, FB, ALP-CS, and PSMD policies on the perishable inventory control instances.**

Instance parameters					Optimality gaps			
c_h	c_d	c_b	γ	l_s	LAH	FB	ALP-CS	PSMD
2	5	10	0.95	-10	6.12	8.92	4.52	4.01
2	5	10	0.99	-10	15.68	11.14	5.04	6.44
2	5	10	0.95	-50	13.1	9.75	4.50	4.40
2	5	10	0.99	-50	19.75	13.91	4.74	2.17
5	10	8	0.95	-10	7.18	7.83	8.56	7.96
5	10	8	0.99	-10	2.40	4.89	4.16	2.17
5	10	8	0.95	-50	12.82	10.18	10.05	11.73
5	10	8	0.99	-50	20.58	15.19	14.16	11.01
2	10	10	0.95	-10	8.31	12.21	6.79	2.40
2	10	10	0.99	-10	14.28	16.87	5.19	3.60
2	10	10	0.95	-30	9.05	9.41	8.71	6.14
2	10	10	0.99	-30	7.51	7.36	8.75	9.14
				Average	11.40	10.64	7.10	5.93

7.4. Consumer electricity storage management

We model the management of electricity storage by a consumer facing stochastic electricity demand and price. This application relates to recent research on the long-term operations of storage in the context of smart grids and power systems. For examples, see Grillo et al. (2012), van de Ven et al. (2013), and Erseghe et al. (2014).

MDP formulation and instances. Let $x \in [0, W]$ represent the amount of electricity stored in a battery with capacity W. The electricity demand and price are denoted by g and e, respectively. The pair (g, e) belongs to the box $[l_g, u_g] \times [l_e, u_e]$, where l_g, u_g, l_e, and u_e are non-negative and satisfy $u_g > l_g$ and $u_e > l_e$. The MDP state is the triple (x, g, e). At each stage, the consumer can charge or discharge the battery. If the decision is to charge, then the demand at the current stage plus the charge amount must be purchased from the grid. If the decision is instead to discharge, then any residual demand after using the discharged electricity is purchased from the grid and excess power is sold into the grid. Our MDP action, denoted by a, represents a charge if $a \geq 0$ and a discharge otherwise. A feasible charge/discharge decision satisfies the battery capacity, that is, $a \in \{a' \big| 0 \leq x + a' \leq W\}$, and changes the electricity in the battery to $x^+ = x + a$. The demand-price pair (g, e) transitions in a random manner to (g', e') following a truncated bivariate normal distribution with support on the box $[l_g, u_g] \times [l_e, u_e]$, mean (μ_g, μ_e), standard deviation (σ_g, σ_e), and correlation $\rho_{g,e}$. Denoting by $\psi_c > 1$ and $\psi_d \in (0, 1)$ the charging and discharging losses, respectively, the MDP cost function is

$$c(e, g, a) := (g + [\psi_c(a)_+ - \psi_d(a)_-])e.$$

We create instances of the above MDP by fixing $\gamma = 0.99$, $l_g = l_e = 0$, $u_g = u_e = 5$, $\mu_g = \mu_e = 2.5$, and $\rho_{g,e} = 0$. We assume $\sigma = \sigma_g = \sigma_e$ and vary σ, W, ψ_c, and ψ_d as shown in Table 3. We chose an initial state that is uniformly distributed in the interval $[0, 5]$ for policy evaluation.

Table 3 PIB and ALP-CS objective function values as percentages of the PSMD lower bound on the storage instances.

Instance parameters					ALP-CS Objective		
ψ_c	ψ_d	W	σ	PIB	Minimum	Average	Maximum
1.1	0.9	1	1	97.96	101.5	103.4	104.2
1.1	0.9	1	2	97.60	101.7	102.6	103.0
1.1	0.9	2	1	98.31	102.9	103.6	104.3
1.1	0.9	2	2	91.11	101.5	103.7	105.4
1.25	0.75	1	1	93.04	103.5	104.4	105.0
1.25	0.75	1	2	97.42	100.1	102.1	104.5
1.25	0.75	2	1	93.95	100.3	100.5	100.9
1.25	0.75	2	2	96.07	102.2	102.8	103.8
		Average		95.58	101.7	102.8	103.9

PSMD and benchmarks. PSMD solves an ALP formulated using the following polynomial basis functions of the state: 1, x, e, g, x^2, e^2, g^2, xe, xg, and ge. The initial distribution q in PSMD and the initial state for policy evaluation were both chosen to be uniform over storage inventory levels. We fixed both H and N equal to 10 as this provided a good trade-off between CPU time and lower bound quality. Please see the Electronic Companion EC.3. We terminated PSMD by imposing a run time limit of one minute as its lower bound stabilized before this time threshold on all the storage instances.

As policy benchmarks, we considered an LAH policy analogous to the one used in §7.3 and the ALP-CS policy. When implementing ALP-CS, we experimented with the same choices for the initial and constraint sampling distributions as outlined in §7.3. Combining a uniform initial distribution with either uniform- or LAH policy-based constraint sampling led to results that were within 0.5% of each other. We thus chose a uniform distribution for constraint sampling as well. We set the number of samples N_E to approximate expectations in ALP constraints equal to 10,000 as it leads to roughly the same stabilized objective function value as a higher value such as 50,000. We allowed one minute for constraint sampling in ALP-CS to be consistent with our PSMD run time limit. To benchmark the PSMD lower bound, we computed a perfect information bound (PIB) and also tracked the ALP-CS objective function value. Further details on parameter choices can be found in the Electronic Companion EC.3.

Results. Table 3 reports the PIB and ALP-CS objective function values as percentages of the PSMD lower bound on eight instances. We solved five ALP-CS models that were created using different random seeds for constraint sampling. The PSMD lower bound improves on PIB by

Table 4 Optimality gaps of the LAH, ALP-CS, and PSMD policies on the storage instances.

Instance parameters				Optimality gaps		
ψ_c	ψ_d	W	σ	LAH	ALP-CS	PSMD
1.1	0.9	1	1	9.07	7.49	7.04
1.1	0.9	1	2	10.86	7.08	5.96
1.1	0.9	2	1	9.89	9.67	9.93
1.1	0.9	2	2	12.59	10.24	11.76
1.25	0.75	1	1	10.61	4.05	5.18
1.25	0.75	1	2	9.12	3.96	3.02
1.25	0.75	2	1	4.73	5.59	5.44
1.25	0.75	2	2	9.63	5.72	5.23
		Average		9.56	6.72	6.69

between 2% and 9%. The ALP-CS optimal objective function value exhibits significant variability across different random trials, which can be as much as 5% on some instances. In particular, we verified that the ALP-CS objective function value after one minute can be larger than an upper bound on the optimal policy value, that is, ALP-CS may not provide a lower bound within a given a run time limit (please see the Electronic Companion EC.3 for details).

Table 4 contains the optimality gaps of the LAH, ALP-CS, and PSMD policies with respect to the PSMD lower bound. The PSMD optimality gaps range between 3.02% and 11.76% and are 6.69% on average. Analogous ranges and averages for ALP-CS and LAH are respectively 3.96%−10.24% and 6.72% and 4.73%−12.59% and 9.56%. We find that the PSMD and ALP-CS policies improve on the LAH policies on almost all the instances and their average policy values are roughly 3% larger. The average performance of PSMD and ALP-CS policies is the same although the performance of each policy on individual instances differ.

Overall, PSMD provides stronger lower bounds than PIB, while ALP-CS may not provide a valid lower bound. These findings are consistent with those in §7.3. PSMD policies improve on LAH policies and are comparable to the ALP-CS policies. The latter finding deviates partially from §7.3 where the PSMD policies also improved on the ALP-CS policies in addition to other benchmarks.

8. Conclusions

The linear programming approach to ADP has been successfully used to tackle high dimensional MDPs arising in several business applications and offers theoretical guarantees. This approach solves an ALP to compute a VFA that can be used to define an operating policy. ALP VFAs also provide a lower bound on the optimal policy cost that is useful for computing the optimality gaps of heuristic policies. Solving ALPs near optimally remains challenging in applications where the cost functions and/or transition dynamics are nonlinear or rich basis functions are needed to obtain good VFAs. We address this issue by introducing a saddle-point based reformulation of an ALP

that endogenizes a state-action density function as a dual decision variable. We develop a mirror-descent solution approach, PSMD, to solve this reformulation based on a functional stochastic gradient. PSMD computes a VFA that is a near optimal solution to ALP, provides a lower bound on the optimal policy cost, and guarantees such a solution and bound in a finite number of iterations with high probability. In theory, our approach offers advantages over row generation and constraint sampling strategies commonly employed for solving ALP. It also exhibits promising numerical performance on applications related to inventory control and energy storage. Specifically, PSMD delivers reliable and high quality lower bounds while the constraint sampling objective function value may provide invalid lower bounds. Moreover, PSMD policies outperform problem-specific benchmark policies and match or improve on the constraint sampling based ALP policies. Our ALP reformulation and PSMD thus extend the class of ALPs that can be solved near optimally.

Endnotes

[1]We caution that the use of sample average approximations may lead to biased gradients if a primal, as opposed to a primal-dual, first order approach is used. A primal approach views (7) as $\max_{\theta\in\Theta} \overline{f}(\theta)$, where $\overline{f}(\theta) := \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a)$, and requires the gradient of $\overline{f}(\theta)$. Since the expectation in $\overline{f}(\theta)$ lies inside a minimization, a sample average approximation of this expectation could lead to a biased estimate of $\overline{f}(\theta)$ and its gradient.

[2]The equivalence follows from $\arg\min_{\theta\in\Theta} \frac{1}{2}\left\|\theta - (\theta_t + \eta_t \mathbb{E}_{y_t}\left[\nabla_\theta f(\theta_t,s,a)\right])\right\|_2^2 = \arg\min_{\theta\in\Theta} \frac{\eta_t^2}{2}\left\|\mathbb{E}_{y_t}\left[\nabla_\theta f(\theta_t,s,a)\right]\right\|^2 + \eta_t\left\langle\theta_t - \theta, \mathbb{E}_{y_t}\left[\nabla_\theta f(\theta_t,s,a)\right]\right\rangle + \frac{1}{2}\|\theta - \theta_t\|_2^2 = \arg\min_{\theta\in\Theta}\left[\eta_t\left\langle\theta_t - \theta, \mathbb{E}_{y_t}\left[\nabla_\theta f(\theta_t,s,a)\right]\right\rangle + \frac{1}{2}\|\theta-\theta_t\|_2^2\right]$. The constant term $\frac{\eta_t^2}{2}\left\|\mathbb{E}_{y_t}\left[\nabla_\theta f(\theta_t,s,a)\right]\right\|^2$ can be dropped from the last equation since it does not effect the optimal solution.

[3]We are using the equality $\nabla_\theta \mathbb{E}_y\left[f(\theta,s,a)\right] = \mathbb{E}_y\left[\nabla_\theta f(\theta,s,a)\right]$, which holds by assumptions 1 and 2 and the compactness of $\mathcal{S}\times\mathcal{A}$.

[4]This closed-form expression can be derived because we use a KL divergence as a regularizer and a prox-function in (9) and (14), respectively. Obtaining a closed form is not viable under the Euclidean prox-function used, for example, in Nemirovski et al. (2009) and Chen et al. (2014b).

[5]Lemma EC.4 in the Electronic Companion EC.1 shows that $\bar{C}$ is non-positive. Since set $\mathcal{S}\times\mathcal{A}$ is full-dimensional, such a ball exists. The Lipschitz continuity of function $f(\cdot,\cdot,\cdot)$ follows from assumptions 1 and 2, and the compactness of the set Θ.

References

Adelman, D. 2003. Price-directed replenishment of subsets: Methodology and its application to inventory routing. *Manufacturing & Service Operations Management* **5**(4) 348–371.

Adelman, D. 2004. A price-directed approach to stochastic inventory/routing. *Operations Research* **52**(4) 499–514.

Adelman, D. 2007. Dynamic bid prices in revenue management. *Operations Research* **55**(4) 647–661.

Adelman, D., C. Barz. 2013. A unifying approximate dynamic programming model for the economic lot scheduling problem. *Mathematics of Operations Research* **39**(2) 374–402.

Adelman, D., D. Klabjan. 2012. Computing near-optimal policies in generalized joint replenishment. *INFORMS Journal on Computing* **24**(1) 148–164.

Adelman, D., A. Mersereau. 2013. Dynamic capacity allocation to customers who remember past service. *Management Science* **59**(3) 592–612.

Bach, E., J. O. Shallit. 1996. *Algorithmic Number Theory: Efficient Algorithms*, vol. 1. MIT press, Boston, MA.

Benjaafar, S., M. ElHafsi, T. Huang. 2010. Optimal control of a production-inventory system with both backorders and lost sales. *Naval Research Logistics* **57**(3) 252–265.

Bertsekas, P. B. 2007. *Dynamic Programming and Optimal Control*, vol. 2. 3rd ed. Athena Scientific, Belmont, MA, USA.

Bhat, B., V. F. Farias, C. C. Moallemi. 2012. Non-parametric approximate dynamic programming via the kernel method. Working paper, Columbia Univ.

Birge, J.R, F. Louveaux. 2011. *Introduction to Stochastic Programming*. Springer Science & Business Media, New York, NY, USA.

Bishop, C. M. 2006. *Pattern Recognition and Machine Learning*. Springer, New York, NY, USA.

Brown, D. B, M. B Haugh. 2017. Information relaxation bounds for infinite horizon markov decision processes. *Operations Research* **65**(5) 1355–1379.

Brown, D. B., J. E. Smith. 2014. Information relaxations, duality, and convex stochastic dynamic programs. *Operations Research* **62**(6) 1394–1415.

Brown, D. B., J. E. Smith, P. Sun. 2010. Information relaxations and duality in stochastic dynamic programs. *Operations Research* **58**(4) 785–801.

Bubeck, S. 2015. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning* **8**(3-4) 231–357.

Chen, X., Z. Pang, L. Pan. 2014a. Coordinating inventory control and pricing strategies for perishable products. *Operations Research* **62**(2) 284–300.

Chen, Y., G. Lan, Y. Ouyang. 2014b. Optimal primal-dual methods for a class of saddle point problems. *SIAM Journal on Optimization* **24**(4) 1779–1814.

de Farias, D. P., B. Van Roy. 2003. The linear programming approach to approximate dynamic programming. *Operations Research* **51**(6) 850–865.

de Farias, D. P., B. Van Roy. 2004. On constraint sampling for the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research* **29**(3) 462–478.

Desai, V. V., V. F. Farias, C. C Moallemi. 2012. Pathwise optimization for optimal stopping problems. *Management Science* **58**(12) 2292–2308.

Duchi, J. 2016. Introductory lectures on stochastic optimization. Lecture notes, Stanford Univ.

Erseghe, T., A. Zanella, C. G. Codemo. 2014. Optimal and compact control policies for energy storage units with single and multiple batteries. *IEEE Transactions on Smart Grid* **5**(3) 1308–1317.

Fan, K. 1953. Minimax theorems. *Proceedings of the National Academy of Sciences* **39**(1) 42–47.

Farias, V. F., B. Van Roy. 2006. Tetris: A study of randomized constraint sampling. *Probabilistic and Randomized Methods for Design Under Uncertainty (Springer-Verlag, London)* 189–201.

Farias, V. F., B. Van Roy. 2007. An approximate dynamic programming approach to network revenue management. Working paper, Stanford Univ.

Gallego, G. 2003. Production management. Lecture notes, Columbia Univ.

Grillo, S., M. Marinelli, S. Massucco, F. Silvestro. 2012. Optimal management strategy of a battery-based storage system to improve renewable energy integration in distribution networks. *IEEE Transactions on Smart Grid* **3**(2) 950–958.

Hazan, E., S. Kale. 2014. Beyond the regret minimization barrier: Optimal algorithms for stochastic strongly-convex optimization. *The Journal of Machine Learning Research* **15**(1) 2489–2512.

Hernández-Lerma, O., J. B. Lasserre. 1996. *Discrete-time Markov Control Processes: Basic Optimality Criteria*, vol. 30. Springer, New York, USA.

Juditsky, A., A. Nemirovski. 2011a. First-order methods for nonsmooth convex large-scale optimization, i: General purpose methods. S. Sra, S. Nowozin, SJ Wright, eds., *Optimization for Machine Learning*. MIT Press, Cambridge, MA, USA, 121–148.

Juditsky, A., A. Nemirovski. 2011b. First-order methods for nonsmooth convex large-scale optimization, ii: Utilizing problem's structure. S. Sra, S. Nowozin, SJ Wright, eds., *Optimization for Machine Learning*. MIT Press, Cambridge, MA, USA, 149–184.

Karaesmen, I. Z., A. Scheller-Wolf, B. Deniz. 2011. Managing perishable and aging inventories: Review and future research directions. *Planning Production and Inventories in the Extended Enterprise*. Springer, New York, NY, USA, 393–436.

Klabjan, D., D. Adelman. 2007. An infinite-dimensional linear programming algorithm for deterministic semi-Markov decision processes on Borel spaces. *Mathematics of Operations Research* **32**(3) 528–550.

Nadarajah, S., F. Margot, N. Secomandi. 2015. Relaxations of approximate linear programs for the real option management of commodity storage. *Management Science* **61**(12) 3054–3076.

Nadarajah, S., N. Secomandi. 2017. Relationship between least squares Monte Carlo and approximate linear programming. *Operations Research Letters* **45**(5) 409–414.

Nahmias, S., S. A. Smith. 1994. Optimizing inventory levels in a two-echelon retailer system with partial lost sales. *Management Science* **40**(5) 582–596.

Nemirovski, A., A. Juditsky, G. Lan, A. Shapiro. 2009. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* **19**(4) 1574–1609.

Patrick, J., M. Puterman, M. Queyranne. 2008. Dynamic multipriority patient scheduling for a diagnostic resource. *Operations Research* **56**(6) 1507–1525.

Petrik, M., G. Taylor, R. Parr, S. Zilberstein. 2010. Feature selection using regularization in approximate linear program for Markov decision processes. *Proceedings of the 27th International Conference on Machine Learning*. Haifa, Israel.

Powell, W. B. 2011. *Approximate Dynamic Programming: Solving the Curses of Dimensionality*. 2nd ed. John Wiley & Sons, Hoboken, NJ, USA.

Puterman, M. L. 1994. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA.

Rabinowitz, G., A. Mehrez, C. Chu, B. E. Patuwo. 1995. A partial backorder control for continuous review (r, q) inventory system with Poisson demand and constant lead time. *Computers & Operations Research* **22**(7) 689–700.

Restrepo, M. 2008. Computational methods for static allocation and real-time redeployment of ambulances. Ph.D. thesis, Cornell Univ.

Robert, C. P., G. Casella. 2004. *Monte Carlo Statistical Methods*. Springer, New York, NY, USA.

Schweitzer, P. J., A. Seidmann. 1985. Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications* **110**(2) 568–582.

Secomandi, N., D. J. Seppi. 2014. Real options and merchant operations of energy and other commodities. *Foundations and Trends in Technology, Information and Operations Management* **6**(3–4) 161–331.

Shapiro, A., D. Dentcheva, A. Ruszczyński. 2009. *Lectures on stochastic programming: modeling and theory*. SIAM, Philadelphia, PA.

Sun, P., K. Wang, P. Zipkin. 2016. Quadratic approximation of cost functions in lost sales and perishable inventory control problems. Working paper, Duke Univ.

TensorFlow. 2018. TensorFlow API documentation. Tech. rep., https://www.tensorflow.org/api_docs/python/tf/contrib/training/HParams. Accessed on February 10, 2018.

Tibshirani, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 267–288.

Topaloglu, H., S. Kunnumkal. 2006. Approximate dynamic programming methods for an inventory allocation problem under uncertainty. *Naval Research Logistics (NRL)* **53**(8) 822–841.

Trick, M. A., S. E. Zin. 1997. Spline approximations to value functions. *Macroeconomic Dynamics* **1**(1) 255–277.

van de Ven, P. M., N. Hegde, L. Massoulié, T. Salonidis. 2013. Optimal control of end-user energy storage. *IEEE Transactions on Smart Grid* **4**(2) 789–797.

Vossen, T. W., D. Zhang. 2015. Reductions of approximate linear programs for network revenue management. *Operations Research* **63**(6) 1352–1371.

Wang, K. 2014. Heuristics for inventory systems based on quadratic approximation of $l^{\#}$-convex value functions. Ph.D. thesis, Duke Univ.

Zhang, D., D. Adelman. 2009. An approximate dynamic programming approach to network revenue management with customer choice. *Transportation Science* **43**(3) 381–394.

Zipkin, P. 2008a. Old and new methods for lost-sales inventory systems. *Operations Research* **56**(5) 1256–1263.

Zipkin, P. 2008b. On the structure of lost-sales inventory models. *Operations Research* **56**(4) 937–944.

Electronic Companion

EC.1. Proofs

The proofs in this appendix use terms defined in the main paper such as the gamma function $\Gamma(\cdot)$ and the exponential function $\exp(\cdot)$; the parameters $\bar{C}$, R, L, n, and $Q_{\mathcal{S}\times\mathcal{A}}$ specified in §5.1; and $y^*_{\lambda,\theta} := \arg\min_{y\in\mathcal{Y}}[\mathbb{E}_y[f(\theta,s,a)] + \lambda D(y,p_u)]$ for given $\theta\in\Theta$ and $\lambda\in(0,1]$. These proofs also require new notation. To ease referencing, we define the most commonly used notation below and present the remaining terms as needed.

1. For any $y\in\mathcal{Y}$, $\theta^*_y := \arg\max_{\theta\in\Theta}[\mathbb{E}_y[f(\theta,s,a)]]$.
2. Given $\theta\in\Theta$, $(s^*_\theta,a^*_\theta)\in\arg\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a)$.
3. $Q_\Theta := \max_{\bar{\theta},\hat{\theta}\in\Theta}\|\bar{\theta}-\hat{\theta}\|^2$ is the maximum width of the compact set Θ.
4. Given $\lambda\in(0,1]$, $Q_D := -\bar{C} - n\log(\lambda)$.
5. Given $\theta\in\Theta$, M_θ is a scalar such that $M_\theta \geq \|\nabla_\theta \hat{f}(\theta,s,a)\|_2$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and all possible samples $\check{s}$ defining $\hat{f}(\theta,s,a)$. The scalar M_θ can be chosen as a finite value because the stochastic gradient $\nabla_\theta\hat{f}(\theta,s,a)$ in (13) is bounded in our setting. This is easy to verify by noting that the b-th element of $\nabla_\theta\hat{f}(\theta,s,a)$ is $\gamma\phi_b(\check{s}) - \phi_b(s) + \mathbb{E}_q[\phi_b(\cdot)]$, where $\phi_b(s)$ is a continuous function over a compact domain $\mathcal{S}$.
6. Given $y\in\mathcal{Y}$, M_y is a scalar such that $M_y \geq \|\hat{f}(\theta,s,a)\|_\infty := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}}|\hat{f}(\theta,s,a)|$ for all $\theta\in\Theta$ almost surely. The parameter M_y can be chosen as a bounded value because the gradient $\nabla_y\mathbb{E}_y\left[\hat{f}(\theta_t,s,a)\right]$ with respect to y (i.e. $\nabla_y\mathbb{E}_y\left[\hat{f}(\theta_t,s,a)\right] = \hat{f}(\theta_t,s,a)$) used in (15) is bounded in our setting. This follows because the function $\hat{f}(\theta_t,s,a)$ is defined by a sample $\check{s}$ that can be bounded on each sample path by a constant M_y, since each realization of the state-action pair (s,a) belongs to the compact domain $\mathcal{S}\times\mathcal{A}$, the continuous functions $c(s,a)$ and $\phi_b(s)$ are defined on this domain and θ also belongs to the compact set Θ.
7. $Q := Q_\Theta + 2Q_D$.
8. $M^2 = M_\theta^2 + M_y^2$.

Proof of Proposition 1 Since $\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a) \leq \inf_{y\in\mathcal{Y}}\mathbb{E}_y[f(\theta,s,a)]$ for any $\theta\in\Theta$, it holds that $F \leq \max_{\theta\in\Theta}\inf_{y\in\mathcal{Y}}\mathbb{E}_y[f(\theta,s,a)]$. To complete the proof, we establish the reverse of this inequality. Fix an arbitrary $\theta\in\Theta$. For this θ and any $r\geq 0$, we define a density function

$$y_{r,\theta}(s,a) := \frac{\mathbf{1}(\|(s,a)-(s^*_\theta,a^*_\theta)\|_2\leq r)}{\int_{\mathcal{S}\times\mathcal{A}}\mathbf{1}(\|(s,a)-(s^*_\theta,a^*_\theta)\|_2\leq r)d(s,a)}$$

over $\mathcal{S}\times\mathcal{A}$, where $\mathbf{1}$ is the 0-1 indicator function. Note that $y_{r,\theta}\in\mathcal{Y}$. Since $f(\theta,s,a)$ is L-Lipschitz continuous on $\mathcal{S}\times\mathcal{A}$ (by Assumption 1), we have $f(\theta,s,a)\leq f(\theta,s^*_\theta,a^*_\theta) + L\|(s,a)-(s^*_\theta,a^*_\theta)\|_2 \leq f(\theta,s^*_\theta,a^*_\theta)+rL$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}$ such that $\|(s,a)-(s^*_\theta,a^*_\theta)\|_2\leq r$. Therefore, $\inf_{y\in\mathcal{Y}}\mathbb{E}_y[f(\theta,s,a)] \leq$

$\mathbb{E}_{y_{r,\theta}}[f(\theta,s,a)] = \int_{\mathcal{S}\times\mathcal{A}} f(\theta,s,a) y_{r,\theta}(s,a) d(s,a) \le f(\theta,s^*_\theta,a^*_\theta) + rL$. Since r can be arbitrarily small, in the limit we have $\inf_{y\in\mathcal{Y}} \mathbb{E}_y[f(\theta,s,a)] \le f(\theta,s^*_\theta,a^*_\theta) = \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a)$. Since this inequality holds for any $\theta\in\Theta$, we have $\max_{\theta\in\Theta}\inf_{y\in\mathcal{Y}} \mathbb{E}_y[f(\theta,s,a)] \le F$. □

We present Lemmas EC.1-EC.7 below, which are needed in the proofs of Proposition 2 and Theorem 1.

LEMMA EC.1. *For any $\lambda\in(0,1]$ and $\theta\in\Theta$ we have*

$$\int_{\mathcal{S}\times\mathcal{A}} \exp\left(-\frac{1}{\lambda} f(\theta,s,a)\right) d(s,a) \ge \frac{\pi^{n/2}}{\Gamma(n/2+1)} (\lambda R)^n \exp(-L(R+Q_{\mathcal{S}\times\mathcal{A}})) \exp\left(-\frac{1}{\lambda}\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a)\right). \tag{EC.1}$$

Proof. Take any $\theta\in\Theta$ and $\lambda\in(0,1]$. By multiplying and dividing the left hand side of (EC.1) by $\exp\big(-1/\lambda \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a)\big)$, we get $\int_{\mathcal{S}\times\mathcal{A}} \exp\left(-\frac{1}{\lambda} f(\theta,s,a)\right) d(s,a)$

$$\begin{aligned}
&= \exp\left(-\frac{1}{\lambda}\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a)\right) \int_{\mathcal{S}\times\mathcal{A}} \exp\left(-\frac{1}{\lambda}\left[f(\theta,s,a) - \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a)\right]\right) d(s,a) \\
&\ge \exp\left(-\frac{1}{\lambda}\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a)\right) \int_{\mathcal{S}\times\mathcal{A}} \exp\left(-\frac{L}{\lambda}\|(s,a)-(s^*_\theta,a^*_\theta)\|_2\right) d(s,a). \qquad \text{(EC.2)}
\end{aligned}$$

The above inequality holds since $f(\cdot,s,a)$ is L-Lipschitz continuous in (s,a) by Assumption 1. We next show that

$$\int_{\mathcal{S}\times\mathcal{A}} \exp\left(-\frac{L}{\lambda}\|(s,a)-(s^*_\theta,a^*_\theta)\|_2\right) d(s,a) \ge \frac{\pi^{n/2}}{\Gamma(n/2+1)} (\lambda R)^n \exp(-L(R+Q_{\mathcal{S}\times\mathcal{A}})). \tag{EC.3}$$

Then using (EC.3) in (EC.2) completes the proof. Since $\mathcal{S}\times\mathcal{A}$ is full-dimensional, there exists $(\hat{s},\hat{a})\in\mathcal{S}\times\mathcal{A}$ such that $\mathcal{B}_R(\hat{s},\hat{a}) := \{(s,a) : \|(s,a)-(\hat{s},\hat{a})\|_2 \le R\} \subseteq \mathcal{S}\times\mathcal{A}$ where $\mathcal{B}_R(\hat{s},\hat{a})$ is a Euclidean ball with radius R centered at $(\hat{s},\hat{a})$. For any $(s,a)\in\mathcal{B}_R(\hat{s},\hat{a})$, we have

$$\|(s,a)-(s^*_\theta,s^*_\theta)\|_2 \le \|(s,a)-(\hat{s},\hat{a})\|_2 + \|(\hat{s},\hat{a})-(s^*_\theta,a^*_\theta)\|_2 \le R + \|(\hat{s},\hat{a})-(s^*_\theta,a^*_\theta)\|_2. \tag{EC.4}$$

Let $(s_\lambda,a_\lambda) := \lambda(\hat{s},\hat{a}) + (1-\lambda)(s^*_\theta,a^*_\theta)$. Then we have $\mathcal{B}_{\lambda R}(s_\lambda,a_\lambda) := \{(s,a) : \|(s,a)-(s_\lambda,a_\lambda)\|_2 \le \lambda R\} = \lambda\mathcal{B}_R(\hat{s},\hat{a}) + (1-\lambda)(s^*_\theta,a^*_\theta) \subseteq \mathcal{S}\times\mathcal{A}$, where the last inclusion holds since $\mathcal{B}_R(\hat{s},\hat{a}) \subseteq \mathcal{S}\times\mathcal{A}$, $(s^*_\theta,a^*_\theta)\in\mathcal{S}\times\mathcal{A}$ and the set $\mathcal{S}\times\mathcal{A}$ is convex. This indicates that for any $(s,a)\in\mathcal{B}_{\lambda R}(s_\lambda,a_\lambda)$, there exists a point $(s',a')\in\mathcal{B}_R(\hat{s},\hat{a})$ such that $(s,a) = \lambda(s',a') + (1-\lambda)(s^*_\theta,a^*_\theta)$. Therefore, for any $(s,a)\in\mathcal{B}_{\lambda R}(s_\lambda,a_\lambda)$, we have

$$\begin{aligned}
\|(s,a)-(s^*_\theta,a^*_\theta)\|_2 &= \|\lambda(s',a') + (1-\lambda)(s^*_\theta,a^*_\theta) - (s^*_\theta,a^*_\theta)\|_2 = \lambda\|(s',a')-(s^*_\theta,a^*_\theta)\|_2 \\
&\le \lambda\left(R + \|(\hat{s},\hat{a})-(s^*_\theta,a^*_\theta)\|_2\right). \qquad \text{(EC.5)}
\end{aligned}$$

The last inequality holds by (EC.4) since $(s',a')\in\mathcal{B}_R(\hat{s},\hat{a})$. Therefore, we have

$$\int_{\mathcal{S}\times\mathcal{A}} \exp\left(-\frac{L}{\lambda}\|(s,a)-(s^*_\theta,a^*_\theta)\|_2\right) d(s,a) \ge \int_{\mathcal{B}_{\lambda R}(s_\lambda,a_\lambda)} \exp\left(-\frac{L}{\lambda}\|(s,a)-(s^*_\theta,a^*_\theta)\|_2\right) d(s,a)$$

$$\begin{aligned}
&\geq \int_{\mathcal{B}_{\lambda R}(s_\lambda,a_\lambda)} \exp\left(-L\left(R+\|(\hat{s},\hat{a})-(s^*_\theta,a^*_\theta)\|_2\right)\right) d(s,a)\\
&= \exp\left(-L\left(R+\|(\hat{s},\hat{a})-(s^*_\theta,a^*_\theta)\|_2\right)\right) \int_{\mathcal{B}_{\lambda R}(s_\lambda,a_\lambda)} 1 d(s,a)\\
&\geq \frac{\pi^{n/2}}{\Gamma(n/2+1)}(\lambda R)^n \exp\left(-L(R+Q_{\mathcal{S}\times\mathcal{A}})\right).
\end{aligned}$$

The first inequality holds since $\mathcal{B}_{\lambda R}(s_\lambda,a_\lambda)\subseteq\mathcal{S}\times\mathcal{A}$, the second inequality follows by (EC.5), and the last inequality holds by the definition of $Q_{\mathcal{S}\times\mathcal{A}}$ and the fact that $\int_{\mathcal{B}_{\lambda R}(s_\lambda,a_\lambda)} 1 d(s,a) = \frac{\pi^{n/2}}{\Gamma(n/2+1)}(\lambda R)^n$ is the volume of the Euclidean ball with radius λR in an n-dimensional space. □

LEMMA EC.2. *For a given $\theta\in\Theta$ and $\lambda\in(0,1]$, a dual minimizer in (9) is*

$$y^*_{\lambda,\theta}(s,a) := \frac{\exp\left(-f(\theta,s,a)/\lambda\right)}{\int \exp\left(-f(\theta,s,a)/\lambda\right) d(s,a)},$$

and $y^*_{\lambda,\theta}\in\mathcal{Y}$.

Proof. The KKT conditions for the convex optimization problem (9) are

$$\begin{aligned}
&f(\theta,s,a)+\lambda\left(1+\log\frac{y^*_{\lambda,\theta}(s,a)}{\bar{p}}\right)+\mu(s,a)+\beta=0, \quad \forall(s,a)\in\mathcal{S}\times\mathcal{A}, \qquad \text{(EC.6)}\\
&\int_{\mathcal{S}\times\mathcal{A}} y^*_{\lambda,\theta}(s,a)d(s,a)=1,\ \beta\in\mathbb{R},\ \mu(s,a)\leq 0,\ \mu(s,a)y^*_{\lambda,\theta}(s,a)=0,\ y^*_{\lambda,\theta}(s,a)>0,\ \forall(s,a)\in\mathcal{S}\times\mathcal{A},
\end{aligned}$$

where $\mu(\cdot,\cdot)$ and β are Lagrange multipliers. Since $y(s,a)>0$ and $\mu(s,a)y(s,a)=0$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$, we conclude that $\mu(s,a)=0$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. Using this observation in (EC.6) and solving it for $y^*_{\lambda,\theta}$ gives

$$y^*_{\lambda,\theta}(s,a)=\bar{p}\exp\left(-\frac{f(\theta,s,a)}{\lambda}-\frac{\beta+\lambda}{\lambda}\right), \quad \forall(s,a)\in\mathcal{S}\times\mathcal{A}. \qquad \text{(EC.7)}$$

The equality $\int_{\mathcal{S}\times\mathcal{A}} y^*_{\lambda,\theta}(s,a)d(s,a)=1$ can be used to verify that $\bar{p}\exp\left(-\frac{\beta+\lambda}{\lambda}\right)=\left(\int_{\mathcal{S}\times\mathcal{A}}\exp\left(-\frac{f(\theta,s,a)}{\lambda}\right)d(s,a)\right)^{-1}$. Using this latter equality in (EC.7) shows that

$$y^*_{\lambda,\theta}(s,a)=\bar{p}\exp\left(-\frac{f(\theta,s,a)}{\lambda}\right)\exp\left(-\frac{\beta+\lambda}{\lambda}\right)=\frac{\exp\left(-\frac{f(\theta,s,a)}{\lambda}\right)}{\int\exp\left(-\frac{f(\theta,s,a)}{\lambda}\right)d(s,a)}.$$

Finally, $y^*_{\lambda,\theta}\in\mathcal{Y}$ because $y^*_{\lambda,\theta}>0$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$ and $\int_{\mathcal{S}\times\mathcal{A}} y^*_{\lambda,\theta}(s,a)d(s,a)=1$, which follow directly from the preceding expression for $y^*_{\lambda,\theta}$. □

LEMMA EC.3. *For any $\lambda\in(0,1]$ and $\theta\in\Theta$ we have*

$$\inf_{y\in\mathcal{Y}}\mathbb{E}_y[f(\theta,s,a)]\geq\min_{y\in\mathcal{Y}}\left[\mathbb{E}_y[f(\theta,s,a)]+\lambda D(y,p_u)\right]+\lambda\bar{C}+n\lambda\log(\lambda).$$

Proof. Using Lemma (EC.2), we know $y^*_{\lambda,\theta}$ solves (9). Evaluating the objection function of (9) at $y^*_{\lambda,\theta}$ and θ, we get $\mathbb{E}_{y^*_{\lambda,\theta}}[f(\theta,s,a)] + \lambda D(y^*_{\lambda,\theta}, p_u)$

$$
\begin{aligned}
&= \lambda \left(\int_{\mathcal{S}\times\mathcal{A}} \exp\left(-f(\theta,s,a)/\lambda\right) d(s,a) \right)^{-1} \left[\int_{\mathcal{S}\times\mathcal{A}} \exp\left(-f(\theta,s,a)/\lambda\right) \log\left(\frac{1}{\bar{p}}\right) d(s,a) \right. \\
&\quad \left. - \log\left(\int_{\mathcal{S}\times\mathcal{A}} \exp\left(-f(\theta,s,a)/\lambda\right) d(s,a) \right) \int_{\mathcal{S}\times\mathcal{A}} \exp\left(-f(\theta,s,a)/\lambda\right) d(s,a) \right] \\
&= \lambda \log\left(\frac{1}{\bar{p}}\right) - \lambda \log\left(\int_{\mathcal{S}\times\mathcal{A}} \exp\left(-f(\theta,s,a)/\lambda\right) d(s,a) \right),
\end{aligned}
$$

where we remind the reader that $\bar{p}$ is the constant value defining the uniform density p_u. Using (EC.1) to bound the second term in the right hand side of the last equality we get $\mathbb{E}_{y^*_{\lambda,\theta}}[f(\theta,s,a)] + \lambda D(y^*_{\lambda,\theta}, p_u)$

$$
\begin{aligned}
&\leq \lambda \log\left(\frac{1}{\bar{p}}\right) - \lambda \log\left(\frac{\pi^{n/2}}{\Gamma(n/2+1)} (\lambda R)^n \exp(-L(R + Q_{\mathcal{S}\times\mathcal{A}})) \exp\left(-\frac{1}{\lambda} \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a) \right) \right) \\
&= \lambda \log\left(\frac{1}{\bar{p}}\right) + \lambda \log\left(\frac{\Gamma(n/2+1)}{\pi^{n/2} R^n} \right) + \lambda L(R + Q_{\mathcal{S}\times\mathcal{A}}) - n\lambda \log(\lambda) + \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a) \\
&= -\lambda \bar{C} - n\lambda \log(\lambda) + \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a), \qquad \text{(EC.8)}
\end{aligned}
$$

where the second and third equalities follow from rearranging and simplifying terms, and recalling that $\bar{C} = \log(\bar{p}) - \log\left(\frac{\Gamma(n/2+1)}{\pi^{n/2}R^n}\right) - L(R + Q_{\mathcal{S}\times\mathcal{A}})$. Therefore, $\inf_{y\in\mathcal{Y}} \mathbb{E}_y[f(\theta,s,a)] - \min_{y\in\mathcal{Y}}[\mathbb{E}_y[f(\theta,s,a)] + \lambda D(y,p_u)] \geq \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a) + \lambda\bar{C} + n\lambda\log(\lambda) - \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} f(\theta,s,a) = \lambda\bar{C} + n\lambda\log(\lambda)$, where the inequality follows from (EC.8). □

LEMMA EC.4. *The parameter* $\bar{C} = \log(\bar{p}) - \log\left(\frac{\Gamma(n/2+1)}{\pi^{n/2}R^n}\right) - L(R + Q_{\mathcal{S}\times\mathcal{A}})$ *is non-positive.*

Proof. The third term $-L(R + Q_{\mathcal{S}\times\mathcal{A}})$ in the definition of $\bar{C}$ is non-positive since $L \geq 0$, $R > 0$, and $Q_{\mathcal{S}\times\mathcal{A}} \geq 0$. In addition, since $\mathcal{S}\times\mathcal{A}$ is full-dimensional, there exists $(\hat{s},\hat{a}) \in \mathcal{S}\times\mathcal{A}$ such that $\mathcal{B}_R(\hat{s},\hat{a}) \subseteq \mathcal{S}\times\mathcal{A}$, where $\mathcal{B}_R(\hat{s},\hat{a})$ denotes a Euclidean ball with radius R centered around $(\hat{s},\hat{a})$. Using $\int_{\mathcal{B}_R(\hat{s},\hat{a})} 1 d(s,a) \leq \int_{\mathcal{S}\times\mathcal{A}} 1 d(s,a)$ we conclude that $\bar{C}$ is non-positive since $\log(\bar{p}) - \log\left(\frac{\Gamma(n/2+1)}{\pi^{n/2}R^n}\right) = \log\left(\frac{\int_{\mathcal{B}_R(\hat{s},\hat{a})} 1 d(s,a)}{\int_{\mathcal{S}\times\mathcal{A}} 1 d(s,a)}\right) \leq 0$. Notice that, in the last equation we used the formula for the volume of a Euclidean ball with radius R in a n-dimensional space that is $\int_{\mathcal{B}_R(\hat{s},\hat{a})} 1 d(s,a) = \frac{\pi^{n/2}R^n}{\Gamma(n/2+1)}$. □

LEMMA EC.5. *Let* η_t, λ_t, *and* $\bar{\lambda}_t$ *be respectively the step length, the regularization parameter, and averaged regularization coefficient used in the* t*-th iteration of Algorithm 1. For a given* $\tilde{\lambda} \in (0,1]$, *we have that* $\bar{\lambda}_T \leq \tilde{\lambda}$ *after* $T \geq \max\left\{4, \left(\frac{8\lambda_0}{\tilde{\lambda}} \log\left(\frac{8\lambda_0}{\tilde{\lambda}}\right)\right)^2 - 2\right\}$ *iterations of Algorithm 1.*

Proof. Since $\eta_t = \dfrac{\eta_0}{\sqrt{t+1}}$ and $\lambda_t = \dfrac{\lambda_0}{\sqrt{t+1}}$, when $T \geq 4$, we have $\sum_{t=0}^{T} \eta_t \geq 2\eta_0(\sqrt{T+2}-1) \geq \eta_0\sqrt{T+2}$ and $\sum_{t=0}^{T} \eta_t\lambda_t = \sum_{t=0}^{T} \dfrac{\eta_0\lambda_0}{t+1} \leq \eta_0\lambda_0(1+\log(T+2)) \leq 2\eta_0\lambda_0\log(T+2)$. Therefore,

$$\bar{\lambda}_T = \frac{1}{\sum_{t=0}^{T}\eta_t}\sum_{t=0}^{T}\eta_t\lambda_t \leq \frac{2\lambda_0\log(T+2)}{\sqrt{T+2}}.$$

When $T+2 \geq \max\left\{6, \left(\dfrac{8\lambda_0}{\tilde{\lambda}}\log\left(\dfrac{8\lambda_0}{\tilde{\lambda}}\right)\right)^2\right\}$, we get

$$\bar{\lambda}_T \leq \frac{\tilde{\lambda}}{2}\log\left(\frac{8\lambda_0}{\tilde{\lambda}}\log\left(\frac{8\lambda_0}{\tilde{\lambda}}\right)\right)/\log\left(\frac{8\lambda_0}{\tilde{\lambda}}\right) \leq \tilde{\lambda},$$

where we used the fact that the function $\log(T+2)/\sqrt{T+2}$ is decreasing for $T \geq 4$ in the first inequality and $\log\left(\dfrac{8\lambda_0}{\tilde{\lambda}}\log\left(\dfrac{8\lambda_0}{\tilde{\lambda}}\right)\right)/\log\left(\dfrac{8\lambda_0}{\tilde{\lambda}}\right) \leq 2$ in the second inequality. □

LEMMA EC.6. *For any $\alpha > 0$ and*

$$\tilde{\lambda} := \begin{cases} \min\left\{\dfrac{\alpha}{32n\log(32n/\alpha)}, -\dfrac{\alpha}{16\bar{C}}\right\} & \text{if } \bar{C} > 0, \\ \dfrac{\alpha}{32n\log(32n/\alpha)} & \text{if } \bar{C} = 0, \end{cases} \tag{EC.9}$$

we have $\tilde{\lambda}\bar{C} + n\tilde{\lambda}\log(\tilde{\lambda}) \geq -\alpha/8$.

Proof. If $\bar{C} = 0$, then $\tilde{\lambda} = \dfrac{\alpha}{32n\log(32n/\alpha)}$ by definition and it follows that

$$\begin{aligned} \tilde{\lambda}\bar{C} + n\tilde{\lambda}\log(\tilde{\lambda}) &= n\tilde{\lambda}\log(\tilde{\lambda}) \\ &= -n\tilde{\lambda}\log\left(\tilde{\lambda}^{-1}\right) \\ &= -n\left(\frac{\alpha}{32n\log(32n/\alpha)}\log\left(\frac{32n\log(32n/\alpha)}{\alpha}\right)\right) \\ &\geq -\alpha/16 \geq -\alpha/8. \end{aligned}$$

The inequality holds since $\log\left(\frac{32n}{\alpha}\log(32n/\alpha)\right)/\log(32n/\alpha) \leq 2$.
If $\bar{C} < 0$, we have $\tilde{\lambda} = \min\left\{\dfrac{\alpha}{32n\log(32n/\alpha)}, -\dfrac{\alpha}{16\bar{C}}\right\}$. In this case,

$$\begin{aligned} \tilde{\lambda}\bar{C} + n\tilde{\lambda}\log(\tilde{\lambda}) &\geq -\frac{\alpha}{16} - n\tilde{\lambda}\log(\tilde{\lambda}^{-1}) \\ &\geq -\frac{\alpha}{16} - n\left(\frac{\alpha}{32n\log(32n/\alpha)}\log\left(\frac{32n\log(32n/\alpha)}{\alpha}\right)\right) \\ &\geq -\frac{\alpha}{16} - \frac{\alpha}{16} = -\frac{\alpha}{8}, \end{aligned}$$

where we use $\tilde{\lambda} \leq -\dfrac{\alpha}{16\bar{C}}$ to obtain the first inequality. The second inequality follows from $\tilde{\lambda} \leq \dfrac{\alpha}{32n\log(32n/\alpha)}$ and the function $\lambda\log(\lambda^{-1})$ increasing on the interval $[0, 0.3]$. The third inequality holds since

$$\log\left(\frac{32n}{\alpha}\log\left(\frac{32n}{\alpha}\right)\right)/\log\left(\frac{32n}{\alpha}\right) \leq 2.$$

□

LEMMA EC.7. *For any $\lambda \in (0,1]$ and $\theta \in \Theta$, we have*

$$D(y^*_{\lambda,\theta}, p_u) \le Q_D, \tag{EC.10}$$

where $Q_D := -\bar{C} - n\log(\lambda)$.

Proof. We proceed to show that $\lambda D(y^*_{\lambda,\theta}, p_u) \le \lambda Q_D = -\lambda\bar{C} - n\lambda\log(\lambda)$ for $\lambda > 0$, which proves the required inequality. Using the definition of $y^*_{\lambda,\theta}$ and Lemma EC.3 we get $\lambda D(y^*_{\lambda,\theta}, p_u) = \mathbb{E}_{y^*_{\lambda,\theta}}[f(\theta,s,a)] + \lambda D(y^*_{\lambda,\theta}, p_u) - \mathbb{E}_{y^*_{\lambda,\theta}}[f(\theta,s,a)] \le \min_{y\in\mathcal{Y}}[\mathbb{E}_y[f(\theta,s,a)] + \lambda D(y,p_u)] - \inf_{y\in\mathcal{Y}}\mathbb{E}_y[f(\theta,s,a)] \le -\lambda\bar{C} - n\lambda\log(\lambda)$. Notice that since $\bar{C} \le 0$ (by Lemma EC.4) and $\lambda \in (0,1]$, the upper bound on $D(y^*_{\lambda,\theta}, p_u)$ is non-negative. □

Proof of Proposition 2 The first part of this Proposition was already proved in Lemma EC.2. We begin here by showing that $F(\lambda) \ge F$ for any $\lambda \in (0,1]$. Take an arbitrary $\theta \in \Theta$. From the definition of $F(\lambda)$, we have

$$\begin{aligned} F(\lambda) &\ge \min_{y\in\mathcal{Y}}[\mathbb{E}_y[f(\theta,s,a)] + \lambda D(y,p_u)] = \mathbb{E}_{y^*_{\theta,\lambda}}[f(\theta,s,a)] + \lambda D(y^*_{\theta,\lambda}, p_u) \\ &\ge \mathbb{E}_{y^*_{\theta,\lambda}}[f(\theta,s,a)] \ge \inf_{y\in\mathcal{Y}}\mathbb{E}[f(\theta,s,a)], \end{aligned}$$

where the second inequality holds since $\lambda > 0$ and $D(y,p_u) \ge 0$ for all $y \in \mathcal{Y}$. Since our choice of θ was arbitrary, it follows that

$$F(\lambda) \ge \max_{\theta\in\Theta}\inf_{y\in\mathcal{Y}}\mathbb{E}[f(\theta,s,a)] = F.$$

Finally, we show that given $\alpha > 0$, if λ is sufficiently small, i.e. $\lambda = \tilde{\lambda}$ for $\tilde{\lambda}$ defined in (EC.9), we have $F(\lambda) \le F + \alpha$. Given $\lambda \in (0,1]$, we have $F = \max_{\theta\in\Theta}\inf_{y\in\mathcal{Y}}\mathbb{E}_y[f(\theta,s,a)] \ge \max_{\theta\in\Theta}\min_{y\in\mathcal{Y}}[\mathbb{E}_y[f(\theta,s,a)] + \lambda D(y,p_u)] + \lambda\bar{C} + n\lambda\log(\lambda) = F(\lambda) + \lambda\bar{C} + n\lambda\log(\lambda)$. The first inequality follows from Lemma EC.3. When $\lambda = \tilde{\lambda}$ (defined in (EC.9)), we have $F \ge F(\tilde{\lambda}) + \tilde{\lambda}\bar{C} + n\tilde{\lambda}\log(\tilde{\lambda}) \ge F(\tilde{\lambda}) - \alpha/8 \ge F(\tilde{\lambda}) - \alpha$, where the second inequality holds by virtue of Lemma EC.6. □

Proof of Theorem 1 We begin by bounding the duality gap $\mathcal{P}(\bar{y}) - \mathcal{D}(\bar{\theta})$ and use this bound to derive the iteration complexity.

Bound on $-\mathbb{E}_{y-y_t}[f(\theta_t,s,a)] - \lambda_t D(y,p_u)$ for any $y \in \mathcal{Y}$: Fix $t \ge 0$. When performing the y-update, $y_{t+1} \in \mathcal{Y}$ solves the convex optimization problem (14). The optimality of y_{t+1} implies

$$\left\langle \eta_t(\hat{f}(\theta_t,\cdot,\cdot) + \lambda_t(1 + \log(y_{t+1}/p_u))) + 1 + \log(y_{t+1}/y_t), y_{t+1} - y \right\rangle \le 0.$$

This indicates that $\mathbb{E}_{y_{t+1}-y}\left[\eta_t\hat{f}(\theta_t,s,a) + \eta_t\lambda_t\log(\frac{y_{t+1}(s,a)}{p_u(s,a)}) + \log(\frac{y_{t+1}(s,a)}{y_t(s,a)})\right] \le 0$. By rearranging the terms in this inequality we get $-\mathbb{E}_y\left[\log(\frac{y_{t+1}(s,a)}{y_t(s,a)})\right]$

$$\begin{aligned} &\le -\eta_t\mathbb{E}_{y_{t+1}-y}\left[\hat{f}(\theta_t,s,a)\right] - \eta_t\lambda_t\mathbb{E}_{y_{t+1}-y}\left[\log\left(\frac{y_{t+1}(s,a)}{p_u(s,a)}\right)\right] - \mathbb{E}_{y_{t+1}}\left[\log\left(\frac{y_{t+1}(s,a)}{y_t(s,a)}\right)\right] \\ &= -\eta_t\mathbb{E}_{y_{t+1}-y}\left[\hat{f}(\theta_t,s,a)\right] - \eta_t\lambda_t D(y_{t+1},p_u) + \eta_t\lambda_t D(y,p_u) - \eta_t\lambda_t D(y,y_{t+1}) - D(y_{t+1},y_t). \end{aligned} \tag{EC.11}$$

The last equality holds since $\mathbb{E}_y\left[\log\left(\frac{y_{t+1}(s,a)}{p_u(s,a)}\right)\right] = D(y,p_u) - D(y,y_{t+1})$. In addition, $D(y,y_{t+1}) - D(y,y_t)$

$$
\begin{aligned}
&= \mathbb{E}_y\left[\log\left(\frac{y(s,a)}{y_{t+1}(s,a)}\right)\right] - \mathbb{E}_y\left[\log\left(\frac{y(s,a)}{y_t(s,a)}\right)\right] \\
&= -\mathbb{E}_y\left[\log\left(\frac{y_{t+1}(s,a)}{y_t(s,a)}\right)\right] \\
&\le -\eta_t\mathbb{E}_{y_{t+1}-y}\left[\hat{f}(\theta_t,s,a)\right] - \eta_t\lambda_t D(y_{t+1},p_u) + \eta_t\lambda_t D(y,p_u) - \eta_t\lambda_t D(y,y_{t+1}) - D(y_{t+1},y_t). \quad \text{(EC.12)}
\end{aligned}
$$

The last inequality holds by (EC.11). By diving both sides of (EC.12) by $\eta_t > 0$, adding $\mathbb{E}_{y_t}[\hat{f}(\theta_t,s,a)]$ to both sides, and rearranging the terms we get

$$
\begin{aligned}
&-\mathbb{E}_y\left[\hat{f}(\theta_t,s,a)\right] - \lambda_t D(y,p_u) + \mathbb{E}_{y_t}\left[\hat{f}(\theta_t,s,a)\right] \\
&\le -(\lambda_t + \frac{1}{\eta_t})D(y,y_{t+1}) - \lambda_t D(y_{t+1},p_u) + \frac{1}{\eta_t}D(y,y_t) - \frac{1}{\eta_t}D(y_{t+1},y_t) - \mathbb{E}_{y_{t+1}-y_t}\left[\hat{f}(\theta_t,s,a)\right] \\
&\le -(\lambda_t + \frac{1}{\eta_t})D(y,y_{t+1}) - \lambda_t D(y_{t+1},p_u) + \frac{1}{\eta_t}D(y,y_t) - \frac{1}{2\eta_t}\|y_{t+1}-y_t\|_1^2 - \mathbb{E}_{y_{t+1}-y_t}\left[\hat{f}(\theta_t,s,a)\right], \\
&\qquad\qquad \text{(EC.13)}
\end{aligned}
$$

where the last inequality holds by Pinsker's inequality (Pinsker 1964): $D(y_{t+1},y_t) \ge \frac{1}{2}\|y_{t+1}-y_t\|_1^2$. Furthermore, by Young's inequality we get

$$
-\frac{1}{2\eta_t}\|y_{t+1}-y_t\|_1^2 - \mathbb{E}_{y_{t+1}-y_t}\left[\hat{f}(\theta_t,s,a)\right] \le \frac{\eta_t}{2}\|\hat{f}(\theta_t,s,a)\|_\infty^2 \le \frac{\eta_t}{2}M_y^2, \qquad \text{(EC.14)}
$$

where the last inequality follows from the definition of M_y, that is, $\|\hat{f}(\theta_t,s,a)\|_\infty \le M_y$ almost surely. Using (EC.14) in (EC.13) and dropping the non-positive term $-\lambda_t D(y,y_{t+1})$, we have

$$
-\mathbb{E}_{y-y_t}\left[\hat{f}(\theta_t,s,a)\right] - \lambda_t D(y,p_u) \le -\frac{1}{\eta_t}D(y,y_{t+1}) - \lambda_t D(y_{t+1},p_u) + \frac{1}{\eta_t}D(y,y_t) + \frac{\eta_t}{2}M_y^2 \qquad \text{(EC.15)}
$$

To obtain a bound on $-\mathbb{E}_{y-y_t}[f(\theta_t,s,a)] - \lambda_t D(y,p_u)$, we need to connect the left hand side of (EC.15) with an analogous term involving $f(\theta_t,s,a)$ instead of $\hat{f}(\theta_t,s,a)$. For this, we employ an auxiliary sequence of $\hat{y}_t \in \mathcal{Y}$ for $t \ge 0$ as follows in the spirit of Nemirovski et al. (2009). Let $\hat{y}_0 = y_0 = p_u$. For $t \ge 0$ we define

$$
\hat{y}_{t+1} \in \arg\min_{y\in\mathcal{Y}}\left[\eta_t\mathbb{E}_y\left[(f(\theta_t,s,a) - \hat{f}(\theta_t,s,a))\right] + D(y,\hat{y}_t)\right]. \qquad \text{(EC.16)}
$$

Starting from (EC.16) and applying arguments analogous to the ones used to obtain (EC.15) beginning from (14), we can show, for any $y \in \mathcal{Y}$, that

$$
-\mathbb{E}_{y-\hat{y}_t}\left[f(\theta_t,s,a) - \hat{f}(\theta_t,s,a)\right] \le -\frac{1}{\eta_t}D(y,\hat{y}_{t+1}) + \frac{1}{\eta_t}D(y,\hat{y}_t) + 2\eta_t M_y^2. \qquad \text{(EC.17)}
$$

Let $\Delta_{1t} := \mathbb{E}_{y_t - \hat{y}_t}\left[f(\theta_t, s, a) - \hat{f}(\theta_t, s, a)\right]$. We sum (EC.15) and (EC.17), and add and subtract the term $\mathbb{E}_{y_t}[f(\theta_t, s, a)]$ on the left hand side of this sum to obtain

$$-\eta_t \mathbb{E}_{y-y_t}[f(\theta_t, s, a)] - \eta_t \lambda_t D(y, p_u) \le \eta_t \Delta_{1t} - \eta_t \lambda_t D(y_{t+1}, p_u) + D(y, y_t) - D(y, y_{t+1}) + D(y, \hat{y}_t) - D(y, \hat{y}_{t+1}) + \frac{5}{2}\eta_t^2 M_y^2. \tag{EC.18}$$

Bound on $\mathbb{E}_{y_t}[f(\theta, s, a) - f(\theta_t, s, a)] + \lambda_t D(y_t, p_u)$ *for any* $\theta \in \Theta$*:* Recall that θ_{t+1} is a minimizer of (10). Starting from this optimization and applying similar arguments used to derive (EC.13) we can show for any $\theta \in \Theta$ that

$$-\left\langle \theta_t - \theta, \nabla_\theta \hat{f}(\theta_t, s, a)\right\rangle \le \frac{1}{2\eta_t}(\|\theta - \theta_t\|_2^2 - \|\theta - \theta_{t+1}\|_2^2 - \|\theta_{t+1} - \theta_t\|_2^2) - \left\langle \theta_t - \theta_{t+1}, \nabla_\theta \hat{f}(\theta_t, s, a)\right\rangle.$$

By Young's inequality we have

$$-\left\langle \theta_t - \theta_{t+1}, \nabla_\theta \hat{f}(\theta_t, s, a)\right\rangle - \frac{1}{2\eta_t}\|\theta_{t+1} - \theta_t\|_2^2 \le \frac{\eta_t}{2}\left\|\nabla_\theta \hat{f}(\theta_t, s, a)\right\|_2^2 \le \frac{\eta_t}{2} M_\theta^2,$$

where the last inequality follows from the definition of M_θ. Combining the preceding two inequalities we establish

$$-\left\langle \theta_t - \theta, \nabla_\theta \hat{f}(\theta_t, s, a)\right\rangle \le \frac{1}{2\eta_t}\left(\|\theta - \theta_t\|_2^2 - \|\theta - \theta_{t+1}\|_2^2\right) + \frac{\eta_t}{2} M_\theta^2. \tag{EC.19}$$

Analogous to the auxiliary sequence used to bound $-\mathbb{E}_{y-y_t}[f(\theta_t, s, a)] - \lambda_t D(y, p_u)$ earlier, we define $\hat{\theta}_t \in \Theta$ for $t \ge 0$. Let $\hat{\theta}_0 = \theta_0$. For $t \ge 0$ we define

$$\hat{\theta}_{t+1} = \arg\min_{\theta \in \Theta}\left[\eta_t \left\langle \hat{\theta}_t - \theta, \mathbb{E}_{y_t}[\nabla_\theta f(\theta_t, s, a)] - \nabla_\theta \hat{f}(\theta_t, s, a)\right\rangle + \frac{1}{2}\|\theta - \hat{\theta}_t\|_2^2\right].$$

Notice that M_θ can be chosen so that $\|\mathbb{E}_{y_t}[\nabla_\theta f(\theta_t, s, a)]\| \le M_\theta$ because the difference $\phi_b(s') - \phi_b(s)$ is bounded for $(s', s) \in \mathcal{S} \times \mathcal{S}$. Using an approach similar to the one involved in deriving (EC.17) and because we can show for any $\theta \in \Theta$ that

$$-\left\langle \hat{\theta}_t - \theta, \mathbb{E}_{y_t}[\nabla_\theta f(\theta_t, s, a)] - \nabla_\theta \hat{f}(\theta_t, s, a)\right\rangle \le \frac{1}{2\eta_t}\left(\left\|\theta - \hat{\theta}_t\right\|_2^2 - \left\|\theta - \hat{\theta}_{t+1}\right\|_2^2\right) + 2\eta_t M_\theta^2. \tag{EC.20}$$

Let $\Delta_{2t} := \left\langle \hat{\theta}_t - \theta_t, \mathbb{E}_{y_t}[\nabla_\theta f(\theta_t, s, a)] - \nabla_\theta \hat{f}(\theta_t, s, a)\right\rangle$. Summing (EC.19) and (EC.20), and adding and subtracting $\langle \theta_t - \theta, \mathbb{E}_{y_t}[\nabla_\theta f(\theta_t, s, a)]\rangle + \lambda_t D(y_t, p_u)$, we get

$$-\eta_t \langle \theta_t - \theta, \mathbb{E}_{y_t}[\nabla_\theta f(\theta_t, s, a)]\rangle + \eta_t \lambda_t D(y_t, p_u) \le \eta_t \Delta_{2t} + \frac{1}{2}\|\theta - \theta_t\|_2^2 - \frac{1}{2}\|\theta - \theta_{t+1}\|_2^2 + \frac{1}{2}\|\theta - \hat{\theta}_t\|_2^2 - \frac{1}{2}\|\theta - \hat{\theta}_{t+1}\|_2^2 + \frac{5}{2}\eta_t^2 M_\theta^2 + \eta_t \lambda_t D(y_t, p_u). \tag{EC.21}$$

Finally, note that $\langle \theta_t - \theta, \mathbb{E}_{y_t}[\nabla_\theta f(\theta_t, s, a)]\rangle = \mathbb{E}_{y_t}[f(\theta_t, s, a) - f(\theta, s, a)]$ since $f(\theta_t, s, a)$ is linear in θ_t.

Bounding the duality gap: We add the inequalities (EC.18) and (EC.21) and sum both sides of the summed inequality over t for $t = 0, 1, \ldots, T$ and then divide both sides by $\sum_{t=0}^{T} \eta_t$. Following the preceding steps we get

$$\begin{aligned}
&\left[\mathbb{E}_{\bar{y}}\left[f(\theta, s, a)\right] - \mathbb{E}_{y}\left[f(\bar{\theta}, s, a)\right] - \sum_{t=0}^{T} \eta_t \lambda_t D(y, p_u)/(\sum_{t=0}^{T} \eta_t)\right] \\
&\leq \frac{1}{\sum_{t=0}^{T} \eta_t}\left(D(y, y_0) + D(y, \hat{y}_0) + \frac{1}{2}\left\|\theta - \theta_0\right\|_2^2 + \frac{1}{2}\|\theta - \hat{\theta}_0\|_2^2\right) + \frac{5\left(M_\theta^2 + M_y^2\right)}{2\left(\sum_{t=0}^{T} \eta_t\right)}\left(\sum_{t=0}^{T} \eta_t^2\right) \\
&+ \frac{1}{\sum_{t=0}^{T} \eta_t}\left(\sum_{t=0}^{T} \eta_t(\Delta_{1t} + \Delta_{2t})\right) \\
&= \frac{1}{\sum_{t=0}^{T} \eta_t}\left(2D(y, p_u) + \left\|\theta - \theta_0\right\|_2^2\right) + \frac{5\left(M_\theta^2 + M_y^2\right)}{2\left(\sum_{t=0}^{T} \eta_t\right)}\left(\sum_{t=0}^{T} \eta_t^2\right) + \frac{1}{\sum_{t=0}^{T} \eta_t}\left(\sum_{t=0}^{T} \eta_t(\Delta_{1t} + \Delta_{2t})\right).
\end{aligned} \tag{EC.22}$$

Since $\eta_t = \eta_0/\sqrt{t+1}$ and $\lambda_t = \lambda_0/\sqrt{t+1}$ for all $t \geq 0$, when $T \geq 4$ we have

$$\sum_{t=0}^{T} \eta_t \geq 2\eta_0(\sqrt{T+2} - 1) \geq \eta_0\sqrt{T+2}, \text{ and } \frac{\sum_{t=0}^{T} \eta_t^2}{\sum_{t=0}^{T} \eta_t} \leq \frac{\eta_0^2\left(1 + \log(T+2)\right)}{\eta_0\sqrt{T+2}} \leq \frac{2\eta_0 \log(T+2)}{\sqrt{T+2}}.$$

Applying these two inequalities to (EC.22) and evaluating (EC.22) at $y = y^*_{\bar{\lambda}_T, \bar{\theta}}$ and $\theta = \theta^*_{\bar{y}}$, we obtain a bound on the duality gap as

$$\begin{aligned}
&\mathcal{P}(\bar{y}) - \mathbb{E}_{y^*_{\bar{\lambda}_T, \bar{\theta}}}\left[f(\bar{\theta}, s, a)\right] - \bar{\lambda}_T D(y^*_{\bar{\lambda}_T, \bar{\theta}}, p_u) \\
\leq &\frac{2}{\sqrt{T+2}\eta_0} D(y^*_{\bar{\lambda}_T, \bar{\theta}}, p_u) + \frac{1}{\sqrt{T+2}\eta_0}\left\|\theta^*_{\bar{y}} - \theta_0\right\|_2^2 + \frac{5\eta_0 \log(T+2)}{\sqrt{T+2}}\left(M_\theta^2 + M_y^2\right) + \frac{1}{\sum_{t=0}^{T} \eta_t}\sum_{t=0}^{T} \eta_t(\Delta_{1t} + \Delta_{2t}) \\
\leq &\frac{1}{\sqrt{T+2}\eta_0}(2Q_D + Q_\Theta) + \frac{5\eta_0 \log(T+2)}{\sqrt{T+2}}\left(M_\theta^2 + M_y^2\right) + \frac{1}{\sum_{t=0}^{T} \eta_t}\sum_{t=0}^{T} \eta_t(\Delta_{1t} + \Delta_{2t}) \\
= &\frac{1}{\sqrt{T+2}\eta_0} Q^2 + \frac{5\eta_0 \log(T+2)}{\sqrt{T+2}} M^2 + \frac{1}{\sum_{t=0}^{T} \eta_t}\sum_{t=0}^{T} \eta_t(\Delta_{1t} + \Delta_{2t}),
\end{aligned} \tag{EC.23}$$

where the second inequality holds by the definitions of Q_Θ and Q_D. Using Lemma EC.3, we get

$$\mathcal{D}(\bar{\theta}) = \inf_{y \in \mathcal{Y}} \mathbb{E}_y[f(\bar{\theta}, s, a)] \geq \min_{y \in \mathcal{Y}}\left[\mathbb{E}_y[f(\bar{\theta}, s, a)] + \bar{\lambda}_T D(y, p_u)\right] + \bar{\lambda}_T \bar{C} + n\bar{\lambda}_T \log(\bar{\lambda}_T).$$

Applying this inequality to (EC.23) and using the definition of $y^*_{\bar{\lambda}_T, \bar{\theta}}$, we have

$$\begin{aligned}
&\mathcal{P}(\bar{y}) - \mathcal{D}(\bar{\theta}) \\
\leq &\frac{1}{\sqrt{T+2}\eta_0} Q^2 + \frac{5\eta_0 \log(T+2)}{\sqrt{T+2}} M^2 + \frac{1}{\sum_{t=0}^{T} \eta_t}\sum_{t=0}^{T} \eta_t(\Delta_{1t} + \Delta_{2t}) - \bar{\lambda}_T \bar{C} - n\bar{\lambda}_T \log(\bar{\lambda}_T).
\end{aligned} \tag{EC.24}$$

Iteration complexity and probability guarantee for $\mathcal{P}(\bar{y})-\mathcal{D}(\bar{\theta})\le\alpha$: Note that $\mathbb{E}[\Delta_{1t}]=\mathbb{E}[\Delta_{2t}]=0$. By Cauchy-Schwarz and triangle inequalities, we can show that $|\Delta_{1t}|\le 4M_y$ and $|\Delta_{2t}|\le 2Q_\Theta M_\theta$. Therefore, the sequence $\{\eta_t(\Delta_{1t}+\Delta_{2t})\}_{t\ge 0}$ is a Martingale difference sequence (Resnick 2014, page 354) bounded by $4\eta_t M_y+2\eta_t Q_\Theta M_\theta$. Hence using $T\ge 4$ and Azuma's inequality (Azuma 1967), we get

$$\begin{aligned}&\text{Prob}\left(\frac{1}{\sum_{t=0}^T\eta_t}\sum_{t=0}^T\eta_t(\Delta_{1t}+\Delta_{2t})\ge\frac{\alpha}{8}\right)\\ &\le\exp\left(\frac{-\alpha^2(\sum_{t=0}^T\eta_t)^2}{2(8)^2(4M_y+2Q_\Theta M_\theta)^2(\sum_{t=0}^T\eta_t^2)}\right)\le\exp\left(\frac{-\alpha^2(T+2)}{4(32M_y+16Q_\Theta M_\theta)^2\log(T+2)}\right).\end{aligned}$$

Choosing

$$\begin{aligned}T+2\ge\overline{T}(\alpha,\delta):=&\max\left\{4,\left(\frac{8Q^2}{\alpha\eta_0}\right)^2,\left(\frac{160\eta_0M^2}{\alpha}\log\left(\frac{160\eta_0M^2}{\alpha}\right)\right)^2,\left(\frac{2\sqrt{2}(32M_y+16Q_\Theta M_\theta)}{\alpha}\right)^2\right.\\ &\left.\log\left(\frac{1}{\delta}\right)\left[2\log\left(\frac{2\sqrt{2}(32M_y+16Q_\theta M_\theta)}{\alpha}\right)+\log\left(\log\left(\frac{1}{\delta}\right)\right)\right],\left(\frac{8\lambda_0}{\tilde{\lambda}}\log\left(\frac{8\lambda_0}{\tilde{\lambda}}\right)\right)^2\right\}\\ =&\mathcal{O}\left(\frac{1}{\alpha^2}\log\left(\frac{1}{\alpha}\right)\log\left(\frac{1}{\delta}\right)\right),\end{aligned}\tag{EC.25}$$

we guarantee

$$\frac{1}{\sqrt{T+2}\eta_0}Q^2+\frac{5\eta_0\log(T+2)}{\sqrt{T+2}}M^2\le\frac{\alpha}{4},\quad\text{Prob}\left(\frac{1}{\sum_{t=0}^T\eta_t}\sum_{t=0}^T\eta_t(\Delta_{1t}+\Delta_{2t})\ge\frac{\alpha}{8}\right)\le\delta,$$

and

$$-\bar{\lambda}_T\bar{C}-n\bar{\lambda}_T\log(\bar{\lambda}_T)\le\frac{\alpha}{8}.$$

This indicates that if the number of iterations is greater than $\overline{T}(\alpha,\delta)$ we have

$$\begin{aligned}\text{Prob}\left(\mathcal{P}(\bar{y})-\mathcal{D}(\bar{\theta})\ge\alpha/2\right)\le\text{Prob}\Bigg(&\frac{1}{\sqrt{T+2}\eta_0}Q^2+\frac{5\eta_0\log(T+2)}{\sqrt{T+2}}M^2+\frac{1}{\sum_{t=0}^T\eta_t}\sum_{t=0}^T\eta_t(\Delta_{1t}+\Delta_{2t})\\ &-\bar{\lambda}_T\bar{C}-n\bar{\lambda}_T\log(\bar{\lambda}_T)\ge\alpha/2\Bigg)\le\delta\end{aligned}\tag{EC.26}$$

Therefore, with probability at least $1-\delta$, it is guaranteed that $\mathcal{P}(\bar{y})-\mathcal{D}(\bar{\theta})\le\alpha/2\le\alpha$. □

Proof of Theorem 2 Assume $\bar{\theta}\in\Theta$, $\bar{y}\in\mathcal{Y}$, and $\bar{\lambda}\in(0,1]$ are the inputs to Algorithm 1. We first show that $l(\bar{\theta})\le\mathcal{D}(\bar{\theta})\le F\le\mathcal{P}(\bar{y})$. Using the definition of $\mathcal{D}(\bar{\theta})$ and F we have

$$\begin{aligned}F=\max_{\theta\in\Theta}\inf_{y\in\mathcal{Y}}\mathbb{E}_y[f(\theta,s,a)]&\ge\inf_{y\in\mathcal{Y}}\mathbb{E}_y[f(\bar{\theta},s,a)]=\mathcal{D}(\bar{\theta})\\ &\ge\min_{y\in\mathcal{Y}}\left[\mathbb{E}_y[f(\bar{\theta},s,a)]+\bar{\lambda}D(y,p_u)\right]+\bar{\lambda}\bar{C}+n\bar{\lambda}\log(\bar{\lambda})\\ &=\mathbb{E}_{y^*_{\bar{\lambda},\bar{\theta}}}[f(\bar{\theta},s,a)]+\bar{\lambda}D(y^*_{\bar{\lambda},\bar{\theta}},p_u)+\bar{\lambda}\bar{C}+n\bar{\lambda}\log(\bar{\lambda})\\ &\ge\mathbb{E}_{y^*_{\bar{\lambda},\bar{\theta}}}[f(\bar{\theta},s,a)]+\bar{\lambda}\bar{C}+n\bar{\lambda}\log(\bar{\lambda})=l(\bar{\theta}),\end{aligned}$$

where the second and third inequalities follow from Lemma EC.3 and $\lambda D(y,p_u)>0$ for all $\lambda\geq 0$ and $y\in\mathcal{Y}$, respectively.

We next show that given $\alpha>0$ and $\delta\in(0,1)$, the solution $(\bar{\theta},\bar{y})$ returned by the PSMD algorithm at termination satisfies $\mathcal{P}(\bar{y})-l(\bar{\theta})\leq\alpha$ with probability of at least $1-\delta$ in at most $\mathcal{O}\left(\frac{1}{\alpha^2}\log(1/\alpha)\log(1/\delta)\right)$ iterations. This result follows directly from Theorem 1. In particular, using this theorem, we know that after $T\geq\overline{T}(\alpha,\delta)$ iterations (see (EC.25) for the definition of $\overline{T}(\alpha,\delta)$), with probability of at least $1-\delta$, we get $\mathcal{P}(\bar{y})-\mathcal{D}(\bar{\theta})\leq\alpha/2$ and $-\bar{\lambda}_T\bar{C}-n\bar{\lambda}_T\log(\bar{\lambda}_T)\leq\alpha/2$. Therefore, using the definitions of $l(\bar{\theta})$ and $\mathcal{D}(\bar{\theta})$ it follows that

$$\mathcal{P}(\bar{y})-l(\bar{\theta})\leq P(\bar{y})-\mathcal{D}(\bar{\theta})-\bar{\lambda}_T\bar{C}-n\bar{\lambda}_T\log(\bar{\lambda}_T)\leq\frac{\alpha}{2}+\frac{\alpha}{2}=\alpha.$$

Furthermore, since $l(\bar{\theta})\leq F\leq\mathcal{P}(\bar{y})$, we get $F-l(\bar{\theta})\leq\mathcal{P}(\bar{y})-l(\bar{\theta})\leq\alpha$. □

Proof of Proposition 3 The proof of this proposition follows standard arguments from the sample average approximation literature. See for example, Shapiro et al. (2009, Chapter 5). Since $\frac{1}{H'}\sum_{h=1}^{H'}\hat{f}^{N'}(\theta,\hat{s}_h,\hat{a}_h)$ is an unbiased sample average approximation of $\mathbb{E}_{\bar{y}}[f(\theta,s,a)]$, it follows that

$$\mathbb{E}_{(H',N')}\left[\frac{1}{H'}\sum_{h=1}^{H'}\hat{f}^{N'}(\theta,\hat{s}_h,\hat{a}_h)\right]=\mathbb{E}_{\bar{y}}[f(\theta,s,a)].$$

Moreover, we have

$$\frac{1}{H'}\sum_{h=1}^{H'}\hat{f}^{N'}(\theta,\hat{s}_h,\hat{a}_h)\leq\max_{\theta\in\Theta}\frac{1}{H'}\sum_{h=1}^{H'}\hat{f}^{N'}(\theta,\hat{s}_h,\hat{a}_h).$$

Taking expectations on both sides and using the unbiased nature of the sample average approximation gives

$$\mathbb{E}_{\bar{y}}[f(\theta,s,a)]=\mathbb{E}_{(H',N')}\left[\frac{1}{H'}\sum_{h=1}^{H'}\hat{f}^{N'}(\theta,\hat{s}_h,\hat{a}_h)\right]\leq\mathbb{E}_{(H',N')}\left[\max_{\theta\in\Theta}\frac{1}{H'}\sum_{h=1}^{H'}\hat{f}^{N'}(\theta,\hat{s}_h,\hat{a}_h)\right].$$

Since this inequality holds for any $\theta\in\Theta$, we can maximize over θ to obtain the required inequality

$$\mathcal{P}(\bar{y})=\max_{\theta\in\Theta}\mathbb{E}_{\bar{y}}[f(\theta,s,a)]\leq\mathbb{E}_{(H',N')}\left[\max_{\theta\in\Theta}\frac{1}{H'}\sum_{h=1}^{H'}\hat{f}^{N'}(\theta,\hat{s}_h,\hat{a}_h)\right]=\hat{\mathcal{P}}(\bar{y}).$$

We prove $\hat{l}(\bar{\theta})\leq\mathcal{D}(\bar{\theta})$ next. Recall that $\hat{y}_{\bar{\lambda},\bar{\theta}}(s,a)$ solves $\min_{y\in\mathcal{Y}}\left[\mathbb{E}_y\left[\hat{f}^{N'}(\bar{\theta},s,a)\right]+\bar{\lambda}D(y,p_u)\right]$ and is accessed using the proportionality expression (see (19))

$$\hat{y}_{\bar{\lambda},\bar{\theta}}(s,a)\propto\exp\left(-\hat{f}^{N'}(\bar{\theta},s,a)/\bar{\lambda}\right).$$

Because $\hat{y}_{\bar{\lambda},\bar{\theta}}(s,a)$ depends on $\hat{f}^{N'}(\bar{\theta},s,a)$ its definition relies on N' independent samples from the transition function $p(\cdot|s,a)$ for each state-action pair (s,a). In other words, we would obtain

a different $\hat{y}_{\bar{\lambda},\bar{\theta}}(s,a)$ density each time we repeat the sampling procedure, that is, $\hat{y}_{\bar{\lambda},\bar{\theta}}(s,a)$ can be viewed as one among a set of random density functions $\hat{\mathcal{Y}}(N')$ constructed in this manner. Let $\mathbb{E}_{\hat{\mathcal{Y}}(N')}$ denote expectation over the density functions in this set. Given a density function $\hat{y}_{\bar{\lambda},\bar{\theta}} \in \hat{\mathcal{Y}}(N')$, we denote by $\mathbb{E}_{H'|y_{\bar{\lambda},\bar{\theta}}}$ the expectation over the sets of H' independent samples (i.e., $(\hat{s}_h, \hat{a}_h)$, $h = 1, 2, \ldots, H'$) from this density. Then we have

$$\begin{aligned}
\mathbb{E}_{(H',N')}\left[\frac{1}{H'}\sum_{h=1}^{H'}\hat{f}^{N'}(\bar{\theta},\hat{s}_h,\hat{a}_h) + \bar{\lambda}D(\hat{y}_{\bar{\lambda},\bar{\theta}},p_u)\right] &= \mathbb{E}_{\hat{\mathcal{Y}}(N')}\left[\mathbb{E}_{H'|\hat{y}_{\bar{\lambda},\bar{\theta}}}\left[\frac{1}{H'}\sum_{h=1}^{H'}\hat{f}^{N'}(\bar{\theta},\hat{s}_h,\hat{a}_h)\right] + \bar{\lambda}D(\hat{y}_{\bar{\lambda},\bar{\theta}},p_u)\right] \\
&= \mathbb{E}_{\hat{\mathcal{Y}}(N')}\left[\mathbb{E}_{\hat{y}_{\bar{\lambda},\bar{\theta}}}\left[\hat{f}^{N'}(\bar{\theta},s,a)\right] + \bar{\lambda}D(\hat{y}_{\bar{\lambda},\bar{\theta}},p_u)\right] \\
&= \mathbb{E}_{\hat{\mathcal{Y}}(N')}\left[\min_{y\in\mathcal{Y}}\left\{\mathbb{E}_y\left[\hat{f}^{N'}(\bar{\theta},s,a)\right] + \bar{\lambda}D(y,p_u)\right\}\right] \\
&\le \mathbb{E}_{\hat{\mathcal{Y}}(N')}\left[\mathbb{E}_{y^*_{\bar{\lambda},\bar{\theta}}}\left[\hat{f}^{N'}(\bar{\theta},s,a)\right] + \bar{\lambda}D(y^*_{\bar{\lambda},\bar{\theta}},p_u)\right] \\
&= \mathbb{E}_{y^*_{\bar{\lambda},\bar{\theta}}}\left[\mathbb{E}_{\hat{\mathcal{Y}}(N')}\left[\hat{f}^{N'}(\bar{\theta},s,a)\right]\right] + \bar{\lambda}D(y^*_{\bar{\lambda},\bar{\theta}},p_u) \\
&= \mathbb{E}_{y^*_{\bar{\lambda},\bar{\theta}}}\left[f(\bar{\theta},s,a)\right] + \bar{\lambda}D(y^*_{\bar{\lambda},\bar{\theta}},p_u), \qquad \text{(EC.27)}
\end{aligned}$$

where the first equality follows from the definition of set $\hat{\mathcal{Y}}(N')$ and conditional expectation; the second equality because $\mathbb{E}_{H'|\hat{y}_{\bar{\lambda},\bar{\theta}}}[\frac{1}{H'}\sum_{h=1}^{H'}\hat{f}^{N'}(\bar{\theta},\hat{s}_h,\hat{a}_h)]$ is an unbiased expectation over independent sample average approximations of $\mathbb{E}_{\hat{y}_{\bar{\lambda},\bar{\theta}}}\left[\hat{f}^{N'}(\bar{\theta},s,a)\right]$; the third from the definition of $\hat{y}_{\bar{\lambda},\bar{\theta}}$; the inequality because $y^*_{\bar{\lambda},\bar{\theta}}$ is feasible but not necessarily optimal to the inner minimization; the fourth equality follows as a result of $y^*_{\bar{\lambda},\bar{\theta}}$ being independent of the collection of N' samples at each state-action pair from $\hat{\mathcal{Y}}(N')$; and the last equality because $\hat{f}^{N'}(\bar{\theta},s,a)$ is a sample average approximation of the function $f(\bar{\theta},s,a)$.

Using (EC.27) and the definition of $\hat{l}(\bar{\theta})$, we have

$$\begin{aligned}
\hat{l}(\bar{\theta}) &\le \mathbb{E}_{(H',N')}\left[\frac{1}{H'}\sum_{h=1}^{H'}\hat{f}^{N'}(\bar{\theta},\hat{s}_h,\hat{a}_h) + \bar{\lambda}D(\hat{y}_{\bar{\lambda},\bar{\theta}},p_u)\right] + \bar{\lambda}\bar{C} + n\bar{\lambda}\log(\bar{\lambda}) \\
&\le \mathbb{E}_{y^*_{\bar{\lambda},\bar{\theta}}}\left[f(\bar{\theta},s,a)\right] + \bar{\lambda}D(y^*_{\bar{\lambda},\bar{\theta}},p_u) + \bar{\lambda}\bar{C} + n\bar{\lambda}\log(\bar{\lambda}) \\
&= \min_{y\in\mathcal{Y}}\left[\mathbb{E}_y[f(\bar{\theta},s,a)] + \bar{\lambda}D(y,p_u)\right] + \bar{\lambda}\bar{C} + n\bar{\lambda}\log(\bar{\lambda}) \\
&\le \mathcal{D}(\bar{\theta})
\end{aligned}$$

where the first inequality follows from the definition of $\hat{l}(\bar{\theta})$ and the non-negativity of $\bar{\lambda}D(\hat{y}_{\bar{\lambda},\bar{\theta}},p_u)$, the second is a consequence of (EC.27), the equality uses the definition of $y^*_{\bar{\lambda},\bar{\theta}}$, and the last inequality is due to Lemma EC.3. □

Proof of Proposition 4 To prove that $a_{FB}(s)$ is determined by the myopic optimization problem stated in the proposition, we first show that the state $s = (z, q_1, q_2, \ldots, q_{J-1})$ collapses to a scalar $Z = z + \sum_{j=1}^{J-1} q_j$. We then establish that the desired myopic optimization is sufficient

to compute an optimal order quantity. The structure of our proof mirrors the one in the lecture notes of Gallego (2003).

The in-transit inventory elements $q_1,\ldots,q_{J-1}$ do not affect the system cost until they become on-hand inventory, that is, they replace z. This feature is evident from the definition of the MDP immediate cost function $c(s,a)$. Therefore, an order quantity a at the current period incurs its first cost J periods in the future and this cost will be a function of the on-hand inventory $z+\sum_{j=1}^{J-1} q_j + a - \sum_{j=0}^{J} G_j$ at that future stage, where G_j is the a realization of random demand at a stage, j periods in the future. The cost associated with this future on-hand inventory discounted back to the current period is

$$\gamma^J c_p a + \gamma^J \mathbb{E}\left[c_h\left(Z+a-G(J)\right)_+ + c_b \min\{\left(G(J)-Z-a\right)_+, -l_s\} + c_l\left(G(J)-Z-a+l_s\right)_+\right]. \quad \text{(EC.28)}$$

We remind that reader that $G(J)=\sum_{j=0}^{J} G_j$. Defining

$$C(x) := \gamma^J \mathbb{E}\left[c_h\left(x-G(J)\right)_+ + c_b \min\{\left(G(J)-x\right)_+, -l_s\} + c_l\left(G(J)-x+l_s\right)_+\right].$$

We rewrite (EC.28) as $\gamma^J c_p a + C(Z+a)$, which is an alternative definition of the immediate cost for the problem. Since its dependence on the state s is only through the scalar variable Z, we obtain the desired state-space collapse and can express the problem of computing an optimal ordering policy using the following stochastic dynamic program with value function $V^*(Z)$:

$$V^*(Z) = \min_{a\geq 0} \gamma^J c_p a + C(Z+a) + \gamma \mathbb{E}\left[V^*(Z+a-G)\right], \quad \forall Z\in\mathbb{R}. \quad \text{(EC.29)}$$

The remaining steps of the proof focus on showing that an optimal action $a^*(Z)$ to (EC.29) solves

$$\min_{a\geq 0}\left[(1-\gamma)\gamma^J c_p a + C(Z+a)\right], \quad \text{(EC.30)}$$

which is the myopic optimization problem determining $a_{FB}(s)$. Consider the function and variable transformations $U^*(Z) := V^*(Z) + \gamma^J c_p Z$ and $Z' = Z + a$, respectively. Applying them on (EC.29) and (EC.30) results in the equivalent problems

$$U^*(Z) = \gamma^{J+1} c_p \mathbb{E}[G] + \min_{Z'\geq Z}\left[(1-\gamma)\gamma^J c_p Z' + C(Z') + \gamma \mathbb{E}\left[U^*(Z'-G)\right]\right], \quad \text{(EC.31)}$$

and

$$\min_{Z'\geq Z}\left[(1-\gamma)\gamma^J c_p Z' + C(Z')\right]. \quad \text{(EC.32)}$$

In obtaining (EC.32) from (EC.30), we dropped the constant term $(\gamma-1)\gamma^J c_p Z$ as it does not affect the optimal solution. Based on the definition of $U^*(\cdot)$, it follows immediately that it is a non-decreasing function. Moreover, due to the convexity of $C(\cdot)$, standard arguments can be used to show that $U^*(\cdot)$ is also a convex function. The convex and non-decreasing nature of $U^*(\cdot)$ will

facilitate our ensuing analysis to prove that the optimal solutions to (EC.31) and (EC.32) are identical, which then implies that the same holds for the pair of equivalent optimization problems (EC.29) and (EC.30).

Consider the optimal solution to an unconstrained version of the optimization in (EC.31):

$$Z^* \in \arg\min_{Z'} \left[(1-\gamma)\gamma^J c_p Z' + C(Z') + \gamma \mathbb{E}\left[U^*(Z'-G)\right]\right].$$

- When $Z^* \geq Z$, clearly we have Z^* being a minimizer in (EC.31). Thus, in this case, we have

$$U^*(Z) = \gamma^{J+1} c_p \mathbb{E}[G] + (1-\gamma)\gamma^J c_p Z^* + C(Z^*) + \gamma \mathbb{E}\left[U^*(Z^*-G)\right]. \tag{EC.33}$$

The unconstrained optimal value in the right-hand side of the above expression is independent of Z. Denoting this constant by U^{min}, we have $U^*(Z) = U^{min}$ for all $Z \leq Z^*$.

- When $Z^* < Z$, the optimal solution to the minimization problem in the definition of $U^*(Z)$ will be Z because $U^*(Z)$ is a convex function.

Next we show that Z^* is also a minimizer of $\min_{Z'}[(1-\gamma)\gamma^J c_p Z' + C(Z')]$. To prove this statement by contradiction, suppose Z^* is not a minimizer of the latter optimization problem. Then there exists a $\hat{Z} \neq Z^*$ such that

$$\min_{Z'} \left[(1-\gamma)\gamma^J c_p Z' + C(Z')\right] = (1-\gamma)\gamma^J c_p \hat{Z} + C(\hat{Z}) < (1-\gamma)\gamma^J c_p Z^* + C(Z^*). \tag{EC.34}$$

Once again we consider two cases.

- Suppose $\hat{Z} < Z^*$. We note that

$$\begin{aligned}
&\gamma^{J+1} c_p \mathbb{E}[G] + (1-\gamma)\gamma^J c_p Z^* + C(Z^*) + \gamma \mathbb{E}\left[U^*(Z^*-G)\right] \\
&\qquad \leq \gamma^{J+1} c_p \mathbb{E}[G] + (1-\gamma)\gamma^J c_p \hat{Z} + C(\hat{Z}) + \gamma \mathbb{E}\left[U^*(\hat{Z}-G)\right] \\
&\qquad < \gamma^{J+1} c_p \mathbb{E}[G] + (1-\gamma)\gamma^J c_p Z^* + C(Z^*) + \gamma \mathbb{E}\left[U^*(Z^*-G)\right],
\end{aligned} \tag{EC.35}$$

where the first inequality follows from the definition of Z^* (i.e. Z^* is the minimizer of (EC.31)) and the second inequality holds since $\hat{Z} - G < Z^* - G \leq Z^*$ and $U^*(Z) = U^{min}$ for all $Z \leq Z^*$. We have a contradiction, which rules out $\hat{Z} < Z^*$.

- Suppose $\hat{Z} \geq Z^*$. Due to the convexity of $U^*(\cdot)$ and the cost function $C(\cdot)$, the optimal solution of $\min_{Z' \geq \hat{Z}}[(1-\gamma)\gamma^J c_p Z' + C(Z') + \gamma \mathbb{E}[U^*(Z'-G)]]$ is $\hat{Z}$. Since $U^*(\cdot)$ is non-decreasing (i.e., $U^*(\hat{Z}-G) \leq U^*(\hat{Z})$), we have

$$\begin{aligned}
U^*(\hat{Z}) &= \gamma^{J+1} c_p \mathbb{E}[G] + (1-\gamma)\gamma^J c_p \hat{Z} + C(\hat{Z}) + \gamma \mathbb{E}\left[U^*(\hat{Z}-G)\right] \\
&\leq \gamma^{J+1} c_p \mathbb{E}[G] + (1-\gamma)\gamma^J c_p \hat{Z} + C(\hat{Z}) + \gamma U^*(\hat{Z}).
\end{aligned} \tag{EC.36}$$

Once again we use the non-decreasing property of the function $U^*(\cdot)$ to write

$$\begin{aligned}(1-\gamma)U^*(Z^*) &\le (1-\gamma)U^*(\hat{Z}) \\ &\le \gamma^{J+1}c_p\mathbb{E}[G] + (1-\gamma)\gamma^J c_p\hat{Z} + C(\hat{Z}) \\ &< \gamma^{J+1}c_p\mathbb{E}[G] + (1-\gamma)\gamma^J c_p Z^* + C(Z^*),\end{aligned}$$

where the second inequality follows from rearranging terms in (EC.36) and last inequality is due to (EC.34). The preceding inequality, equation (EC.33), and $U^*(Z^*) = U^*(Z^* - G) = U^{min}$ result in the following contradiction:

$$\begin{aligned}U^{min} &< \gamma^{J+1}c_p\mathbb{E}[G] + (1-\gamma)\gamma^J c_p Z^* + C(Z^*) + \gamma U^*(Z^*) \\ &= \gamma^{J+1}c_p\mathbb{E}[G] + (1-\gamma)\gamma^J c_p Z^* + C(Z^*) + \gamma\mathbb{E}[U^*(Z^* - G)] = U^{min}.\end{aligned}$$

This contradiction rules out $\hat{Z} \ge Z^*$.

Thus, we have shown that Z^* solves $\min_{Z' \ge Z}(1-\gamma)\gamma^J c_p Z' + C(Z')$ when $Z^* \ge Z$. If $Z^* < Z$, then Z is the optimal solution since $(1-\gamma)\gamma^J c_p Z' + C(Z')$ is convex in Z'. □

EC.2. Termination of PSMD based on a relative optimality gap

In this section, we discuss how PSMD can be terminated using a relative optimality gap. We define a relative optimality gap with respect to a reference feasible solution as done in the math programming literature (see, e.g., Balas and Saxena 2008 in integer programming, and Aravkin et al. 2016 in first order methods). Recall that the absolute optimality gap of a feasible ALP solution (τ, θ) is $\text{AGAP}(\tau,\theta) := (\tau^* - \tau) + \sum_{b=1}^{B}(\theta_b^* - \theta_b)\mathbb{E}_q[\phi_b(s)]$, where (τ^*, θ^*) is an optimal ALP solution. Suppose we have a reference feasible ALP solution $(\hat{\tau}, \hat{\theta})$. Given a target relative optimality gap $\alpha_r \in (0, 1]$, we stop PSMD when it finds an improved feasible ALP solution $(\bar{\tau}, \bar{\theta})$ such that

$$\frac{\text{AGAP}(\bar{\tau}, \bar{\theta})}{\text{AGAP}(\hat{\tau}, \hat{\theta})} \le \alpha_r. \tag{EC.37}$$

By virtue of this condition, the feasible solution $(\bar{\tau}, \bar{\theta})$ closes the absolute optimality gap of the feasible solution $(\hat{\tau}, \hat{\theta})$ by at least $(1 - \alpha_r) \times 100\%$.

The left-hand side of the condition (EC.37) is not computable as it requires knowledge of the unknown optimal solution (τ^*, θ^*). We thus construct a conservative approximation of this ratio by upper bounding and lower bounding its numerator and denominator, respectively. Specifically, we leverage the inequality $\text{AGAP}(\bar{\tau}, \bar{\theta}) \le \mathcal{P}(\bar{y}) - l(\bar{\theta})$, which follows from the the definition of the primal-dual gap in Theorem 2, and the relationship $l(\bar{\theta}) - \hat{\tau} - \sum_{b=1}^{B}\hat{\theta}_b\mathbb{E}_q[\phi_b(s)] \le \text{AGAP}(\hat{\tau}, \hat{\theta})$,

which follows from $l(\bar{\theta})$ being a lower bound on the optimal ALP objective function value F (also due to Theorem 2). The resulting stopping criterion is

$$\frac{\mathcal{P}(\bar{y}) - l(\bar{\theta})}{l(\bar{\theta}) - \hat{\tau} - \sum_{b=1}^{B} \hat{\theta}_b \mathbb{E}_q \left[\phi_b(s)\right]} \leq \alpha_r, \tag{EC.38}$$

which is a sufficient condition for a feasible ALP solution $(\bar{\tau}, \bar{\theta})$ to satisfy (EC.37).

Proposition EC.1 establishes that the solution returned by PSMD satisfies the relative optimality gap condition (EC.38) in a finite number of iterations with high probability. Our iteration complexity depends on the parameter $\bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r) := \alpha_r \mathrm{AGAP}(\hat{\tau}, \hat{\theta})/(1+\alpha_r)$, which captures the intuitive effect that achieving relative optimality gap reductions is more difficult when the absolute optimality gap of the reference solution $(\hat{\tau}, \hat{\theta})$ is small. Recall the definition of $\bar{\tau}(\bar{\theta})$ in (6).

Proposition EC.1. *Let $\alpha_r > 0$ and $\delta \in (0,1)$. Suppose we have a reference feasible ALP solution $(\hat{\tau}, \hat{\theta})$. Then PSMD returns the pair $(\bar{\theta}, \bar{y})$ such that the feasible ALP solution $(\bar{\tau}(\bar{\theta}), \bar{\theta})$ satisfies (EC.38), and hence (EC.37), in at most*

$$\mathcal{O}\left(\frac{1}{\bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)^2} \log\left(\frac{1}{\bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)}\right) \log\left(\frac{1}{\delta}\right)\right)$$

iterations with probability of at least $1-\delta$.

Proof. The inequality $\mathcal{P}(\bar{y}) - l(\bar{\theta}) \leq \bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)$ can be guaranteed to hold after

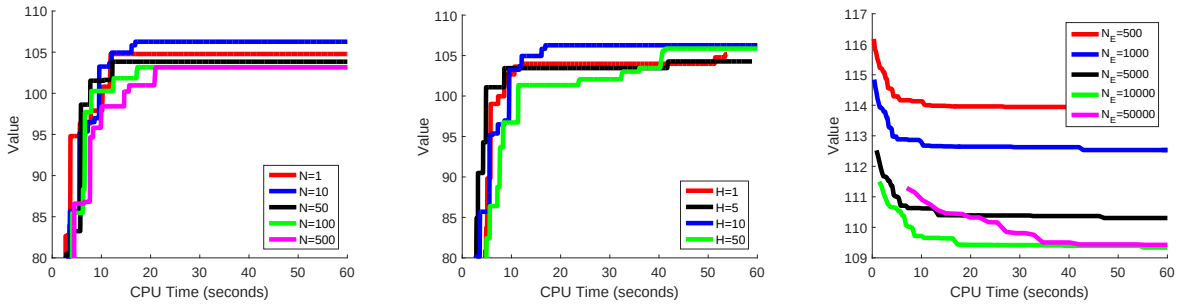
$$\begin{aligned} T+2 \geq & \max\left\{4, \left(\frac{8Q^2}{\bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)\eta_0}\right)^2, \left(\frac{160\eta_0 M^2}{\bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)} \log\left(\frac{160\eta_0 M^2}{\bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)}\right)\right)^2, \left(\frac{2\sqrt{2}(32M_y + 16Q_\Theta M_\theta)}{\bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)}\right)^2 \right. \\ & \left. \log\left(\frac{1}{\delta}\right)\left[2\log\left(\frac{2\sqrt{2}(32M_y + 16Q_\theta M_\theta)}{\bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)}\right) + \log\left(\log\left(\frac{1}{\delta}\right)\right)\right], \left(\frac{8\lambda_0}{\tilde{\lambda}} \log\left(\frac{8\lambda_0}{\tilde{\lambda}}\right)\right)^2 \right\} \\ = & \mathcal{O}\left(\frac{1}{\bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)^2} \log\left(\frac{1}{\bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)}\right) \log\left(\frac{1}{\delta}\right)\right), \end{aligned}$$

iterations by following the same steps as the proof of Theorem 1 with α replaced by $\bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)$. Moreover,

$$\frac{\mathrm{AGAP}(\bar{\tau}, \bar{\theta})}{\mathrm{AGAP}(\hat{\tau}, \hat{\theta})} \leq \frac{\mathcal{P}(\bar{y}) - l(\bar{\theta})}{l(\bar{\theta}) - \hat{\tau} - \sum_{b=1}^{B} \hat{\theta}_b \mathbb{E}_q \left[\phi_b(s)\right]} \leq \frac{\bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)}{\mathrm{AGAP}(\hat{\tau}, \hat{\theta}) - \bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)} = \alpha_r,$$

where the second inequality follows from $F - l(\bar{\theta}) \leq \mathcal{P}(\bar{y}) - l(\bar{\theta}) \leq \bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)$ and the equality holds by simplifying expressions after writing $\mathrm{AGAP}(\hat{\tau}, \hat{\theta})$ in terms of $\bar{\alpha}(\hat{\tau}, \hat{\theta}, \alpha_r)$ and α_r. □

When implementing the relative gap stopping criterion, we can replace the upper bound $\mathcal{P}(\bar{y})$ and the lower bound $l(\bar{\theta})$ in (EC.38) by $\hat{\mathcal{P}}(\bar{y})$ and $\hat{l}(\bar{\theta})$, respectively, as discussed in §6. If a reference feasible ALP solution is not available, one can be constructed as follows: (i) Run the PSMD algorithm for a fixed number of iterations (100 iterations for example) and obtain (θ, y); (ii) compute $\hat{\mathcal{P}}(y)$ as described in §6; and (iii) define the reference feasible ALP solution $(\hat{\tau}, \hat{\theta})$ as $\hat{\theta} = \theta$ and $\hat{\tau} = \hat{\mathcal{P}}(y) - \sum_{b=1}^{B} \theta_b \mathbb{E}_q[\phi_b(s)]$, where feasibility follows since $\hat{\mathcal{P}}(y) \geq \bar{\tau}(\theta) + \sum_{b=1}^{B} \theta_b \mathbb{E}_q[\phi_b(s)]$ by Proposition 3.



(a) PSMD lower bound trajectories for different N (H fixed at 10).

(b) PSMD lower bound trajectories for different H (N fixed at 10).

(c) ALP-CS objective function value trajectories for different N_E.

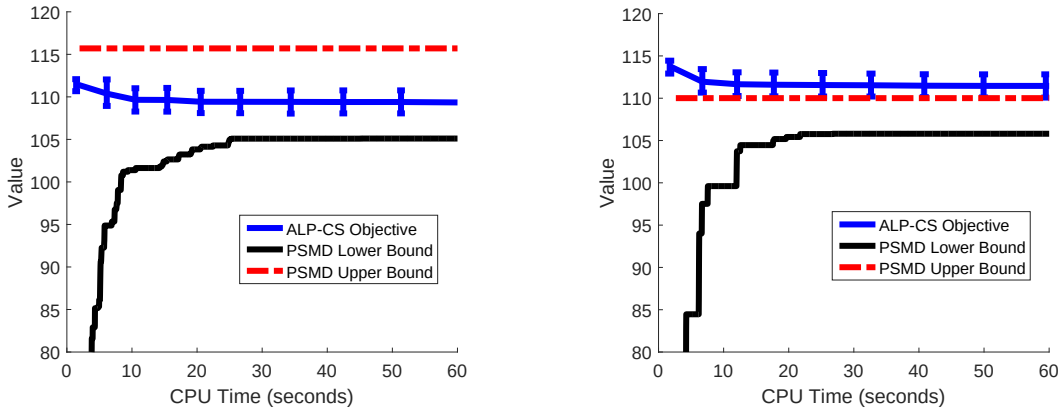
Figure EC.1 Changes in the PSMD lower bound and ALP-CS objective function values on a representative consumer energy storage instance.

EC.3. Addendum to numerical results

We provide computational results supplementing §7.4. Figures EC.1(a) and EC.1(b) show the PSMD lower bound behavior as functions of time for different values of N and H, respectively, on the consumer energy storage instance corresponding to row one of Table 3. These results are representative of the behavior of PSMD on other instances and shows that $N = H = 10$ provides bounds of good quality in reasonable time. We thus used this choice for our PSMD implementation.

Figure EC.1(c) shows the behavior of the ALP-CS objective function value with CPU time for different values of N_E, which denotes the number of samples used to approximate the expectations in the ALP constraints. We chose N_E equal to 10,000 as it decreased the ALP-CS objective the fastest.

Finally, in Figure EC.2 we investigate the behavior of the ALP-CS objective and PSMD lower bound against a PSMD upper bound on two consumer energy storage instances corresponding to rows one and five of Table 3. As expected, the PSMD lower bound is always below the upper bound. On the instance considered in Figure EC.2(a) the ALP-CS objective function value is below the upper bound but this is not the case for the instance corresponding to Figure EC.2(b). Thus, the ALP-CS objective may not provide a lower bound on the optimal objective function value.



(a) Instance defined in row one of Table 3.

(b) Instance defined in row five of Table 3.

Figure EC.2 Comparison of the PSMD upper bound against the PSMD lower bound and the ALP-CS objective function value on two representative energy storage instances.

References

Aravkin, A. Y., J. V. Burke, D. Drusvyatskiy, M. P. Friedlander, S. Roy. 2016. Level-set methods for convex optimization. *arXiv preprint arXiv:1602.01506* .

Azuma, K. 1967. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series* **19**(3) 357–367.

Balas, E., A. Saxena. 2008. Optimizing over the split closure. *Mathematical Programming* **113**(2) 219–240.

Nemirovski, A., A. Juditsky, G. Lan, A. Shapiro. 2009. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization* **19**(4) 1574–1609.

Pinsker, M. S. 1964. *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco, USA.

Resnick, S. I. 2014. *A Probability Path*. Springer Science & Business Media, New York, NY, USA.

Shapiro, A., D. Dentcheva, A. Ruszczyński. 2009. *Lectures on stochastic programming: modeling and theory*. SIAM, Philadelphia, PA.