Large-Scale Markov Decision Problems via the Linear Programming Dual

Yasin Abbasi-Yadkori, Adobe Research
Peter L. Bartlett, UC Berkeley
Xi Chen, NYU
Alan Malek, Simons Institute

January 8, 2019

Abstract

We consider the problem of controlling a fully specified Markov decision process (MDP), also known as the planning problem, when the state space is very large and calculating the optimal policy is intractable. Instead, we pursue the more modest goal of optimizing over some small family of policies. Specifically, we show that the family of policies associated with a low-dimensional approximation of occupancy measures yields a tractable optimization. Moreover, we propose an efficient algorithm, scaling with the size of the subspace but not the state space, that is able to find a policy with low *excess loss* relative to the best policy in this class. To the best of our knowledge, such results did not exist in the literature previously. We bound excess loss in the average cost and discounted cost cases, which are treated separately. Preliminary experiments show the effectiveness of the proposed algorithms in a queueing application.

1 Introduction

The Markov Decision Process planning problem is to find a good policy given complete knowledge of the transition dynamics and loss function. Much work has been done by the reinforcement learning community; the earliest approaches with convergence guarantees date back to value iteration [Bellman, 1957], policy iteration [Howard, 1960], and other dynamic programming ideas. Another thread has been the linear programming formulation [Manne, 1960]. In general, the planning problem is well understood for state-spaces small enough to permit computation of the value function [Bertsekas, 2007]. However, in large state space problems, both the dynamic programming and linear program approaches are computationally infeasible as complexity scales quadratically with the number of states.

A popular approach to large-scale problems is to search for the optimal value function within the linear span of a small number of features with the hope that the optimal value function will be well approximated and will lead to a near optimal policy. Two popular methods are Approximate Dynamic Programming (ADP) and Approximate Linear Programming (ALP). For a survey on

theoretical results for ADP, see [Bertsekas and Tsitsiklis, 1996], [Bertsekas, 2007, Vol. 2, Chapter 6], and more recent papers [Sutton et al., 2009b,a, Maei et al., 2009, 2010].

Our goal is to find an almost-optimal policy in some low dimensional space such that the complexity scales with the low dimensional space but is sublinear in the size of the state space. In contrast, all prior work on ALP either scales badly or requires access to samples from a distribution that depends on the optimal policy. To accomplish this, we will use randomized algorithms to optimize policies that are parameterized by linear functions in the dual LP. We provide performance bounds in the average loss and discounted loss cases. In particular, we introduce new proof techniques and tools for average cost and discounted cost MDP problems and use these techniques to derive a reduction to stochastic convex optimization with accompanying error bounds.

1.1 Markov Decision Process

Markov decision processes have become a popular approach to modeling an agent interacting with an environment, and, most notably, are the model assumed by reinforcement learning. Using $[N] = \{1, \ldots, N\}$, an MDP is parameterized by:

1. a discrete state space $\{1, 2, \ldots, \mathcal{X}\}$,
2. a discrete action space $\{1, 2, \ldots, \mathcal{A}\}$,
3. transition dynamics $P : [\mathcal{X}] \times [\mathcal{A}] \to \triangle_{[\mathcal{X}]}$ that describes the distribution of the next states x' given a current state and action (x, a), and
4. loss function $\ell : [\mathcal{X}] \times [\mathcal{A}] \to [0, 1]$ that provides the cost of taking an action in a given state.

The (fully observed) state encapsulates all the persistent information of the environment, and the influence of the agent is captured through the transition distribution, which is a function of the current state and the current action.

A policy $\pi : [\mathcal{X}] \to \triangle_{[\mathcal{A}]}$ gives a distribution over actions for every possible state, and the goal of the learner is to identify a policy with small loss. Throughout, we will use x and a to refer to specific states and actions, respectively. Given some random variable X_0 for the starting distribution and some fixed policy π, the distribution of the random variable of the initial action A_0 is fixed. Then, given the transition dynamics P and π, we can calculate the random trajectory $X_1, A_1, X_2, A_2, \ldots$. The random variables X_t and A_t will always refer to the random state and actions induced by a fixed policy π, the transition dynamics, and initial distribution of X_0. Using this random variable notation, we will write $P(X_{t+1} = x' | X_t = x, A_t = a)$ to refer to the x'th entry of $P(x, a)$, i.e. the probability of transitioning to state x' from state x when action a is taken.

How can we evaluate a policy? The two most common metrics are average cost and discounted cost. Average cost is roughly the expected loss of the policy once the Markov chain has reached stationarity and disregards the transient dynamics. Discounted cost minimizes the cost where future

losses t rounds into the future are discounted by γ^t, where $\gamma \in (0,1)$ is some discounting factor. Therefore, discounted cost emphasized the short-term cost and roughly only considers $1/(1-\gamma)$ rounds into the future. Precisely,

$$\lambda_\pi(x) \stackrel{\text{def}}{=} \lim_{n\to\infty} \mathbb{E}\left[\frac{1}{n}\sum_{t=0}^{n} \ell(X_t, \pi(X_t)) \;\middle|\; X_0 = x\right] \qquad \text{(average cost), and} \tag{1}$$

$$J_\pi(x) \stackrel{\text{def}}{=} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \ell(X_t, \pi(X_t)) \;\middle|\; X_0 = x\right] \qquad \text{(discounted cost)} \tag{2}$$

The initial state is very relevant for J but irrelevant for λ under the usual regularity (it is sufficient to assume that the induced Markov chain is recurrent [Puterman, 1994]). We study the average cost in Section 2 and the discounted cost in Section 4.

1.2 Notation

It will be convenient to be able to write the transition dynamics as a matrix multiplication. For vectors $v \in \mathbb{R}^{\mathcal{X}\mathcal{A}}$ over state-action pairs, we will write $v(x,a)$ for the element corresponding to state x and action a. The specific mapping from (x,a) to $\{1,\ldots,\mathcal{X}\mathcal{A}\}$ is irrelevant, so just pick one and fix your favorite. We can then define the matrix $P \in \mathbb{R}^{\mathcal{X}\mathcal{A}\times\mathcal{X}}$ to have row $P(x,a)$ in the (x,a) position; therefore, if v is a probability distribution over X_t, A_t, then $v^\intercal P \in \triangle_{\mathcal{X}}$ is the distribution over X_{t+1}. We can also define the vector of losses ℓ to have value $\ell(x,a)$ in position x,a.

Given a vector v and a matrix $M \in \mathbb{R}^{\mathcal{X}\times\mathcal{X}}$, we will use $v(i)$ for the ith component of vector v and $M_{i,:}$, $M_{:,j}$, and M_{ij} for the the ith row, jth column, and element in the i,j position of M, respectively. For matrices $M \in \mathbb{R}^{\mathcal{X}\mathcal{A}\times\mathcal{X}}$, where the first index is over state-action pairs, we will define $M_{(x,a),:}$ to be the row corresponding to (x,a) and $M_{:,x}$ to be the column, over state-action pairs, corresponding to the xth column.

Any distribution over state-action pairs μ defines a policy π_μ with

$$\pi_\mu(a|x) = \frac{\mu(x,a)}{\sum_{a' \in [\mathcal{A}]} \mu(x,a')}, \tag{3}$$

with $\pi_\mu(a|x) = \mathcal{A}^{-1}$ if $\mu(x,a) = 0$ for all a. This is simply the conditional distribution of A given X. We will also define the marginalization matrix $B \in \{0,1\}^{\mathcal{X}\mathcal{A}\times\mathcal{X}}$ to be the binary matrix such that the xth coordinate of $v^\intercal B$ is $\sum_a v(x,a)$. If v is a probability distribution over X_t, A_t, then $v^\intercal B$ is the marginal of X_t.

For some fixed policy π, we would also like to refer to the induced state transition matrix, P^π, defined by

$$(P^\pi)_{x,x'} = \sum_a P(X_{t+1} = x' | X_t = x, A_t = a)\pi(A_t = a | X_t = x),$$

so that if $X_t \sim v$, then $X_{t+1} \sim P^{\pi\intercal} v$ if policy π is used.

We will use the norms $\|v\|_{1,c} = \sum_i c_i |v_i|$ and $\|v\|_{\infty,c} = \max_i c_i |v_i|$ (for a positive vector c). The constant one and zero vector are $\mathbf{1}$ and $\mathbf{0}$, and $\wedge$ and $\vee$ refer to the element-wise minimum and maximum. We can then compactly define $[v]_- = v \wedge 0$ and $[v]_+ = v \vee 0$ as the negative and positive parts of a vector v, respectively. Finally, $v \leq w$ for two vectors means element-wise inequality, i.e. $v_i \leq w_i$ for all i.

1.3 Linear Programming for Average Cost

For the average cost, let $h \in \mathbb{R}^{\mathcal{X}}$ be a vector and $\lambda \in \mathbb{R}$ a scalar. The *Bellman operator for average cost* is

$$\boldsymbol{L}h(x) \stackrel{\text{def}}{=} \min_{a \in [\mathcal{A}]} \left[\ell(x,a) + \sum_{x' \in \mathcal{X}} P_{(x,a),x'} h(x') \right],$$

and h and λ correspond to an optimal policy if they satisfy the Bellman optimality equation,

$$\lambda + h(x) = \boldsymbol{L}h(x) \quad \forall x.$$

We will call such an h and λ the differential value function and the average cost, respectively. When the Bellman optimality equation is satisfied, the greedy policy (taking the action that achieves the minimum in the operator with probability 1) achieves the optimal loss [Puterman, 1994].

The Bellman optimality equation was first recast as a linear program by Manne [1960], who noted that, if λ and h satisfy $\boldsymbol{L}h \geq h + \lambda \mathbf{1}$, then we must have $\lambda \leq \lambda^*$, where λ^* is the average cost of the optimal policy. Therefore, the optimal λ and h are the solution to

$$\begin{aligned} &\max_{\lambda, h} \lambda\,, \\ &\text{s.t.} \quad h + \lambda \mathbf{1} \leq \boldsymbol{L}h. \end{aligned}$$

Now, notice that $h(x) + \lambda \leq \min_a \left[\ell(x,a) + \sum_y P(y|x,a) h(y) \right]$ is equivalent to requiring $h(x) + \lambda \leq \ell(x,a) + \sum_y P(y|x,a) h(y)$ for all x and a. In our matrix notation, this is precisely $B(\lambda \mathbf{1} + h) \leq \ell + Ph$. Hence, the Bellman optimality equation is equivalent to the linear program

$$\begin{aligned} &\max_{\lambda, h} \lambda\,, \\ &\text{s.t.} \quad B(\lambda \mathbf{1} + h) \leq \ell + Ph\,. \end{aligned} \tag{4}$$

A standard computation shows that the dual of LP (4) has the form of

$$\begin{aligned} &\min_{\mu \in \mathbb{R}^{\mathcal{X}\mathcal{A}}} \mu^\intercal \ell \,, \\ &\text{s.t.} \quad \mu^\intercal \mathbf{1} = 1,\, \mu \geq \mathbf{0},\, \mu^\intercal (P - B) = \mathbf{0} \,, \end{aligned} \tag{5}$$

The dual variable, μ, has an important interpretation: it is a stationary distribution over state-action pairs under its implied policy.

The first two constraints ensure that μ is a probability distribution over state-action space and the third constraint forces μ to be a stationary distribution under π_μ. Intuitively, if $X_t \sim \mu^\intercal B$, then $\mu^\intercal P$ is the distribution of X_{t+1} under policy π_μ; hence, the third constraint implies that X_t and X_{t+1} have the same distribution. If μ is a stationary distribution, then the average loss under μ is exactly $\mu^\intercal \ell$.

1.4 Linear Programming for Discounted Cost

There are analogous notions for the discounted cost setting. We define a value function $J : [\mathcal{X}] \to \mathbb{R}$ as a mapping from states to discounted costs. The hope is to find J^*, where $J^*(x)$ is the discounted cost starting in state x if the optimal policy is used.

We define the *Bellman operator for discounted cost*

$$\boldsymbol{L}^\gamma J(x) \stackrel{\text{def}}{=} \min_{a \in [\mathcal{A}]} \left[\ell(x,a) + \gamma \sum_{x' \in [\mathcal{X}]} P_{(x,a),x'} J(x') \right]$$

and the optimal value function will be the fixed point of the Bellman operator,

$$L^\gamma J^* = J^*.$$

It is easy to check that $J \leq \boldsymbol{L}^\gamma J$ implies $J \leq J^*$, and therefore, for any strictly positive vector $\alpha \in \mathbb{R}^{\mathcal{X}}$, the optimal value function is the solution to the linear program

$$\begin{aligned} &\max_J \alpha^\intercal J \\ &\text{s.t.} \quad \boldsymbol{L}^\gamma J \geq J. \end{aligned} \tag{6}$$

We also have an interpretable dual LP. Let α be such that $\alpha \geq 0$ and $\alpha^\intercal \mathbf{1} = 1$. The linear

program for discounted MDPs in the dual space has the form of

$$\min_{\nu \in \mathbb{R}^{\mathcal{X}\mathcal{A}}} \nu^\intercal \ell \,, \tag{7}$$
$$\text{s.t.} \quad (B - \gamma P)^\intercal \nu = \alpha, \quad \nu \geq 0, \quad \nu^\intercal \mathbf{1} = \frac{1}{1-\gamma}.$$

Unlike the average cost case, the dual variable ν cannot be interpreted as a stationary distribution. However, it can be thought of as the discounted number of visits, as made explicit in the following theorem from Puterman [1994]:

Theorem 1. *1. For each randomized Markovian policy π and state x and action a, define $\nu_\pi(x,a)$ by*

$$\nu_\pi(x,a) = \sum_{x'} \alpha(x') \sum_{t=1}^{\infty} \gamma^{t-1} P^\pi(x_t = x, a_t = a \mid x_1 = x') \,.$$

Then ν_π is a feasible solution to the dual problem.

2. Suppose ν is a feasible solution to the dual problem, then, for each $x \in [\mathcal{X}]$, $\sum_a \nu(x,a) > 0$. Define the randomized stationary policy π_ν by

$$\pi_\nu(a|x) = \frac{\nu(x,a)}{\sum_{a'} \nu(x,a')} \,.$$

Then, ν_{π_ν} is a feasible solution to the dual LP and $\nu_{\pi_\nu} = \nu$.

Thus, we can approximately solve the planning problem if we find a vector z such that the discounted cost of the policy defined by z, namely $\ell^\intercal \nu_{\pi_z}$, is small. To handle possibly negative entries of z, we more generally define

$$\pi_z(a|x) = \frac{[z(x,a)]_+}{\sum_{a'} [z(x,a')]_+} \,.$$

In this case, the precise relationship between ν_{π_z} and the value function can be found in Puterman [1994]: for any vector z,

$$\sum_{x,a} \nu_{\pi_z}(x,a) = \frac{1}{1-\gamma} \quad \text{and} \quad \nu_{\pi_z}^T \ell = \alpha^T J_{\pi_z}, \tag{8}$$

where J is the value function corresponding to policy π_z.

1.5 Approximate Linear Programming

If we ignore computational constraints, we can solve the planning problem by solving the linear programs (5) and (7). Unfortunately, state spaces are frequently very large and often grow expo-

nentially with the complexity of the system (e.g. number of queues in the queuing network), and therefore any method polynomial in $\mathcal{X}$ becomes intractable. The general method of solving the planning problem with an approximate solution to the linear program is called Approximate Linear Programming (ALP). As any general optimality guarantee is impossible with computation sublinear in $\mathcal{X}$ without special knowledge of the problem, we instead aim for optimality with respect to some smaller policy class.

We take the less common approach of reducing the dimensionality by placing a subspace restriction of the dual variables. Let $\Phi \in \mathbb{R}^{\mathcal{X}\mathcal{A}\times d}$ by a feature matrix and μ_0 some known stationary distribution (that can be taken to be zero but allows a user to start with a good policy). For the average cost case, we will limit our search to $\mu = \mu_0 + \Phi\theta$ for $\theta \in \Theta \subset \mathbb{R}^d$; that is, we will study the *approximate average cost dual LP*,

$$\begin{aligned} &\min_{\theta\in\Theta} \ (\mu_0 + \Phi\theta)^\intercal \ell \,, \\ &\text{s.t.} \quad (\mu_0 + \Phi\theta)^\intercal \mathbf{1} = 1,\ \mu_0 + \Phi\theta \geq \mathbf{0},\ (\mu_0 + \Phi\theta)^\intercal (P - B) = \mathbf{0} \,. \end{aligned} \tag{9}$$

we will only consider θ that sum to 1 and will restrict Θ to lie in $\{x \in \mathbb{R}^d : x^\intercal \mathbf{1} = 1\}$. This restriction is without loss of generality, since we may always renormalize Φ.

For every θ, we associate a policy

$$\pi_\theta(a|x) = \frac{[\mu_0(x,a) + \Phi_{(x,a),:}\theta]_+}{\sum_{a'}[\mu_0(x,a') + \Phi_{(x,a'),:}\theta]_+} \tag{10}$$

and a stationary distribution μ_θ the actual stationary distribution of running policy π_θ. Thus, the average cost corresponding to the policy π_θ is $\ell^\intercal \mu_\theta$.

For the discounted cost case with feature matrix Φ, we restrict the dual variable to $\nu = \Phi\theta$ and define the *approximate discounted cost dual LP*

$$\begin{aligned} &\min_{\theta\in\Theta} \ \ell^\intercal \Phi\theta \,, \\ &\text{s.t.} \quad (\Phi\theta)^\intercal \mathbf{1} = \frac{1}{1-\gamma}, \quad (B - \gamma P)^\intercal \Phi\theta = \alpha, \quad \Phi\theta \geq 0. \end{aligned}$$

For every θ, we define a policy

$$\pi_\theta(a|x) = \frac{[\Phi_{(x,a),:}\theta]_+}{\sum_{a'}[\Phi_{(x,a'),:}\theta]_+}, \tag{11}$$

and let ν_θ be the corresponding dual variable (i.e. the discounted number of visits); hence, $\ell^\intercal \nu_\theta$ is the discounted cost as in (8). In the discounted case, we will restrict Θ to lie in $\{x \in \mathbb{R}^d : x^\intercal \mathbf{1} = (1-\gamma)^{-1}\}$.

1.6 Problem Definition

The goal of the paper is to find a θ such that the associated policy π_θ is close to the policy corresponding with the best $\theta \in \Theta$ in an efficient manner and while avoiding complexity proportional to $\mathcal{X}$. This goal is formalized by the following definition.

Definition 1 (Efficient Large-Scale Dual ALP)**.** *For an MDP specified by ℓ and P with the dual variables ξ_θ corresponding to $\theta \in \Theta$, the efficient large-scale dual ALP problem is to find a $\widehat{\theta}$ such that*

$$\ell^\intercal \xi_{\widehat{\theta}} \leq \min \{\ell^\intercal \xi_\theta : \xi_\theta \text{ feasible for } (5) \text{ or } (7) \} + O(\epsilon) \tag{12}$$

in time polynomial in d and $1/\epsilon$. The model of computation allows access to arbitrary entries of Φ, ℓ, P, μ_0, $P^\intercal\Phi$, and $\ell^\intercal\Phi$ in unit time.

The computational complexity cannot scale with $\mathcal{X}$ and we do not assume any knowledge of the optimal policy. In fact, as we shall see, we solve a harder problem, which we define as follows.

Definition 2 (Expanded Efficient Large-Scale Dual ALP)**.** *Let $V : \mathbb{R}^d \to \mathbb{R}_+$ be some "violation function" that represents how far ξ_θ is from satisfying the constraints of (5) or (7) and has $V(\theta) = 0$ if θ is feasible.*

The expanded efficient large-scale dual ALP problem is to produce parameters $\widehat{\theta}$ such that

$$\ell^\intercal \xi_{\widehat{\theta}} \leq \min_{\theta \in \Theta} \{\ell^\intercal \xi_\theta + V(\theta)\} + O(\epsilon), \tag{13}$$

in time polynomial in d and $1/\epsilon$, under the same model of computation as in Definition 1.

Note that the expanded problem is strictly more general as guarantee (13) implies guarantee (12). Also, many feature vectors Φ may not admit any feasible points. In this case, the dual ALP problem is trivial, but the expanded problem is still meaningful.

In particular, we desire an agnostic learning guarantee, where the true average cost of running the policy corresponding to $\widehat{\theta}$ to be close to the true average cost of the best policy in the class, regardless of how well the policy class models the optimal value function. To the best of our knowledge, such a guarantee does not exist in the literature.

Having access to arbitrary entries of the quantities in Definition 1 arises naturally in many situations. In many cases, entries of $P^\intercal\Phi$ are easy to compute. For example, suppose that for any state x' there are a small number of state-action pairs (x, a) such that $P(x'|x, a) > 0$. Consider Tetris; although the number of board configurations is large, each state has a small number of possible neighbors. Dynamics specified by graphical models with small connectivity also satisfy this constraint. Computing entries of $P^\intercal\Phi$ is also feasible given reasonable features. If a feature ϕ_i is a stationary distribution, then $P^\intercal\phi_i = B^\intercal\phi_i$. Otherwise, it is our prerogative to design sparse feature vectors, hence making the multiplication easy. We shall see an example of this setting later.

1.7 Related Work

Approximate linear programming, proposed by Schweitzer and Seidmann [1985], constrained the value function in the linear program to a low-dimensional subspace. In the discounted cost setting, the first theoretical analysis of ALP methods, by de Farias and Van Roy [2003a], analyzed the discounted primal LP (7) performance when only value functions of the form $J = \Psi w$, for some feature matrix Ψ, are considered. Roughly, they show that the ALP solution w^* has the family of error bound indexed by a vector $u \in \mathbb{R}^{\mathcal{X}}$

$$\|\boldsymbol{J}^* - \Psi w^*\|_{1,c} \leq \inf_w \frac{2c^\intercal u}{1-\gamma\beta(u)} \|\boldsymbol{J}^* - \Psi w\|,$$

where c is a "state-relevance" vector and $\beta_u = \gamma \max_{x,a} \sum_{x'} P_{(x,a),x'} u(x')/u(x)$ is a "goodness-of-fit" parameter that measures how well u represents a stationary distribution. Unfortunately, c and u are typically hard to choose (for example, a good choice of c would be the stationary distribution under w^*, which we do not know); but more importantly, the bound can be vacuous if Ψ does not model the optimal value function well and $\|J_* - \Psi w\|$ is always large. In particular, the problem we are considering in Definition 2 requires an additive bound with respect to the optimal parameter.

There are also computational concerns with the ALP, as the number of constraints remains $O(\mathcal{X}\mathcal{A})$. One solution, proposed by de Farias and Van Roy [2004], was to sample a small number of constraints and solve the resulting LP; this resulted in an error bound of the form

$$\|J_* - \Psi\widehat{w}\|_{1,c} \leq \|J_* - \Psi w_*\|_{1,c} + \epsilon \|J_*\|_{1,c},$$

but the required number of sampled constraints needs to be a function of the stationary distribution of the optimal policy.

Desai et al. [2012] proposed a different relaxation by defining the *Smoothed Approximate Linear Program*, which only requires a soft feasibility and solves the linear program

$$\begin{aligned} &\max_J c^\intercal \Psi w \\ &\text{s.t.} \quad \boldsymbol{L}^\gamma \Psi w + s \geq \Psi w,\ s \geq 0,\ \nu_{\pi_*,\alpha}^\intercal s \leq D \end{aligned}$$

which is exact LP (6) with $J = \Psi w$, the Bellman optimality constraint relaxed with a slack variable s, and additional bounds places on s. Here, $\nu_{\pi_*,\alpha}^\intercal$ is the stationary distribution of the optimal policy and D a violation budget, so the method requires some knowledge of the optimal policy. Despite this, the method remains computationally efficient and able to produce an agnostic approximation bound

$$\|J_* - \Psi w_*\| \leq \inf_w \|J_* - \Psi w\| O(1).$$

However, their results do not easily extend to bounding the true error of running the policy associ-

ated with w^*, $\left\|J_{\nu_{\Psi w_*}} - J_*\right\|$, without choosing c as a function of w^*, which is itself a function of c. Petrik and Zilberstein [2009] proposed two different constraint relaxations schemes for the ALP, but did not show better approximations to the true solution, but rather focused on the better empirical performance. Yet another relaxation of the primal LP was proposed by Lakshminarayanan et al. [2018], who generalized previous constraint sampling approaches. A bound for the discounted loss of the policy associated with the solution to this relaxed LP is presented and neatly decomposes into an estimation error that tends to zero and an approximation error between the optimal LP solution and the optimal relaxed LP solution.

In the average cost setting, largely thought to be more difficult, shares a similar history is that the first theoretical analysis for ALP was by de Farias and Van Roy [2003b]. They proposed a two stage LP. The first approximates the optimal average cost and the second uses this estimate to try and learn the differential cost function h. The method suffers from the same problem as the discounted cost case is that we can only guarantee that $\lambda_{\hat{w}} - \lambda^*$, the excess loss of running the policy associated with $\hat{w}$, is small when we tune the LP with knowledge of $\mu_{\hat{w}}$, the stationary distribution.

Subsequent work in de Farias and Van Roy [2006] took a different approach by viewing the average cost LP as a perturbed discounted cost LP, which is easier to analyze. Again, the span of the feature vectors needs to approximate the optimal policy in order for the excess loss guarantee to be meaningful. More recently, Veatch [2013] proposed a relaxation, similar to the smoothed ALP but with the total constraint violation terms entering the objective instead of facing a hard constraint, and derived similar loss bounds.

Recently, Chen et al. [2018] analyzed a linearly parameterized ALP where the state and action spaces are both parameterized by linear features and the value function is assumed to be well approximated by linear function of the state features. They propose an efficient algorithm but suffer the same drawback and retain an error term of the form $\min_w \|\Psi w - J^*\|$. Additionally, Banijamali et al. [2019] study the related problem of optimizing policies in the convex hull of base policies. This problem can be seen as a special case of the usual ALP formulation when all features correspond to the stationary distribution of policies.

To the best of our knowledge, no work has been able to show a bound of the form (13), as all the previous bounds are only meaningful when the approximate policy class can closely approximate the optimal policy. We are also the first to prove theoretical guarantees when the dual variables of the LP are restricted to a linear class, though such a parameterization appeared previously by Wang et al. [2008], albeit without theoretical guarantees. See Section D in the appendix for a more thorough literature review and precise statements of prior bounds.

1.8 Our Contributions

We prove that if we parameterize the policy space by using the approximate dual LPs, then we can solve the expanded efficient large-scale dual ALP problem for both average cost and discounted cost. In the average cost setting, we require a (standard) assumption that the distribution of states under any policy converges quickly to its stationary distribution, but no such assumption is needed in the discounted cost setting. We also show that it suffices to solve the approximate dual LPs by approximately minimizing a surrogate loss function equal to the sum of the objective and a scaled violation function.

We begin with the average cost in Section 2 and prove that, for some parameter $H > 0$, any $\epsilon > 0$ and $\delta > 0$, the excess loss bound

$$\mu_{\widehat{\theta}}^{\intercal}\ell \leq \min_{\theta} \mu_{\theta}^{\intercal}\ell + HV(\theta) + O\left(\frac{1}{H}\log\left(\frac{1}{\delta}\right)\right) + O(\epsilon)$$

holds with probability at least $1-\delta$, where $V(\theta) = \|[\mu_0 + \Phi\theta]_-\|_1 + \|(P-B)^{\intercal}(\mu_0 + \Phi\theta)\|_1$. The $V(\theta)$ term is zero for feasible points (that is, points in the intersection of the feasible set of LP (9) and the span of the features). For points outside the feasible set, these terms measure the extent of constraint violations for the vector $\mu_0 + \Phi\theta$, which indicate how well stationary distributions can be represented by the chosen features.

However, optimizing the excess loss bound to obtain the guarantee of Definition 2 requires us to tune H correctly (in particular, setting $H \approx V(\theta)^{-1/2}$). Unfortunately, the convex surrogate is not jointly convex in θ and H. In Section 3, we present and analyze a *meta-algorithm* that solves the convex surrogate for a grid on H values and returns a $\widehat{\theta}$ that has

$$\ell^{\intercal}\mu_{\widehat{\theta}} \leq \min_{\theta\in\Theta} \ell^{\intercal}\mu_{\theta} + O\left(\sqrt{V(\theta)}\right) + O(\epsilon).$$

We emphasize that this bound is on the loss of actually running the π_θ policy, which could differ from the surrogate used in the optimization, $\ell^{\intercal}(\mu_0 + \Phi\theta)$. The run-time, up to logarithmic factors, is $O(\epsilon^{-4})$ for both algorithms; we essentially can tune H for a small logarithmic cost.

As we have seen in the related works section, all previous guarantees for efficient ADP algorithms only had meaningful guarantees when the policy class closely approximates the true value function, and many algorithms required tuning (say, of the state relevance weights) with knowledge of the optimal policy or stationary distribution. These restrictions render previous guarantees meaningless in many modern reinforcement learning systems, where the optimal value function is completely unknown and it is hopeless to try to engineer features that can approximate it [Goodfellow et al., 2016]. Our algorithm have guarantees that are meaningful in this setting, as we can obtain near-optimal excess loss within the policy class; in fact, one can use the stationary distribution of existing policies (based on DQN, heuristics, etc.) as feature vectors and improve upon them.

We then turn to the discounted cost problem in Section 4. We propose an algorithm and show that it guarantees a bound on the discounted cost of the form

$$\ell^\intercal \nu_{\widehat{\theta}_T} \leq \ell^\intercal \nu_\theta + \left(\frac{6}{1-\gamma} + H\right) V(\theta) + O\left(\frac{1}{H(1-\gamma)}\right) + O(\epsilon).$$

Furthermore, the *meta-algorithm*, with minimal modification, solves the Expanded Efficient Large-Scale Dual ALP problem by obtaining the bound

$$\ell^\intercal \nu_{\theta_{\hat{k}}} \leq \min_\theta \ell^\intercal \nu_\theta + O\left(\sqrt{V(\theta)}\right) + O(\epsilon),$$

where the violation function for the discounted cost is $V(\theta) = \| [\Phi\theta]_- \|_1 + \|(B - \gamma P)^T \Phi\theta - \alpha\|$.

Section 5 then demonstrates the effectiveness of our method on a well studied example from queuing theory, the Rybko-Stolyar queue. We show that using two simple heuristic policies with a small number of simple features provides good performance.

2 The Dual ALP for Average Cost

Is this section, we propose and analyze our solution to the Expanded large-scale MDP problem for average cost. As discussed in the introduction, there are two main challenges for solving the planning problem in its LP formulation: the optimization is in dimension $\mathcal{X}$, and there are $O(\mathcal{X}\mathcal{A})$ constraints, which is intractable in the large state-space setting.

We solve the two challenges by projecting the dual LP onto a subspace and by approximately solving the optimization using stochastic gradient descent, respectively. Unlike previous approaches for the primal LP, we show that an approximate solution in the dual allows us to bound the excess loss, i.e. one that controls the error between our approximate solution and the best solution in some approximate policy class, and thereby solve Equation (13). We also provide some interpretation of the approximations we make.

Recall that, for a matrix Φ and a known stationary distribution μ_0 (which may be set to zero if no distribution is known), we defined the dual ALP

$$\begin{aligned} &\min_\theta (\mu_0 + \Phi\theta)^\intercal \ell\,, \\ &\text{s.t.} \quad (\mu_0 + \Phi\theta)^\intercal \mathbf{1} = 1,\ \mu_0 + \Phi\theta \geq \mathbf{0},\ (\mu_0 + \Phi\theta)^\intercal (P - B) = \mathbf{0} \end{aligned}$$

and associated every θ with the policy

$$\pi_\theta(a|x) = \frac{[\mu_0(x,a) + \Phi_{(x,a),:}\theta]_+}{\sum_{a'} [\mu_0(x,a') + \Phi_{(x,a'),:}\theta]_+}\,.$$

We denote the stationary distribution of this policy μ_θ, which is only equal to $\mu_0 + \Phi\theta$ if θ is in the

feasible set.

2.1 A Reduction to Stochastic Convex Optimization

Unfortunately, the ALP (9) still has $O(\mathcal{X}\mathcal{A})$ constraints and cannot be solved exactly. Instead, we will use the penalty method to form an unconstrained convex optimization that will act as a surrogate for the original problem and show that it is a finite sum, e.g. equal to $\sum_{i=1}^{N} f_i(\theta)$. Therefore, we can apply the extensive literature of solving finite sum problems with stochastic subgradient descent methods.

To this end, for a constant $H \geq 1$, define the following convex cost function by adding a multiple of the total constraint violations to the objective of the LP (9):

$$\begin{aligned} c(\theta) &\stackrel{\text{def}}{=} \ell^\intercal(\mu_0 + \Phi\theta) + H \left\|[\mu_0 + \Phi\theta]_-\right\|_1 + H \left\|(P-B)^\intercal(\mu_0 + \Phi\theta)\right\|_1 \\ &= \ell^\intercal(\mu_0 + \Phi\theta) + H \left\|[\mu_0 + \Phi\theta]_-\right\|_1 + H \left\|(P-B)^\intercal\Phi\theta\right\|_1 \\ &= \ell^\intercal(\mu_0 + \Phi\theta) + H \sum_{(x,a)} \left|[\mu_0(x,a) + \Phi_{(x,a),:}\theta]_-\right| + H \sum_{x'} \left|(P-B)^\intercal_{:,x'}\Phi\theta\right| \ . \end{aligned} \tag{14}$$

We justify using this surrogate function as follows. Suppose we find a near optimal vector $\widehat{\theta}$ such that $c(\widehat{\theta}) \leq \min_{\theta\in\Theta} c(\theta) + O(\epsilon)$. We will prove

1. that $\left\|[\mu_0 + \Phi\widehat{\theta}]_-\right\|_1$ and $\left\|(P-B)^\intercal(\mu_0 + \Phi\widehat{\theta})\right\|_1$ are small and $\mu_0 + \Phi\widehat{\theta}$ is close to $\mu_{\widehat{\theta}}$ (Lemma 3), and
2. that $\ell^\intercal(\mu_0 + \Phi\widehat{\theta}) \leq \min_{\theta\in\Theta} c(\theta) + O(\epsilon)$.

As we will show, these two facts imply that with high probability, for any $\theta \in \Theta$,

$$\mu_{\widehat{\theta}}^\intercal \ell \leq \mu_\theta^\intercal \ell + \frac{1}{\epsilon}\left\|[\mu_0 + \Phi\theta]_-\right\|_1 + \frac{1}{\epsilon}\left\|(P-B)^\intercal(\mu_0 + \Phi\theta)\right\|_1 + O(\epsilon).$$

Unfortunately, calculating the gradients of $c(\theta)$ is $O(\mathcal{X}\mathcal{A})$. Instead, we construct unbiased estimators and use stochastic subgradient descent. Let T be the number of iterations of our algorithm, q_1 and q_2 be distributions over the state-action and state space, respectively (we will later discuss how to choose them), and $((x_t, a_t))_{t=1\ldots T}$ and $(x'_t)_{t=1\ldots T}$ be i.i.d. samples from these distributions. At round t, the algorithm estimates subgradient $\nabla c(\theta)$ by

$$g_t(\theta) = \ell^\intercal\Phi - H\frac{\Phi_{(x_t,a_t),:}}{q_1(x_t,a_t)}\mathbb{I}\{\mu_0(x_t,a_t) + \Phi_{(x_t,a_t),:}\theta < 0\} + H\frac{(P-B)^\intercal_{:,x'_t}\Phi}{q_2(x'_t)}s((P-B)^\intercal_{:,x'_t}\Phi\theta). \tag{15}$$

This estimate is fed to the projected subgradient method, which in turn generates a vector θ_t. After T rounds, we average vectors $(\theta_t)_{t=1\ldots T}$ and obtain the final solution $\widehat{\theta}_T = \sum_{t=1}^{T}\theta_t/T$. Vector $\mu_0 + \Phi\widehat{\theta}_T$ defines a policy, which in turn defines a stationary distribution $\mu_{\widehat{\theta}_T}$. The algorithm is shown in Figure 1.

Input: Constants S and H, number of rounds T, step size η.
Let Π_Θ be the Euclidean projection onto Θ.
Initialize $\theta_1 = 0$.
for $t := 1, 2, \ldots, T$ **do**
  Sample $(x_t, a_t) \sim q_1$ and $x'_t \sim q_2$.
  Compute subgradient estimate g_t (15).
  Update $\theta_{t+1} = \Pi_\Theta(\theta_t - \eta_t g_t)$.
end for
$\widehat{\theta}_T = \frac{1}{T}\sum_{t=1}^{T} \theta_t$.
Return policy $\pi_{\widehat{\theta}_T}$.

Figure 1: The Stochastic Subgradient Method for Markov Decision Processes

2.2 Excess Loss bound

We now turn towards proving the main result of this section, Theorem 2, which requires a (standard) assumption that any policy quickly converges to its stationary distribution.

Assumption A1 ***(Fast Mixing)*** For any policy π, there exists a constant $\tau(\pi) > 0$ such that for all distributions μ and μ' over the state space, $\left\|\mu^\intercal P^\pi - \mu'^\intercal P^\pi\right\|_1 \leq e^{-1/\tau(\pi)} \left\|\mu - \mu'\right\|_1$.

Define

$$C_1 = \max_{(x,a)\in[\mathcal{X}]\times[\mathcal{A}]} \frac{\left\|\Phi_{(x,a),:}\right\|}{q_1(x,a)}, \qquad C_2 = \max_{x\in[\mathcal{X}]} \frac{\left\|(P-B)^\intercal_{:,x}\Phi\right\|}{q_2(x)}.$$

These constants appear in our excess loss bounds, so we would like to choose distributions q_1 and q_2 such that C_1 and C_2 are small. Several common scenarios permit convenient C_1 and C_2:

- **Sparseness of** P If there is $C' > 0$ such that for any (x, a) and i, $\Phi_{(x,a),i} \leq C'/(\mathcal{X}\mathcal{A})$ and each column of P has only N non-zero elements, then we can simply choose q_1 and q_2 to be uniform distributions and

$$\frac{\left\|\Phi_{(x,a),:}\right\|}{q_1(x,a)} \leq C', \qquad \frac{\left\|(P-B)^\intercal_{:,x}\Phi\right\|}{q_2(x)} \leq C'(N+\mathcal{A}).$$

- **Features as stationary distributions** If every feature is the stationary distribution of some policy, then we can choose $q_1(x,a) \propto \min\{\Phi_{(x,a),y} : y \in \mathcal{X}\}$ and $\left\|(P-B)^\intercal_{:,x}\Phi\right\|$ vanishes.

- **Exponential distributions** If $\Phi_{:,i}$ are exponential distributions and feature values at neighboring states are close to each other, then we can choose q_1 and q_2 to be appropriate exponential distributions so that $\left\|\Phi_{(x,a),:}\right\| / q_1(x,a)$ and $\left\|(P-B)^\intercal_{:,x}\Phi\right\| / q_2(x)$ are always bounded.

- **One step look-ahead** When the columns of Φ are close to their *one step look-ahead*, there exists a constant $C'' > 0$ such that for any x, $\|P_{:,x}^{\intercal}\Phi\| / \|B_{:,x}^{\intercal}\Phi\| < C''$. If we are also able to compute $Z_1 = \sum_{(x,a)} \left\|\Phi_{(x,a),:}\right\|$ and $Z_2 = \sum_x \|B_{:,x}^{\intercal}\Phi\|$, then it is natural to take $q_1(x,a) = \left\|\Phi_{(x,a),:}\right\| / Z_1$ and $q_2(x) = \|B_{:,x}^{\intercal}\Phi\| / Z_2$.

In what follows, we assume that such distributions q_1 and q_2 are known.

Minimizing the convex surrogate function does not guarantee a feasible solution to the original dual LP. Therefore, we define the following non-feasibility penalties which roughly correspond to how far $\Phi\theta$ is from the simplex and how far $\Phi\theta$ is from a stationary distribution, respectively:

$$V_1(\theta) \stackrel{\text{def}}{=} \sum_{(x,a)} \left|[\mu_0(x,a) + \Phi_{(x,a),:}\theta]_-\right| \text{ and}$$
$$V_2(\theta) \stackrel{\text{def}}{=} \|(P-B)^{\intercal}(\Phi\theta)\|_1 = \sum_{x'} \left|(P-B)_{:,x'}^{\intercal}\Phi\theta\right|.$$

The rest of the section proves the following theorem, our main guarantee for the stochastic subgradient method.

Theorem 2. *Consider an expanded efficient large-scale dual ALP problem and some error tolerance $\epsilon > 0$ and desired maximum probability of error $\delta > 0$. Then running the stochastic subgradient method (shown in Figure 1) with*

$$T \geq \max\left\{\frac{H^2}{\epsilon^2}, 40S^2\log\frac{1}{\delta}\right\} \quad \text{and} \quad \eta = \left(\sqrt{d} + H(C_1+C_2)\right)\frac{S}{\sqrt{T}},$$

yields a $\widehat{\theta}_T$ where

$$\ell^{\intercal}\mu_{\widehat{\theta}_T} \leq \ell^{\intercal}\mu_{\theta} + 2\left(H + O(1)\right)\left(V_1(\theta) + V_2(\theta)\right) + O\left(\frac{1}{H}\right) + O\left(\epsilon\right),$$

holds with probability at least $1-\delta$. In particular, for the choice of $H = \epsilon^{-1}$, the bound becomes

$$\ell^{\intercal}\mu_{\widehat{\theta}_T} \leq \ell^{\intercal}\mu_{\theta} + O\left(\frac{1}{\epsilon}\right)\left(V_1(\theta) + V_2(\theta)\right) + O\left(\epsilon\right). \tag{16}$$

Constants hidden in the big-O notation are polynomials in S, d, C_1, C_2, $\log(1/\delta)$, $\log(V_1(\theta)+V_2(\theta))$, $\tau(\mu_\theta)$, and $\tau(\mu_{\widehat{\theta}_T})$.

Functions V_1 and V_2 are bounded by small constants for any set of normalized features: for any

$\theta \in \Theta$,

$$
\begin{aligned}
V_1(\theta) &\le \|\mu_0\|_1 + \|\Phi\theta\|_1 \le 1 + \sum_{(x,a)} \left|\Phi_{(x,a),:}\theta\right| \le 1 + Sd\,, \\
V_2(\theta) &\le \sum_{x'} \left|P_{:,x'}^{\mathsf{T}}(\mu_0 + \Phi\theta)\right| + \sum_{x'} \left|B_{:,x'}^{\mathsf{T}}(\mu_0 + \Phi\theta)\right| \\
&\le \left(\sum_{x'} P_{:,x'}\right)^{\mathsf{T}} [\mu_0 + \Phi\theta]_+ + \left(\sum_{x'} B_{:,x'}\right)^{\mathsf{T}} [\mu_0 + \Phi\theta]_+ \\
&= 2[\mu_0 + \Phi\theta]_+^{\mathsf{T}}\mathbf{1} \\
&\le 2\,|\mu_0 + \Phi\theta|^{\mathsf{T}}\,\mathbf{1} \\
&= 2 + 2S\,.
\end{aligned}
$$

Thus V_1 and V_2 can be very small given a carefully designed set of features. The output $\widehat{\theta}_T$ is a random vector as the algorithm is based on a stochastic convex optimization method. The above theorem shows that with high probability the policy implied by this output is near optimal.

The optimal choice for ϵ is $\epsilon = \sqrt{V_1(\theta_*) + V_2(\theta_*)}$, where θ_* is the minimizer of RHS of (16) and not known in advance. One could think of parameterizing the optimization problem by H, but the problem is not jointly convex in H and θ. Nevertheless, we present methods that recover a $O(\sqrt{V_1(\theta_*) + V_2(\theta_*)})$ error bound using a grid based method in Section 3.

2.3 Analysis

This section provides the necessary technical tools and a proof of the main result. We break the proof into two main ingredients. First, we demonstrate that a good approximation to the surrogate loss gives a feature vector that is almost a stationary distribution; this is Lemma 3. Second, we justify the use of unbiased gradients in Theorem 4 and Lemma 6. The section concludes with the proof of Theorem 2. Long, technical proofs have been moved to Section A when we felt that their inclusion did not add much insight.

The first ingredient shows that we can relate the magnitude of the constraint violation of θ to the difference between $\Phi\theta$ and μ_θ, which quantifies how far $\Phi\theta$ is from a stationary distribution.

Lemma 3. *Let $u \in \mathbb{R}^{\mathcal{X}\mathcal{A}}$ be a vector, $\mathcal{N}$ be the set of points (x,a) where $u(x,a) < 0$, and $\mathcal{S}$ be the complement of $\mathcal{N}$. Assume*

$$
\sum_{x,a} u(x,a) = 1,\ \sum_{(x,a)\in\mathcal{N}} |u(x,a)| \le \epsilon',\ \|u^{\mathsf{T}}(P - B)\|_1 \le \epsilon''.
$$

The vector $[u]_+ / \|[u]_+\|_1$ defines a policy, which in turn defines a stationary distribution μ_u. We have that

$$
\|\mu_u - u\|_1 \le \tau(\mu_u)\log(1/\epsilon')(2\epsilon' + \epsilon'') + 3\epsilon'\,.
$$

The second ingredient is the validity of the subgradient estimates. We assume access to estimates of the subgradient of a convex cost function. Error bounds can be obtained from results in the stochastic convex optimization literature; the following theorem, a high-probability version of Lemma 3.1 of Flaxman et al. [2005] for stochastic convex optimization, is sufficient. We note that the variance reduced stochastic gradient descent literature (e.g. SAGA or SVGR) cannot be directly applied since a full gradient calculation is impossible, and most complexity upper bounds are at least $O(\sqrt{\mathcal{X}\mathcal{A}}/\epsilon)$ [Xiao and Zhang, 2014], which is inappropriate for our setting.

Theorem 4. *Consider a bounded set $\mathcal{Z} \subset \mathbb{R}^d$ of radius Z (i.e. $\|z\| \leq Z$ for all $z \in \mathcal{Z}$) and a sequence of real-valued convex cost functions $(f_t)_{t=1,2,\ldots,T}$. Let $z_1, z_2, \ldots, z_T \in \mathcal{Z}$ be the stochastic gradient decent path defined by defined by $z_1 = 0$ and $z_{t+1} = \Pi_{\mathcal{Z}}(z_t - \eta f'_t)$, where $\Pi_{\mathcal{Z}}$ is the Euclidean projection onto $\mathcal{Z}$, $\eta > 0$ is a learning rate, and $f'_1, \ldots, f'_T$ are bounded unbiased subgradient estimates; that is, $\mathbb{E}[f'_t|z_t] = \nabla f(z_t)$ and $\|f'_t\| \leq F$ for some $F > 0$. Then, for $\eta = Z/(F\sqrt{T})$ and any $\delta \in (0,1)$,*

$$\sum_{t=1}^{T} f_t(z_t) - \min_{z\in\mathcal{Z}} \sum_{t=1}^{T} f_t(z) \leq ZF\sqrt{T} + \sqrt{(1+4Z^2T)\left(2\log\frac{1}{\delta} + d\log\left(1+\frac{Z^2T}{d}\right)\right)} \tag{17}$$

with probability at least $1-\delta$.

Proof. Let $z_* = \arg\min_{z\in\mathcal{Z}} \sum_{t=1}^{T} f_t(z)$ and $\eta_t = f'_t - \nabla f_t(z_t)$. Define function $h_t : \mathcal{Z} \to \mathbb{R}$ by $h_t(z) = f_t(z) + z\eta_t$. Notice that $\nabla h_t(z_t) = \nabla f_t(z_t) + \eta_t = f'_t$. By Theorem 1 of Zinkevich [2003], we get that

$$\sum_{t=1}^{T} h_t(z_t) - \sum_{t=1}^{T} h_t(z_*) \leq \sum_{t=1}^{T} h_t(z_t) - \min_{z\in\mathcal{Z}} \sum_{t=1}^{T} h_t(z) \leq ZF\sqrt{T}\,.$$

Thus,

$$\sum_{t=1}^{T} f_t(z_t) - \sum_{t=1}^{T} f_t(z_*) \leq ZF\sqrt{T} + \sum_{t=1}^{T} (z_* - z_t)\eta_t\,.$$

Let $S_t = \sum_{s=1}^{t-1}(z_* - z_s)\eta_s$, which is a self-normalized sum [de la Peña et al., 2009]. By Corollary 3.8 and Lemma E.3 of Abbasi-Yadkori [2012], we get that for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\begin{aligned}
|S_t| &\leq \sqrt{\left(1 + \sum_{s=1}^{t-1}(z_t - z_*)^2\right)\left(2\log\frac{1}{\delta} + d\log\left(1+\frac{Z^2t}{d}\right)\right)} \\
&\leq \sqrt{(1+4Z^2t)\left(2\log\frac{1}{\delta} + d\log\left(1+\frac{Z^2t}{d}\right)\right)}\,.
\end{aligned}$$

Thus,

$$\sum_{t=1}^{T} f_t(z_t) - \min_{z\in\mathcal{Z}} \sum_{t=1}^{T} f_t(z) \le ZF\sqrt{T} + \sqrt{(1+4Z^2T)\left(2\log\frac{1}{\delta} + d\log\left(1+\frac{Z^2T}{d}\right)\right)}\,.$$

□

Remark 5. *Let B_T denote the RHS of (17). If all cost functions are equal to f, then by convexity of f and an application of Jensen's inequality, we obtain that $f(\sum_{t=1}^{T} z_t/T) - \min_{z\in\mathcal{Z}} f(z) \le B_T/T$.*

The last step before giving the proof of Theorem 2 is to apply Theorem 4 to our convex surrogate function, $c(\theta)$.

Lemma 6. *Under the same conditions as in Theorem 2 and any $\delta \in (0,1)$*

$$c(\widehat{\theta}_T) - \min_{\theta\in\Theta} c(\theta) \le \frac{S(\sqrt{d} + H(C_1+C_2))}{\sqrt{T}} + \sqrt{\frac{1+4S^2T}{T^2}\left(2\log\frac{1}{\delta} + d\log\left(1+\frac{S^2T}{d}\right)\right)} \tag{18}$$

with probability at least $1-\delta$,

The proof (in the appendix) consists of checking that conditions of Theorem 4 are satisfied With both ingredients in place, we can prove our main result.

Proof of Theorem 2. Let b_T be the RHS of (18). Using the trivial fact that $\sqrt{a+b} \le 2\sqrt{a} + 2\sqrt{b}$, we can easily derive

$$b_T \le \frac{S}{\sqrt{T}}\left((\sqrt{d} + H(C_1+C_2)) + 2\sqrt{10\log\frac{1}{\delta}} + 2\sqrt{5d\log\left(1+\frac{S^2T}{d}\right)}\right) + O\left(\frac{1}{T}\right). \tag{19}$$

Lemma 6 implies that with high probability for any $\theta \in \Theta$,

$$\ell^\intercal(\mu_0 + \Phi\widehat{\theta}_T) + H\,V_1(\widehat{\theta}_T) + H\,V_2(\widehat{\theta}_T) \le \ell^\intercal(\mu_0 + \Phi\theta) + H\,V_1(\theta) + H\,V_2(\theta) + b_T\,. \tag{20}$$

From (20), we get that

$$V_1(\widehat{\theta}_T) \le \frac{1}{H}\left(2(1+S) + H\,V_1(\theta) + H\,V_2(\theta) + b_T\right) \stackrel{\text{def}}{=} \epsilon'\,, \tag{21}$$

$$V_2(\widehat{\theta}_T) \le \frac{1}{H}\left(2(1+S) + H\,V_1(\theta) + H\,V_2(\theta) + b_T\right) \stackrel{\text{def}}{=} \epsilon''\,. \tag{22}$$

Inequalities (21) and (22) and Lemma 3 give the following bound:

$$\left|\ell^\intercal\mu_{\widehat{\theta}_T} - \ell^\intercal(\mu_0 + \Phi\widehat{\theta}_T)\right| \le \tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon')(2\epsilon' + \epsilon'') + 3\epsilon'\,, \tag{23}$$

and we can similarly bound

$$|\ell^\intercal \mu_\theta - \ell^\intercal(\mu_0 + \Phi\theta)| \leq \tau(\mu_\theta)\log(1/V_1(\theta))(2V_1(\theta) + V_2(\theta)) + 3V_1(\theta). \tag{24}$$

Combining these two equations with (20) gives the final result:

$$\begin{aligned}
\ell^\intercal \mu_{\widehat{\theta}_T} &\leq \ell^\intercal(\mu_0 + \Phi\widehat{\theta}_T) + \tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon')(2\epsilon' + \epsilon'') + 3\epsilon' \\
&\leq \ell^\intercal(\mu_0 + \Phi\theta_T) + \tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon')(2\epsilon' + \epsilon'') + 3\epsilon' + HV_1(\theta) + HV_2(\theta) + b_T \\
&\leq \ell^\intercal\mu_\theta + \tau(\mu_\theta)\log(1/V_1(\theta))(2V_1(\theta) + V_2(\theta)) + 3V_1(\theta) \\
&\quad + \tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon')(2\epsilon' + \epsilon'') + 3\epsilon' + HV_1(\theta) + HV_2(\theta) + b_T \\
&\leq \ell^\intercal\mu_\theta + 2\left(V_1(\theta) + V_2(\theta)\right)\left(3 + \tau(\mu_\theta)\log(1/V_1(\theta)) + \tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon') + H\right) \\
&\quad + \left(2\tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon') + 3\right)\frac{2(1+S)}{H} + (2\tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon') + 3)\frac{b_T}{H} + b_T.
\end{aligned}$$

Using the form of b_T above, we find the excess loss bound

$$\begin{aligned}
\ell^\intercal \mu_{\widehat{\theta}_T} &\leq \ell^\intercal\mu_\theta + 2\left(V_1(\theta) + V_2(\theta)\right)\left(3 + \tau(\mu_\theta)\log(1/V_1(\theta)) + \tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon') + H\right) \\
&\quad + \left(2\tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon') + 3\right)\frac{2(1+S)}{H} + \frac{S}{\sqrt{T}}H(C_1 + C_2) \\
&\quad + \left(\frac{2\tau(\mu_{\widehat{\theta}_T})\log(1/\epsilon') + 3}{H} + 2\right)\frac{S}{\sqrt{T}}\sqrt{10\log\frac{1}{\delta}} \\
&\quad + O\left(\frac{\log(T)}{\sqrt{T}}\right) + O\left(\frac{1}{\sqrt{T}H}\right) \qquad (25) \\
&\leq \ell^\intercal\mu_\theta + 2\left(V_1(\theta) + V_2(\theta)\right)(H + O(1)) + O\left(\frac{1}{H}\right) + O\left(\frac{H}{\sqrt{T}}\right) \\
&\quad + O\left(\frac{1}{H\sqrt{T}}\right)\sqrt{\log\frac{1}{\delta}} + O\left(\frac{\log(T)}{\sqrt{T}}\right) \qquad (26)
\end{aligned}$$

Now, recall that we set

$$T = \max\left\{\frac{H^2}{\epsilon^2}, 40S^2\log\frac{1}{\delta}\right\},$$

which finally yields that with high probability, for any $\theta \in \Theta$,

$$\ell^\intercal\mu_{\widehat{\theta}_T} \leq \ell^\intercal\mu_\theta + 2\left(H + O(1)\right)\left(V_1(\theta) + V_2(\theta)\right) + O\left(\frac{1}{H}\right) + O\left(\epsilon\right),$$

as claimed. □

2.4 Comparison with Previous results

With a precise statement of our main result, we return to compare Theorem 2 from de Farias and Van Roy [2006]. Their approach is to relate the original MDP to a perturbed version [1] and then analyze the corresponding ALP. Let Ψ be a feature matrix that is used to estimate value functions. Recall that λ_* is the average loss of the optimal policy and λ_w is the average loss of the greedy policy with respect to value function Ψw. Let h^*_γ be the differential value function when the restart probability in the perturbed MDP is $1-\gamma$. For vector v and positive vector u, define the weighted maximum norm $\|v\|_{\infty,u} = \max_x u(x)\,|v(x)|$. de Farias and Van Roy [2006] prove that for appropriate constants $C, C' > 0$ and weight vector u,

$$\lambda_{w_*} - \lambda_* \leq \frac{C}{1-\gamma} \min_w \left\| h^*_\gamma - \Psi w \right\|_{\infty,u} + C'(1-\gamma)\,. \tag{27}$$

This bound has similarities to bound (16): tightness of both bounds depends on the quality of feature vectors in representing the relevant quantities (stationary distributions in (16) and value functions in (27)). Once again, we emphasize that the algorithm proposed by de Farias and Van Roy [2006] is computationally expensive and requires access to a distribution that depends on optimal policy.

3 Average Cost Meta-Algorithm

The previous section proved that Algorithm 1 found a $\widehat{\theta}_T$ with

$$\mu_{\widehat{\theta}_T}^\intercal \ell \leq \min_{\theta\in\Theta} \left(\ell^\intercal \mu_\theta + 2\left(H + O(1)\right)\left(V_1(\theta) + V_2(\theta)\right) + O\left(\frac{1}{H}\right) + O\left(\epsilon\right) \right),$$

where H is a hyperparameter and ϵ is some error tolerance. If one has reason to believe that the violation terms $V_1(\theta) + V_2(\theta)$ are negligible (for example, if the features are close to stationary distributions), then one can set $H = \epsilon^{-1}$. However, we wish to be adaptive to the size of the constrain violations around the optimum θ^*, and ideally obtain the excess loss bound

$$\ell^\intercal \mu_{\widehat{\theta}_T} \leq \min_{\theta\in\Theta} \ell^\intercal \mu_\theta + O\left(\sqrt{V_1(\theta) + V_2(\theta)}\right) + O(\epsilon),$$

which would imply that we have solved the Expanded Efficient Large-Scale Dual ALP problem (Definition 2) with violation $V(\theta) = \sqrt{V_1(\theta) + V_2(\theta)}$.

Unfortunately, we must jointly optimize over θ and H and the objective is not jointly convex. We avoid this difficulty with a *meta-algorithm*, proposed and analyzed in this section.

This meta-algorithm, detailed in Figure 2, uses Algorithm 1 to approximate $\widehat{\theta}$ over a grid

[1] In a perturbed MDP, the state process restarts with a certain probability to a *restart distribution*. Such perturbed MDPs are closely related to discounted MDPs.

Input: Upper bound $V_{\max}$ on $V_1(\theta) + V_2(\theta)$, error tolerance $\epsilon > 0$, error probability $\delta > 0$, constraint estimation distributions q_1 and q_2
Initialize $H_0 \leftarrow \beta \left(\sqrt{V_{\max}}\right)^{-1}$ and $i \leftarrow 0$
while $H_i \leq \frac{2\beta}{\epsilon}$ **do**
  Set $H_{i+1} \leftarrow H_i + \epsilon \left(V_{\max} + \frac{\beta}{H_i^2}\right)^{-1}$
  Set $i \leftarrow i + 1$
end while
Set $K \leftarrow i$
for $k = 0, 1, \ldots, K$ **do**
  Obtain $\widehat{\theta}_k$ from Algorithm 1 with $T = \max\left\{\frac{H_k^2}{\epsilon^2}, 40S^2 \log\frac{K}{\delta}\right\}$
  Set $n \leftarrow \frac{8(S(C_1+1)+SC_2)^2}{\epsilon^2} \log\left(\frac{4K}{\delta}\right)$
  Sample $y_1, \ldots, y_n \sim q_1$ and $(x_1, a_1), \ldots, (x_n, a_n) \sim q_2$
  Set $\widehat{V}_k \leftarrow \frac{1}{n} \sum_{i=1}^{n} \left[\frac{[\mu_0(x_i, a_i) + \Phi_{(x_i,a_i),:}\hat{\theta}_k]_-}{q_1(x,a)} + \frac{\left|(P-B)^\intercal_{:,y_i} \Phi \hat{\theta}_k\right|}{q_2(y_i)} \right]$
end for
Set $\hat{k} \leftarrow \arg\min_k \ell^\intercal \Phi \hat{\theta}_k + H_k \widehat{V}_k + \frac{\beta}{H_k}$
Return policy $\pi_{\widehat{\theta}_{\hat{k}}}$

Figure 2: The Meta-algorithm

$H_1, \ldots, H_K$ of H values. It takes as inputs a bound on the violation function $V_{\max}$, a desired error tolerance ϵ, and desired probability tolerance δ. The algorithm then carefully chooses a grid $H_1, \ldots, H_K$, and, for each $i = 1, \ldots, K$, computes $\hat{\theta}_i$, the output of Algorithm 1 with parameter $H = H_i$, and $\widehat{V}_i$, an approximation to $V_1(\hat{\theta}_i) + V_2(\hat{\theta}_i)$. It then returns $\hat{\theta}_{\hat{k}}$, where

$$\hat{k} \stackrel{\text{def}}{=} \arg\min_k \ell^\intercal \Phi \widehat{\theta}_k + H_k \widehat{V}_k + \frac{\beta}{H_k}.$$

Intuitively, this two-step procedure approximately computes

$$\min_{\theta \in \Theta, H \in \mathbb{R}} \left(\ell^\intercal \mu_\theta + H(V_1(\theta) + V_2(\theta)) + \frac{\beta}{H} \right), \tag{28}$$

which produces a bound that satisfies Definition 2.

Throughout this section, we use the following notation. We define $c(H, \theta) \stackrel{\text{def}}{=} \ell^\intercal \Phi \theta + H(V_1(\theta) + V_2(\theta))$, $\theta_H^* \stackrel{\text{def}}{=} \arg\min_\theta c(H, \theta)$, and $F(H) = c(H, \theta_H^*) + \frac{\beta}{H}$. Hence, the optimization (28) is equal to $\min_{H,\theta} c(H, \theta) + \frac{\beta}{H} = \min_H F(H)$.

3.1 Estimating the Error Functions

To run the Grid Algorithm, we need to be able to estimate the constraint violations $V_1(\theta) + V_2(\theta)$. Similar to the gradient estimate, we estimate $V_1 + V_2$ by importance-weighted sampling. For some n and samples $y_1, \ldots, y_n \sim q_1$ and $(x_1, a_1), \ldots, (x_n, a_n) \sim q_2$, define

$$\widehat{V}_n(\theta) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \frac{[\mu_0(x_i, a_i) + \Phi_{(x_i,a_i),:}\theta]_-}{q_1(x_i, a_i)} + \frac{|(P - B)^\intercal_{:,y_i} \Phi\theta|}{q_2(y_i)}. \tag{29}$$

Since $V_1(\theta) = \sum_{(x,a)} \left|[\mu_0(x, a) + \Phi_{(x,a),:}\theta]_-\right|$ and $V_2(\theta) = \sum_{x'} \left|(P - B)^\intercal_{:,x'} \Phi\theta\right|$, this estimate is clearly unbiased. Also, we earlier assumed the existence of constants $C_1 = \max_{(x,a)\in[\mathcal{X}]\times[\mathcal{A}]} \frac{\left\|\Phi_{(x,a),:}\right\|}{q_1(x,a)}$ and $C_2 = \max_{x\in[\mathcal{X}]} \frac{\left\|(P-B)^\intercal_{:,x}\Phi\right\|}{q_2(x)}$, and so we can bound

$$\frac{[\mu_0(x_i, a_i) + \Phi_{(x_i,a_i),:}\theta]_-}{q_1(x, a)} + \frac{|(P - B)^\intercal_{:,y_i} \Phi\theta|}{q_2(y_i)} \leq S(C_1 + 1) + SC_2$$

which gives us concentration of $\widehat{V}$ around V. In particular, applying Hoeffding's inequality yields:

Lemma 7. *Given $\epsilon > 0$ and $\delta \in [0, 1]$, for any θ, the violation function estimate $\widehat{V}_n(\theta)$ has*

$$\left|\widehat{V}_n(\theta) - (V_1(\theta) + V_2(\theta))\right| \leq \epsilon$$

with probability at least $1 - \delta$ as long as we choose $n \geq \frac{(S(C_1+1)+SC_2)^2}{2\epsilon^2} \log\left(\frac{2}{\delta}\right)$.

3.2 Choosing the Coarseness of the Grid

We wish to construct the sequence H_k such that $\max_{H_k \leq H \leq H_{k+1}} F(H)$ is always $\frac{\epsilon}{2}$, and hence we need control of the smoothness of $F(H)$. Recall that we will choose H to approximately balance the two terms $HV(\theta) + \frac{\beta}{H} \leq HV_{\max} + \frac{\beta}{H}$, and so it suffices to only search for $H \geq \frac{\beta}{\sqrt{V_{\max}}}$. The maximum H will be determined by ϵ.

Lemma 8. *Let $\epsilon > 0$ be some desired error tolerance and $V_{\max}$ be some upper bound on $V_1(\theta)+V_2(\theta)$; we can always take $V_{\max} = 3+S(d+2)$. Consider the H_k sequence defined in Algorithm 2 by the base case $H_0 \stackrel{\text{def}}{=} \beta\left(\sqrt{V_{\max}}\right)^{-1}$, induction step $H_{k+1} \stackrel{\text{def}}{=} H_k + \epsilon\left(V_{\max} + \frac{\beta}{H_k^2}\right)^{-1}$, and terminal condition $K \stackrel{\text{def}}{=} \min\left\{i \in \mathbb{N} : H_i \geq \frac{2\beta}{\epsilon}\right\}$. The grid $H_0, \ldots, H_K$ has the property that*

$$\max_{H,H' \in [H_k, H_{k+1}]} \left|F(H) - F(H')\right| \leq \epsilon. \tag{30}$$

Additionally, we have $K = O(\log(1/\epsilon))$.

Proof. Our first goal is to bound $\max_{H,H' \in [H_i, H_{i+1}]} |F(H) - F(H')|$. We first note that $c(H, \theta^*_H)$, which is a function of H only, is increasing since

$$\begin{aligned} c(H, \theta^*_H) &= \min_\theta \ell^\intercal \Phi \theta + H(V_1(\theta) + V_2(\theta)) \\ &\leq \min_\theta \ell^\intercal \Phi \theta + (H + \delta)(V_1(\theta) + V_2(\theta)) \\ &= c(H + \delta, \theta^*_{H+\delta}). \end{aligned}$$

We also note that $c(H, \theta^*_H)$ is sublinear in H, and indeed

$$\begin{aligned} c(H + \delta, \theta^*_{H+\delta}) &= \min_\theta \ell^\intercal \Phi \theta + (H + \delta)(V_1(\theta) + V_2(\theta)) \\ &\leq \ell^\intercal \Phi \theta^*_H + (H + \delta)(V_1(\theta^*_H) + V_2(\theta^*_H)) \\ &= c(H, \theta^*_H) + \delta(V_1(\theta^*_H) + V_2(\theta^*_H)) \\ &\leq c(H, \theta^*_H) + \delta V_{\max}. \end{aligned}$$

The two observations imply that

$$\max_{H,H' \in [H_i, H_{i+1}]} \left| c(H', \theta^*_{H'}) - c(H, \theta^*_H) \right| \leq c(H_i, \theta^*_{H_i}) + V_{\max} \left(H_{i+1} - H_i \right),$$

and hence we may bound

$$\begin{aligned} \max_{H,H' \in [H_i, H_{i+1}]} \left| F(H) - F(H') \right| &\leq \left| c(H_{i+1}, \theta^*_{H_{i+1}}) - c(H_i, \theta^*_{H_i}) \right| + \beta \max_{H_i \leq H \leq H_{i+1}} \left| \frac{1}{H} - \frac{1}{H'} \right| \\ &\leq (H_{i+1} - H_i) V_{\max} + \beta \left(\frac{1}{H_i} - \frac{1}{H_{i+1}} \right). \end{aligned}$$

We now check that the grid has the property that

$$V_{\max}(H_{i+1} - H_i) + \beta \left(\frac{1}{H_i} - \frac{1}{H_{i+1}} \right) \leq \epsilon$$

for all $i \geq 0$. Defining $\Delta_i = H_{i+1} - H_i$, we see that $\Delta_i = \epsilon \left(V_{\max} + \frac{\beta}{H_i^2} \right)^{-1}$ for all i. The left hand side of the above condition is equal to

$$V_{\max} \Delta + \beta \left(\frac{1}{H_i} - \frac{1}{H_i + \Delta} \right) = \Delta \left(V_{\max} + \frac{\beta}{H_i(H_i + \Delta)} \right) \leq \Delta \left(V_{\max} + \frac{\beta}{H_i^2} \right) = \epsilon,$$

giving us the desired condition.

Lastly, we calculate an upper bound on K, the number of grid points needed. We can write

$$H_{i+1} = H_i \left(1 + \frac{\epsilon}{V_{\max} H_i + \frac{\beta}{H_i}}\right),$$

and using the bounds $H_K \leq \frac{2\beta}{\epsilon}$ and $H_i \geq \beta V_{\max}^{-\frac{1}{2}}$, we have that $V_{\max} H_i + \frac{\beta}{H_i} \leq 2\beta V_{\max}/\epsilon + \sqrt{V_{\max}}$, which implies that

$$H_k \geq H_0 \left(1 + \frac{\epsilon}{2\beta V_{\max}/\epsilon + \sqrt{V_{\max}}}\right)^k.$$

Since we defined $K \stackrel{\text{def}}{=} \min\left\{i \in \mathbb{N} : H_i \geq \frac{2\beta}{\epsilon}\right\}$, we conclude that $K > K'$, where K' is the smallest index such that

$$\begin{aligned} & H_0 \left(1 + \frac{\epsilon}{2\beta V_{\max}/\epsilon + \sqrt{V_{\max}}}\right)^{K'} \geq \frac{2\beta}{\epsilon} \\ \Leftrightarrow \quad & K' \geq \frac{\log\left(\frac{2\sqrt{V_{\max}}}{\epsilon}\right)}{\log\left(1 + \frac{\epsilon}{2\beta V_{\max}/\epsilon + \sqrt{V_{\max}}}\right)}, \end{aligned}$$

leading to the conclusion that $K = O(\log(1/\epsilon))$. □

3.3 Meta-Algorithm Excess Loss Bound

Combining the results from the last two section yields the following theorem.

Theorem 9. *For some $\epsilon > 0$ and $\delta \in [0, 1]$, the Meta-Algorithm specified in Figure 2 has excess loss*

$$\mu_{\widehat{\theta}_T}^\intercal \ell \leq \mu_\theta^\intercal \ell + O\left(\sqrt{V_1(\theta) + V_2(\theta)}\right) + O(\epsilon) \tag{31}$$

with probability at least $1 - \delta$. It requires $O\left(\epsilon^{-4}\right)$ subgradient steps and $O\left(\epsilon^{-2}\log(1/\delta)\right)$ samples to estimate the constraint violations.

In particular, adapting to the optimal H only introduces logarithmic terms to the run time.

4 The Dual ALP for Discounted Cost

We now change settings to discounted cost and try to find a policy with discounted cost almost as low as the best in the class. Most of the tools from the average cost carry over with small modifications, and we will focus on presenting the results in this section with most of the theorem proofs presented in the appendix.

Recall that the LP we intend to approximately solve is

$$\begin{aligned}&\min_{\theta\in\mathbb{R}^d}\ \ell^\intercal\Phi\theta\,,\\ &\text{s.t.}\quad (B-\gamma P)^\intercal\Phi\theta=\alpha,\quad \Phi\theta\geq 0.\end{aligned}$$

This LP has another interpretation. The dual of the approximate dual is

$$\begin{aligned}&\max_{J\in\mathbb{R}^{\mathcal{X}}}\ \alpha^\intercal J\\ &\text{s.t.}\Phi^\intercal\left(\ell+(\gamma P-B)J-z\right)=0,\\ &z\geq 0,\end{aligned}$$

which can be viewed as the original primal with constraint aggregation.

Approximately solving the LP Analogous to V_1 and V_2, we define, relative to a feature matrix Φ, the constraint violation functions

$$\begin{aligned}V_3(\theta)&\stackrel{\text{def}}{=}\|\left[\Phi\theta\right]_-\|_1\quad\text{and}\\ V_4(\theta)&\stackrel{\text{def}}{=}\|(B-\gamma P)^T\Phi\theta-\alpha\|\end{aligned}$$

so that we can approximate the solution of the LP by minimizing the convex surrogate

$$\begin{aligned}c^\gamma(\theta)&\stackrel{\text{def}}{=}\ell^\intercal\Phi\theta+H\left(V_3(\theta)+V_4(\theta)\right)\\ &=\ell^\intercal\Phi\theta+H\|\left[\Phi\theta\right]_-\|_1+H\|(B-\gamma P)^\intercal\Phi\theta-\alpha\|_1\\ &=\ell^\intercal\Phi\theta+H\sum_{(x,a)}\left[\Phi_{(x,a),:}\theta\right]_-+H\sum_{x'}\left|(B-\gamma P)^\intercal_{:,x'}\Phi\theta-\alpha\right|\end{aligned}\tag{32}$$

with some constant H and the constraint set $\Theta=\{\theta:\|\theta\|_2\leq S\}$.

We will minimize (32) through stochastic subgradient descent by sampling $(x_t,a_t)\sim q_3\in\triangle_{\mathcal{X}\times\mathcal{A}}$ and $x'_t\sim q_4\in\triangle_{\mathcal{X}}$ and calculating the unbiased estimator of the subgradient,

$$g^\gamma_t(\theta)=\ell^\intercal\Phi-H\frac{\Phi_{(x_t,a_t),:}}{q_3(x_t,a_t)}\mathbb{I}\{\Phi_{(x_t,a_t),:}\theta<0\}+H\frac{(P-\gamma B)^\intercal_{:,x'_t}\Phi}{q_4(x'_t)}\operatorname{sgn}((P-\gamma B)^\intercal_{:,x'_t}\Phi\theta).\tag{33}$$

The algorithm for the average cost case is exactly the same as Figure 1 with g^γ_t instead of g_t. Recall that we are using the shorthand

$$J_\theta=J_{\pi_{\Phi\theta}}\quad\text{and}\quad\nu_\theta=\nu_{\pi_{\Phi\theta}}.$$

Thus, our objective is to show that $\alpha^\intercal J_{\widehat{\theta}_T}$ is small.

A key difference between the average and discounted cases is the interpretation for the dual variables, μ and ν. In the average case, the feasible μ exactly corresponded to stationary distributions and therefore the average loss was precisely $\ell^\intercal \mu$. However, in the discounted case, the dual variables ν correspond to the expected discounted number of visits to each state and $\ell^\intercal \nu = \alpha^\intercal J$, where J is the value function corresponding to policy π_ν.

4.1 A Excess Loss Bound for a Fixed H

Unlike the average cost case, the discounted cost case does not need a fast mixing assumption. Instead, we assume that the operator 1-norm of Φ is upper bounded by some constant C:

$$\|\Phi\|_1 = \max_{x:\|x\|_1=1} \|\Phi x\|_1 = \max_{1\le j\le d} \sum_{(x,a)} |\Phi_{(x,a),j}| \le C. \tag{34}$$

We also need to assume coverage of the constraint sampling distribution, analogously to the average cost case. We assume existence of constants C_3 and C_4 such that

$$C_3 \ge \max_{(x,a)\in[\mathcal{X}]\times[\mathcal{A}]} \frac{\left\|\Phi_{(x,a),:}\right\|}{q_3(x,a)}, \qquad C_4 \ge \max_{x\in[\mathcal{X}]} \frac{\|(P-\gamma B)^\intercal_{:,x}\Phi\|}{q_4(x)}.$$

Special structure may suggest natural choices of sampling distributions to ensure small C_3 and C_4. For example, if P is sparse with support on only N elements and if there is $C' > 0$ such that for any (x,a) and i, $\Phi_{(x,a),i} \le C'/(\mathcal{X}\mathcal{A})$ and each column of P has only N non-zero elements, we can choose q_3 and q_4 to be uniform distributions and we can bound

$$\frac{\left\|\Phi_{(x,a),:}\right\|}{q_3(x,a)} \le C', \qquad \frac{\|(P-\gamma B)^\intercal_{:,x}\Phi\|}{q_4(x)} \le C'(N+\mathcal{A}).$$

Finally, note that we can always upper bound the constraint violation functions. For any $\theta \in \Theta$,

$$\begin{aligned}
V_3(\theta) &\le \|\Phi\theta\|_1 \le \sum_{j=1}^{d} \sum_{(x,a)} \left|\Phi_{(x,a),j}\right| |\theta_j| \le C\|\theta\|_1 \le C\sqrt{d}\|\theta\|_2 \le \sqrt{d}\, CS, \text{ and} \\
V_4(\theta) &\le \sum_{x'} \left|B^\intercal_{:,x'}(\Phi\theta)\right| + \gamma \sum_{x'} \left|P^\intercal_{:,x'}(\Phi\theta)\right| + \|\alpha\|_1 \\
&\le \sum_{(x,a)} \left(\sum_{x'} B_{(x,a),x'}\right) |(\Phi\theta)_{(x,a)}| + \gamma \sum_{(x,a)} \left(\sum_{x'} P_{(x,a),x'}\right) |(\Phi\theta)_{(x,a)}| + 1 \\
&= (1+\gamma)\|\Phi\theta\|_1 + 1 \\
&\le (1+\gamma)\sqrt{d}\, CS + 1.
\end{aligned}$$

We can combine both statements and obtain

$$V_3(\theta) + V_4(\theta) \leq 1 + \sqrt{d}CS(2+\gamma) \leq 4\sqrt{d}CS, \tag{35}$$

as long as $C \geq$ and $S \geq 1$.

The method we propose for optimizing π_θ in the discounted cost setting is to apply stochastic subgradient descent (from Figure 1) to subgradients $g^\gamma(\theta_t)$ defined in (33). Our algorithm for optimizing discounted cost MDPs is just Figure 1 run with subgradient $g^\gamma(\theta_t)$ (defined in (33)) instead of $g(\theta)$.

We now present the excess loss bound for discounted cost and a fixed H.

Theorem 10. *Consider an expanded efficient large-scale dual ALP problem and some error tolerance $\epsilon > 0$, desired maximum probability of error $\delta > 0$, and parameter $H \geq 1$. Running the stochastic subgradient method (Figure 1 with $g^\gamma(\theta)$) with*

$$T = \frac{S^2}{\epsilon^2}\left(H(C_3+C_4) + \sqrt{d} + 2\sqrt{10\log\frac{1}{\delta}} + 2\sqrt{5d\log\left(1+\frac{S^2T}{d}\right)}\right)^2 \tag{36}$$

and constant learning rate $\eta = S/(G'\sqrt{T})$, where $G' = \sqrt{d} + H(C_3+C_4)$, yields a $\widehat{\theta}_T$ with

$$\ell^\intercal \nu_{\widehat{\theta}_T} \leq \ell^\intercal \nu_\theta + \left(\frac{6}{1-\gamma} + H\right)(V_3(\theta) + V_4(\theta)) + \frac{6\sqrt{d}CS}{H(1-\gamma)} + O(\epsilon).$$

Constants hidden in the big-O notation are polynomials in S, d, C_3, C_4, and C.

Because the proof is very similar to the average cost section, it has been deferred to Section B.

4.2 Error Bound

Previous ADP literature concentrated on showing that the optimal value is well approximated if the feature space contains elements close to the optimum; i.e. $|\alpha^\intercal J_{\widehat{\theta}_T} - \alpha^\intercal J^*|$ was bounded in terms of $\min_\theta \|\Phi\theta - \nu^*\|_1$. Theorem 10 is certainly more general, as it remains non-trivial even if $\min_\theta \|\Phi\theta - \nu^*\|_1$ is large. For completeness, we provide a corollary of this form.

Corollary 11. *Under the same conditions as Theorem 10,*

$$\alpha^\intercal J_{\widehat{\theta}_T} - \alpha^\intercal J^* \leq C_3\left(\frac{1}{1-\gamma} + \frac{1}{\epsilon}\right)\min_\theta \|\Phi\theta - \nu^*\|_1 + C_2\frac{\epsilon}{1-\gamma}. \tag{37}$$

Proof of Corollary 11. Let θ^* be one of the vectors minimizing $\|\Phi\theta - \nu^*\|_1$. Theorem 10 gives

$$\alpha^\intercal J_{\theta_T} - \alpha^\intercal J_{\theta^*} \leq C_1\left(\frac{1}{1-\gamma} + \frac{1}{\epsilon}\right)(V_3(\theta^*) + V_4(\theta^*)) + C_2\frac{\epsilon}{1-\gamma},$$

Since $\nu^* \geq 0$ and by the simple fact that $[x]_- \leq |y - x|$ for any $y \geq 0$, we have

$$V_3(\theta^*) \leq \|\Phi\theta - \nu^*\|_1. \tag{38}$$

For the term $V_4(\theta^*)$, since ν^* is feasible (i.e., $(B - \gamma P)^\intercal \nu^* = \alpha$)

$$\begin{aligned} V_4(\theta^*) &\leq \|(B - \gamma P)^\intercal(\Phi\theta^* - \nu^*)\|_1 + \|(B - \gamma P)^\intercal \nu^* - \alpha\|_1 = \|(B - \gamma P)^\intercal(\Phi\theta^* - \nu^*)\|_1 \\ &\leq \|(B - \gamma P)^\intercal\|_1 \|\Phi\theta - \nu^*\|_1 \leq (\|B^\intercal\|_1 + \gamma\|P^\intercal\|_1) \|\Phi\theta - \nu^*\|_1 \\ &= (1 + \gamma)\|\Phi\theta - \nu^*\|_1, \end{aligned} \tag{39}$$

where $\|\cdot\|_1$ is the matrix operator 1-norm. Therefore, we have,

$$\alpha^\intercal J_{\pi_{[\Phi\widehat{\theta}_T]_+}} - \alpha^\intercal J_{\pi_{[\Phi\theta^*]_+}} \leq C_1 \left(\frac{1}{1-\gamma} + \frac{1}{\epsilon}\right)(2 + \gamma)\|\Phi\theta^* - \nu^*\|_1 + C_2 \frac{\epsilon}{1-\gamma}.$$

Next, we bound $\alpha^\intercal J_{\pi_{[\Phi\theta^*]_+}} - \alpha^\intercal J^*$. Since $\alpha^\intercal J_{\pi_{[\Phi\theta^*]_+}} = \ell^\intercal \nu_{\pi_{[\Phi\theta^*]_+}}$ and $\alpha^\intercal J^* = \ell^\intercal \nu^*$ and by Lemma 13,

$$\begin{aligned} \alpha^\intercal J_{\pi_{[\Phi\theta^*]_+}} - \alpha^\intercal J^* &\leq \|\ell\|_\infty \|\nu_{\pi_{[\Phi\theta^*]_+}} - \nu^*\|_1 \leq \|\nu_{\pi_{[\Phi\theta^*]_+}} - \Phi\theta^*\|_1 + \|\Phi\theta^* - \nu^*\|_1 \\ &\leq \frac{3V_3(\theta^*) + V_4(\theta^*)}{1-\gamma} + \|\Phi\theta^* - \nu^*\|_1 \leq \frac{5}{1-\gamma}\|\Phi\theta^* - \nu^*\|_1. \end{aligned}$$

where the last inequality is due to (38) and (39). The theorem statement follows from combining these two results. □

4.3 The Meta-Algorithm for Discounted Cost

Analogously to the average cost case, setting H correctly yields a excess loss bound of $O\left(\sqrt{V_3(\theta^*) + V_3(\theta^*)}\right) + O(\epsilon)$. The excess loss bound from Theorem 10 suggests that we want H and θ to optimize

$$\ell^\intercal \Phi\theta + \left(\frac{6}{1-\gamma} + H\right)(V_3(\theta) + V_4(\theta)) + \frac{\beta}{H},$$

where we have defined $\beta \stackrel{\text{def}}{=} \frac{6\sqrt{d}CS}{(1-\gamma)}$. The Meta-Algorithm for discounted cost, presented in Figure 3, operates in a manner very similar to the average cost case: a grid $H_1, \ldots, H_K$ is chosen, the corresponding $\widehat{\theta}_k$ are computer, then $\pi_{\widehat{\theta}_{\hat{k}}}$, where

$$\hat{k} \stackrel{\text{def}}{=} \arg\min_k \ell^\intercal \Phi\widehat{\theta}_k + \left(H_k + \frac{1}{1-\gamma}\right)\widehat{V}_k + \frac{\beta}{H_k},$$

is returned. We can prove the following bound for the meta-algorithm.

Theorem 12. *For some $\epsilon > 0$ and $\delta \in [0, 1]$, the Meta-Algorithm for discounted cost (Figure 3 has*

Input: Upper bound $V_{\max}$ on $V_3(\theta) + V_4(\theta)$, error tolerance $\epsilon > 0$, error probability $\delta > 0$, constraint estimation distributions q_3 and q_4

Initialize $H_0 \leftarrow \beta \left(\sqrt{V_{\max}}\right)^{-1}$ and $i \leftarrow 0$

while $H_i \leq \frac{2\beta}{\epsilon}$ **do**

Set $H_{i+1} \leftarrow H_i + \epsilon \left(V_{\max} + \frac{\beta}{H_i^2}\right)^{-1}$

Set $i \leftarrow i + 1$

end while

$K \leftarrow i$

for $k = 0, 1, \ldots, K$ **do**

Obtain $\widehat{\theta}_k$ from Algorithm 1 with $T = O\left(H_k^2 S^2 \log\left(\frac{1}{\delta}\right)\right)$ set by (36)

Set $n \leftarrow \frac{(S(C_3+2C_4))^2}{2\epsilon^2} \log\left(\frac{4K}{\delta}\right)$

Sample $y_1, \ldots, y_n \sim q_3$ and $(x_1, a_1), \ldots, (x_n, a_n) \sim q_4$

Set $\widehat{V}_k \leftarrow \frac{1}{n} \sum_{i=1}^{n} \left[\frac{[\mu_0(x_i, a_i) + \Phi_{(x_i,a_i),:} \hat{\theta}_k]_-}{q_3(x, a)} + \frac{\left|(P - \gamma B)^{\intercal}_{:,y_i} \Phi \hat{\theta}_k\right|}{q_4(y_i)} \right]$

end for

Set $\hat{k} \leftarrow \arg\min_k \ell^{\intercal} \Phi \hat{\theta}_k + \left(H_k + \frac{1}{1-\gamma}\right) \widehat{V}_k + \frac{\beta}{H_k(1-\gamma)}$

Return policy $\pi_{\widehat{\theta}_{\hat{k}}}$

Figure 3: The Meta-algorithm for Discounted Cost

excess loss

$$\ell^{\intercal} \nu_{\theta_{\hat{k}}} \leq \min_{\theta} \ell^{\intercal} \nu_{\theta} + O\left(\sqrt{V_3(\theta) + V_4(\theta)}\right) + O(\epsilon),$$

with probability at least $1 - \delta$. *It requires* $O\left(\epsilon^{-4}\right)$ *subgradient steps and* $O\left(\epsilon^{-2} \log(1/\delta)\right)$ *samples to estimate the constraint violations.*

For the proof and technical details, please see Section C.

5 Experiments

In this section, we apply both algorithms to the four-dimensional discrete-time queuing network illustrated in Figure 4. This network has a relatively long history; see, e.g. Rybko and Stolyar [1992] and more recently de Farias and Van Roy [2003a] (c.f. Section 6.2). There are four queues, $\mu_1, \ldots, \mu_4$, each with state $0, \ldots, B$. Since the cardinality of the state space is $\mathcal{X} = (1 + B)^4$, even a modest B results in huge state-spaces. For time t, let $X_t \in [\mathcal{X}]$ be the state and $s_{i,t} \in \{0, 1\}$, $i = 1, 2, 3, 4$ denote whether queue i is being served. Server 1 only serves queue 1 or 4, server 2 only serves queue 2 or 3, and neither server can idle. Thus, $s_{1,t} + s_{4,t} = 1$ and $s_{2,t} + s_{3,t} = 1$. The dynamics are as follows. At each time t, the following random variables are sampled independently: $A_{1,t} \sim$ Bernoulli(a_1), $A_{3,t} \sim$ Bernoulli(a_3), and $D_{i,t} \sim$ Bernoulli$(d_i s_{i,t})$ for $i = 1, 2, 3, 4$. Using

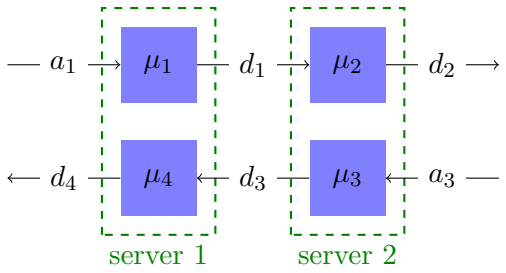

Figure 4: The 4D queuing network. Customers arrive at queue μ_1 or μ_3 then are referred to queue μ_2 or μ_4, respectively. Server 1 can either process queue 1 or 4, and server 2 can only process queue 2 or 3.

$e_1, \ldots, e_4$ to denote the standard basis vectors, the dynamics are:

$$X'_{t+1} = X_t + A_{1,t}e_1 + A_{3,t}e_3 + D_{1,t}(e_2 - e_1) - D_{2,t}e_2 + D_{3,t}(e_4 - e_3) - D_{4,t}e_4,$$

and $X_{t+1} = \max(\mathbf{0}, \min(\mathbf{B}, X'_{t+1}))$ (i.e. all four states are thresholded from below by 0 and above by B). The loss function is the total queue size: $\ell(X_t) = ||X_t||_1$. We compared our method against two common heuristics. In the first, denoted LONGER, each server operates on the queue that is longer with ties broken uniformly at random (e.g. if queue 1 and 4 had the same size, they are equally likely to be served). In the second, denoted LBFS (last buffer first served), the downstream queues always have priority (server 1 will serve queue 4 unless it has length 0, and server 2 will serve queue 2 unless it has length 0). These heuristics are common and have been used as benchmarks for queuing networks (e.g. de Farias and Van Roy [2003a]).

We used $a_1 = a_3 = .08$, $d_1 = d_2 = .12$, and $d_3 = d_4 = .28$, and buffer sizes $B_1 = B_4 = 38$, $B_2 = B_3 = 25$ as the parameters of the network.. The asymmetric size was chosen because server 1 is the bottleneck and tend to have longer queues. The first two features are features of the stationary distributions corresponding to two heuristics. We also included two types of non-stationary-distribution features. For every interval $(0, 5], (6, 10], \ldots, (45, 50]$ and action A, we added a feature ψ with $\phi(x, a) = 1$ if $\ell(x, a)$ is in the interval and $a = A$. To define the second type, consider the three intervals $I_1 = [0, 10]$, $I_2 = [11, 20]$, and $I_3 = [21, 25]$. For every 4-tuple of intervals $(J_1, J_2, J_3, J_4) \in \{I_1, I_2, I_3\}^4$ and action A, we created a feature ψ with $\psi(x, a) = 1$ only if $x_i \in J_i$ and $a = A$. Every feature was normalized to sum to 1. In total, we had 372 features which is about a 10^4 reduction in dimension from the original problem.

We ran our stochastic subgradient descent algorithm with $I = 1000$ sampled constraints and constraint gain $H = 2$. Our learning rate began at 10^{-4} and halved every 2000 iterations. The results of our algorithm are plotted in Figure 5, where $\widehat{\theta}_t$ denotes the running average of θ_t. The left plot is of the LP objective, $\ell^\intercal(\mu_0 + \Phi\widehat{\theta}_t)$. The middle plot is of the sum of the constraint violations, $\left\|[\mu_0 + \Phi\widehat{\theta}_t]_-\right\|_1 + \left\|(P - B)^\intercal \Phi\widehat{\theta}_t\right\|_1$. Thus, $c(\widehat{\theta}_t)$ is a scaled sum of the first two plots. Finally, the

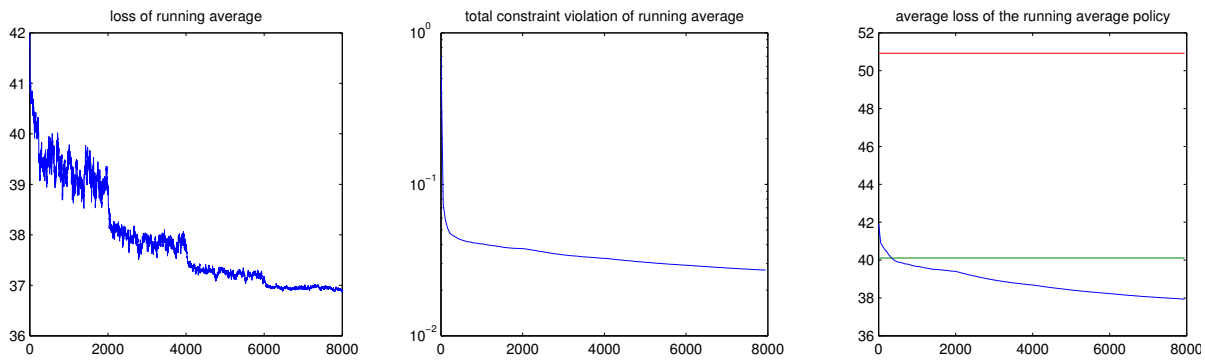

Figure 5: The left plot is of the linear objective of the running average, i.e. $\ell^\intercal \Phi \widehat{\theta}_t$. The center plot is the sum of the two constraint violations of $\widehat{\theta}_t$, and the right plot is $\ell^\intercal \tilde{\mu}_{\widehat{\theta}_t}$ (the average loss of the derived policy). The two horizontal lines correspond to the loss of the two heuristics, LONGER and LBFS.

right plot is of the average losses, $\ell^\intercal \mu_{\widehat{\theta}_t}$ and the two horizontal lines correspond to the loss of the two heuristics, LONGER and LBFS. The right plot demonstrates that, as predicted by our theory, minimizing the surrogate loss $c(\theta)$ does lead to lower average losses.

All previous algorithms (including de Farias and Van Roy [2003a]) work with value functions, while our algorithm works with stationary distributions. Due to this difference, we cannot use the same feature vectors to make a direct comparison. The solution that we find in this different approximating set is slightly worse than the solution of de Farias and Van Roy [2003a].

6 Conclusion

This paper demonstrated the feasibility of solving the MDP planning problem with a parametric policy class based on an approximate dual LP. Unlike previous approaches, we were able to prove *excess loss bounds*, that is, bounds relative to the best policy in our parametric class. We obtained results for both the average cost and discounted cost settings as well as empirical justification.

There are several promising directions. First, are such excess loss bounds possible in the primal formulation?

Another drawback to our methods is that we need a backwards simulator, that is, access to every state with positive probability of transitioning into a state x. Are there alternative formulations that remove this requirement?

References

Y. Abbasi-Yadkori. *Online Learning for Linearly Parametrized Control Problems*. PhD thesis, University of Alberta, 2012.

Ershad Banijamali, Yasin Abbasi-Yadkori, Mohammad Ghavamzadeh, and Nikos Vlassis. Optimizing over a restricted policy class in Markov decision processes. In *AISTATS*, 2019.

R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

D. P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2007.

D. P. Bertsekas and J. Tsitsiklis. *Neuro-Dynamic Programming*. Athena scientific optimization and computation series. Athena Scientific, 1996.

Yichen Chen, Lihong Li, and Mengdi Wang. Scalable bilinear π learning using state and action features. *arXiv preprint arXiv:1804.10328*, 2018.

D. P. de Farias and B. Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51, 2003a.

D. P. de Farias and B. Van Roy. Approximate linear programming for average-cost dynamic programming. In *Advances in Neural Information Processing Systems (NIPS)*, 2003b.

D. P. de Farias and B. Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research*, 29, 2004.

D. P. de Farias and B. Van Roy. A cost-shaping linear program for average-cost approximate dynamic programming with performance guarantees. *Mathematics of Operations Research*, 31, 2006.

V. H. de la Peña, T. L. Lai, and Q-M. Shao. *Self-normalized processes: Limit theory and Statistical Applications*. Springer, 2009.

V. V. Desai, V. F. Farias, and C. C. Moallemi. Approximate dynamic programming via a smoothed linear program. *Operations Research*, 60(3):655–674, 2012.

A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, 2005.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

C. Guestrin, M. Hauskrecht, and B. Kveton. Solving factored mdps with continuous and discrete variables. In *Twentieth Conf. Uncertainty in Artificial Intelligence*, 2004.

M. Hauskrecht and B. Kveton. Linear program approximations to factored continuous-state markov decision processes. In *Advances in Neural Information Processing Systems*, 2003.

R. A. Howard. *Dynamic Programming and Markov Processes*. MIT, 1960.

Chandrashekar Lakshminarayanan, Shalabh Bhatnagar, and Csaba Szepesvári. A linearly relaxed approximate linear program for markov decision processes. *IEEE Transactions on Automatic Control*, 63(4):1185–1191, 2018.

H. R. Maei, Cs. Szepesvári, S. Bhatnagar, D. Precup, D. Silver, and R. S. Sutton. Convergent temporal-difference learning with arbitrary smooth function approximation. In *Advances in Neural Information Processing Systems*, 2009.

H. R. Maei, Cs. Szepesvári, S. Bhatnagar, and R. S. Sutton. Toward off-policy learning control with function approximation. In *Proceedings of the 27th International Conference on Machine Learning*, 2010.

A. S. Manne. Linear programming and sequential decisions. *Management Science*, 6(3):259–267, 1960.

M. Petrik and S. Zilberstein. Constraint relaxation in approximate linear programs. In *Proc. 26th Internat. Conf. Machine Learning (ICML)*, 2009.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1st edition, 1994.

A. N. Rybko and A. L. Stolyar. Ergodicity of stochastic processes describing the operation of open queueing networks. *Problemy Peredachi Informatsii*, 28(3):3–26, 1992.

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

R. S. Sutton, H. R. Maei, D. Precup, S. Bhatnagar, D. Silver, Cs. Szepesvári, and E. Wiewiora. Fast gradient-descent methods for temporal-difference learning with linear function approximation. In *Proceedings of the 26th International Conference on Machine Learning*, 2009a.

R. S. Sutton, Cs. Szepesvári, and H. R. Maei. A convergent O(n) algorithm for off-policy temporal-difference learning with linear function approximation. In *Advances in Neural Information Processing Systems*, 2009b.

M. H. Veatch. Approximate linear programming for average cost mdps. *Mathematics of Operations Research*, 38(3), 2013.

T. Wang, D. Lizotte, M. Bowling, and D. Schuurmans. Dual representations for dynamic programming. *Journal of Machine Learning Research*, pages 1–29, 2008.

Lin Xiao and Tong Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, 2003.

7 Acknowledgments

We gratefully acknowledge the support of the NSF through grant CCF-1115788 and of the ARC through an Australian Research Council Australian Laureate Fellowship (FL110100281).

A Deferred Proofs for Average Cost

Proof of Lemma 3. Let $f = u^\intercal(P - B)$. From $\|u^\intercal(P-B)\|_1 \leq \epsilon''$, we get that for any $x' \in [\mathcal{X}]$,

$$\sum_{(x,a)\in\mathcal{S}} u(x,a)(P-B)_{(x,a),x'} = -\sum_{(x,a)\in\mathcal{N}} u(x,a)(P-B)_{(x,a),x'} + f(x')$$

such that $\sum_{x'} |f(x')| \leq \epsilon''$. Let $h = [u]_+ / \|[u]_+\|_1$. Let $H' = \|h^\intercal(B-P)\|_1$. We write

$$\begin{aligned}
H' &= \sum_{x'} \left| \sum_{(x,a)\in\mathcal{S}} h(x,a)(B-P)_{(x,a),x'} \right| \\
&= \frac{1}{1+\epsilon'} \sum_{x'} \left| \sum_{(x,a)\in\mathcal{S}} u(x,a)(B-P)_{(x,a),x'} \right| \\
&= \frac{1}{1+\epsilon'} \sum_{x'} \left| -\sum_{(x,a)\in\mathcal{N}} u(x,a)(B-P)_{(x,a),x'} + f(x') \right| \\
&\leq \frac{1}{1+\epsilon'} \left(\sum_{x'} \left| -\sum_{(x,a)\in\mathcal{N}} u(x,a)(B-P)_{(x,a),x'} \right| + \sum_{x'} |f(x')| \right) \\
&\leq \frac{1}{1+\epsilon'} \left(\epsilon'' + \sum_{(x,a)\in\mathcal{N}} \sum_{x'} |u(x,a)| \left|(B-P)_{(x,a),x'}\right| \right) \\
&\leq \frac{1}{1+\epsilon'} \left(\epsilon'' + \sum_{(x,a)\in\mathcal{N}} 2\,|u(x,a)| \right) \leq \frac{2\epsilon' + \epsilon''}{1+\epsilon'} \\
&\leq 2\epsilon' + \epsilon'' \,.
\end{aligned}$$

Vector h is almost a stationary distribution in the sense that

$$\|h^\intercal(B-P)\|_1 \le 2\epsilon' + \epsilon'' \,. \tag{40}$$

Let $\|w\|_{1,\mathcal{S}} = \sum_{(x,a)\in\mathcal{S}} |w(x,a)|$. First, we have that

$$\|h-u\|_1 \le \left\|h - \frac{u}{1+\epsilon'}\right\|_1 + \left\|u - \frac{u}{1+\epsilon'}\right\|_{1,\mathcal{S}} \le 2\epsilon' \,.$$

Next we bound $\|\mu_h - h\|_1$. Using $\nu_0 = h$ as the initial state distribution, we will show that as we run policy h (equivalently, policy μ_h), the state distribution converges to μ_h and this vector is close to h. From (40), we have $\mu_0^\intercal P = h^\intercal B + v_0$, where v_0 is such that $\|v_0\|_1 \le 2\epsilon' + \epsilon''$. Let M^h be a $\mathcal{X} \times (\mathcal{X}\mathcal{A})$ matrix that encodes policy h, $M^h_{(i,(i-1)\mathcal{A}+1)\text{-}(i,iA)} = h(\cdot|x_i)$. Other entries of this matrix are zero. We have

$$h^\intercal P M^h = (h^\intercal B + v_0)M^h = h^\intercal B M^h + v_0 M^h = h^\intercal + v_0 M^h \,,$$

where we used the fact that $h^\intercal B M^h = h^\intercal$. Let $\mu_1^\intercal = h^\intercal P M^h$ which is the state-action distribution after running policy h for one step. Let $v_1 = v_0 M^h P = v_0 P^h$ and notice that as $\|v_0\|_1 \le 2\epsilon' + \epsilon''$, we also have that $\|v_1\|_1 = \left\|P^{h\intercal} v_0^\intercal\right\|_1 \le \|v_0\|_1 \le 2\epsilon' + \epsilon''$. Thus,

$$\mu_1^\intercal P = h^\intercal P + v_1 = h^\intercal B + v_0 + v_1 \,.$$

By repeating this argument for k rounds, we obtain

$$\mu_k^\intercal = h^\intercal + (v_0 + v_1 + \cdots + v_{k-1})M^h$$

and it is easy to see that

$$\left\|(v_0 + v_1 + \cdots + v_{k-1})M^h\right\|_1 \le \sum_{i=0}^{k-1} \|v_i\|_1 \le k(2\epsilon' + \epsilon'').$$

Thus, $\|\mu_k - h\|_1 \le k(2\epsilon' + \epsilon'')$. Now, notice that μ_k is the state-action distribution after k rounds of policy μ_h. By the mixing assumption, $\|\mu_k - \mu_h\|_1 \le e^{-k/\tau(h)}$, so the choice of $k = \tau(h)\log(1/\epsilon')$ yields $\|\mu_h - h\|_1 \le \tau(h)\log(1/\epsilon')(2\epsilon' + \epsilon'') + \epsilon'$.

□

Proof of Lemma 6. We prove the lemma by showing that conditions of Theorem 4 are satisfied.

The assumptions allow an easy bound on the subgradient estimate:

$$\|g_t\| \le \|\ell^\intercal \Phi\| + H\frac{\left\|\Phi_{(x_t,a_t),:}\right\|}{q_1(x_t,a_t)} + H\frac{\left\|(P-B)^\intercal_{:,x'_t}\Phi\right\|}{q_2(x'_t)} \le \sqrt{d} + H(C_1+C_2)\,.$$

Also, we show that the subgradient estimate is unbiased:

$$\begin{aligned}
\mathbb{E}\left[g_t(\theta\right] &= \ell^\intercal\Phi - H\sum_{(x,a)} q_1(x,a)\frac{\Phi_{(x,a),:}}{q_1(x,a)}\mathbb{I}\{\mu_0(x,a)+\Phi_{(x,a),:}\theta<0\} \\
&\qquad + H\sum_{x'} q_2(x')\frac{(P-B)^\intercal_{:,x'}\Phi}{q_2(x')}\operatorname{sgn}((P-B)^\intercal_{:,x'}\Phi\theta) \\
&= \ell^\intercal\Phi - H\sum_{(x,a)}\Phi_{(x,a),:}\mathbb{I}\{\mu_0(x,a)+\Phi_{(x,a),:}\theta<0\} + H\sum_{x'}(P-B)^\intercal_{:,x'}\Phi\operatorname{sgn}((P-B)^\intercal_{:,x'}\Phi\theta) \\
&= \nabla_\theta c(\theta)\,.
\end{aligned}$$

The result then follows from Theorem 4 and Remark 5.

It is also convenient to bound the norm of the gradient. If $\mu_0(x,a)+\Phi_{(x,a),:}\theta \ge 0$, then $\nabla_\theta\left|[\mu_0(x,a)+\Phi_{(x,a),:}\theta]_-\right| = 0$. Otherwise, $\nabla_\theta\left|[\mu_0(x,a)+\Phi_{(x,a),:}\theta]_-\right| = -\Phi_{(x,a),:}$. Calculating,

$$\begin{aligned}
\nabla_\theta c(\theta) &= \ell^\intercal\Phi + H\sum_{(x,a)}\nabla_\theta\left|[\mu_0(x,a)+\Phi_{(x,a),:}\theta]_-\right| + H\sum_{x'}\nabla_\theta\left|(P-B)^\intercal_{:,x'}\Phi\theta\right| \\
&= \ell^\intercal\Phi - H\sum_{(x,a)}\Phi_{(x,a),:}\mathbb{I}\{\mu_0(x,a)+\Phi_{(x,a),:}\theta<0\} + H\sum_{x'}(P-B)^\intercal_{:,x'}\Phi\operatorname{sgn}((P-B)^\intercal_{:,x'}\Phi\theta)\,,
\end{aligned} \tag{41}$$

where $\operatorname{sgn}(z) = \mathbb{I}\{z>0\} - \mathbb{I}\{z<0\}$ is the sign function. Let $\pm$ denote the plus or minus sign (the exact sign does not matter here). We have that

$$\|\nabla_\theta c(\theta)\| \le H\sqrt{\sum_{i=1}^d\left(\sum_{x'}\left(\pm\sum_{(x,a)}(P-B)_{(x,a),x'}\Phi_{(x,a),i}\right)\right)^2} + \|\ell^\intercal\Phi\| + H\sqrt{\sum_{i=1}^d\left(\sum_{(x,a)}\left|\Phi_{(x,a),i}\right|\right)^2}\,.$$

Thus,

$$\begin{aligned}
\|\nabla_\theta c(\theta)\| &\le \sqrt{\sum_{i=1}^d(\ell^\intercal\Phi_{:,i})^2} + H\sqrt{d} + H\sqrt{\sum_{i=1}^d\left(\sum_{(x,a)}\left(\pm\sum_{x'}(P-B)_{(x,a),x'}\right)\Phi_{(x,a),i}\right)^2} \\
&\le \sqrt{d} + H\sqrt{d} + H\sqrt{\sum_{i=1}^d\left(2\sum_{(x,a)}\left|\Phi_{(x,a),i}\right|\right)^2} = \sqrt{d}(1+3H)\,,
\end{aligned}$$

where we used $|\ell^\intercal \Phi_{:,i}| \leq \|\ell\|_\infty \|\Phi_{:,i}\|_1 \leq 1$. □

Proof of Theorem 9. By Theorem 2, running Algorithm 1 for a given H_k with $T_k = \max\left\{16\frac{H_k^2}{\epsilon^2}, 160S^2\log\left(\frac{2K}{\delta}\right)\right\}$ produces a $\widehat{\theta}_k$ with

$$c(H_k, \widehat{\theta}_k) \leq c(H_k, \theta_k^*) + H_k V(\theta^*) + \frac{\beta}{H_k} + \frac{\epsilon}{4},$$

where $\theta_k^* = \min_\theta C(H_k, \theta)$, and the probability of error for any single $\widehat{\theta}_k$ is guaranteed to be at most $\frac{\delta}{2K}$. Hence, the union bound implies that the total probability of error of any $\widehat{\theta}_k$ is at most $\frac{\delta}{2}$. Similarly, with our choice of $n = \frac{8(S(C_1+1)+SC_2)^2}{\epsilon^2}\log\left(\frac{4K}{\delta}\right)$, Lemma 7 guarantees that $\left|V_1(\widehat{\theta}_k) + V_2(\widehat{\theta}_k) - \widehat{V}_k\right| \leq \frac{\epsilon}{4}$ holds for all k simultaneously with probability at least $1 - \frac{\delta}{2}$

With these two observations, we can bound the suboptimality of the objective. Recalling that $\hat{k}$ is the minimizer of $\ell^\intercal\Phi\widehat{\theta}_k + H_k\widehat{V}_k + \frac{\beta}{H_k}$, and using k^* as the minimizer of $c(H_k, \theta_k^*) + \frac{\beta}{H_k}$, we have

$$\begin{aligned}
\ell^\intercal\Phi\widehat{\theta}_{\hat{k}} + H_{\hat{k}}\widehat{V}_{\hat{k}} + \frac{\beta}{H_{\hat{k}}} &= \min_k \ell^\intercal\Phi\widehat{\theta}_k + H_k\widehat{V}_k + \frac{\beta}{H_k} \\
&\leq \ell^\intercal\Phi\widehat{\theta}_{k^*} + H_{k^*}\widehat{V}_{k^*} + \frac{\beta}{H_{k^*}} \\
&\leq \ell^\intercal\Phi\widehat{\theta}_{k^*} + H_{k^*}(V_1(\hat{\theta}_{k^*}) + V_2(\hat{\theta}_{k^*})) + \frac{\beta}{H_{k^*}} + \frac{\epsilon}{4} && \text{(Lemma 7)} \\
&= c(H_{k^*}, \widehat{\theta}_{k^*}) + \frac{\beta}{H_{k^*}} + \frac{\epsilon}{4} \\
&\leq c(H_{k^*}, \theta_{k^*}^*) + \frac{\beta}{H_{k^*}} + \frac{\epsilon}{2} \\
&= \min_k c(H_k, \theta_k^*) + \frac{\beta}{H_k} + \frac{\epsilon}{2} \\
&\leq \min_{H,\theta} c(H, \theta) + \frac{\beta}{H} + \epsilon. && \text{(Lemma 8)}
\end{aligned}$$

One final application of the union bound guarantees that the statement holds with probability $1 - (\frac{\delta}{2} + \frac{\delta}{2})$. Hence, the Meta-algorithm minimizes the objective to within ϵ.

We next relate the suboptimality of the objective optimization to the suboptimality of the true loss $\ell^\intercal\mu_{\theta_{\hat{k}}}$. Since all quantities are non-negative, this implies that $\left|\frac{\beta}{H_{\hat{k}}} - \frac{\beta}{H^*}\right| \leq \epsilon$. Finally, we can put together the excess loss bound. To apply Lemma 3 and bound the distance between $\ell^\intercal\Phi\mu_{\widehat{\theta}_{\hat{k}}}$ and $\ell^\intercal\Phi\widehat{\theta}_{\hat{k}}$, we first need to bound $V_1(\widehat{\theta}_{\hat{k}})$ and $V_2(\widehat{\theta}_{\hat{k}})$. Using the bounded suboptimality of $\widehat{\theta}_{\hat{k}}$ as an optimizer of $c(H_{\hat{k}}, \theta)$, we have

$$\begin{aligned}
\ell^\intercal\Phi\widehat{\theta}_{\hat{k}} + H_{\hat{k}}\left(V_1(\widehat{\theta}_{\hat{k}}) + V_2(\widehat{\theta}_{\hat{k}})\right) &\leq \ell^\intercal\Phi\theta_{\hat{k}}^* + H_{\hat{k}}\left(V_1(\theta_{\hat{k}}^*) + V_2(\theta_{\hat{k}}^*)\right) + \frac{\epsilon}{2} \\
&\leq \ell^\intercal\Phi\theta^* + H^*\left(V_1(\theta^*) + V_2(\theta^*)\right) + \epsilon
\end{aligned}$$

and can conclude that

$$\begin{aligned}
V_1(\widehat{\theta}_{\hat{k}}) &\leq \frac{1}{H_{\hat{k}}}\left(2(S+1)+\sqrt{V_1(\theta^*)+V_2(\theta^*)}\right) \\
&\leq \left(\frac{1}{H^*}+\epsilon\right)\left(2(S+1)+\sqrt{V_1(\theta^*)+V_2(\theta^*)}\right) \\
&= (2(S+1)+\epsilon)\sqrt{V_1(\theta^*)+V_2(\theta^*)}+(V_1(\theta^*)+V_2(\theta^*))+2(S+1)\epsilon.
\end{aligned}$$

Completely analogous reasoning gives the same bound on $V_2(\widehat{\theta}_{\hat{k}})$.

Then, applying Lemma 3, we have

$$\begin{aligned}
\ell^\intercal \Phi \mu_{\theta_{\hat{k}}} &\leq \ell^\intercal \Phi \widehat{\theta}_{\hat{k}} + 4\tau(\mu_{\theta_{\hat{k}}})\log(1/\epsilon')\left((2(S+1)+\epsilon)\sqrt{V_1(\theta^*)+V_2(\theta^*)}+(V_1(\theta^*)+V_2(\theta^*))+2(S+1)\epsilon\right) \\
&\leq \ell^\intercal \Phi \widehat{\theta}^* + 4\tau(\mu_{\theta_{\hat{k}}})\log(1/\epsilon')\left((2(S+1)+\epsilon)\sqrt{V_1(\theta^*)+V_2(\theta^*)}+(V_1(\theta^*)+V_2(\theta^*))+2(S+1)\epsilon\right) \\
&\quad + H^*(V_1(\theta^*)+V_2(\theta^*)) + \frac{\beta}{H^*} + \epsilon \\
&\leq \ell^\intercal \mu_{\theta^*} + 4\tau(\mu_{\theta_{\hat{k}}})\log(1/\epsilon')\left((2(S+1)+\epsilon)\sqrt{V_1(\theta^*)+V_2(\theta^*)}+(V_1(\theta^*)+V_2(\theta^*))+2(S+1)\epsilon\right) \\
&\quad + H^*(V_1(\theta^*)+V_2(\theta^*)) + \frac{\beta}{H^*} + \epsilon + (V_1(\theta^*)+V_2(\theta^*)).
\end{aligned}$$

Plugging in $H^* = \left(\sqrt{V_1(\theta)+V_2(\theta)}\right)^{-1}$ produces

$$\ell^\intercal \mu_{\theta_{\hat{k}}} \leq \min_\theta \ell^\intercal \mu_\theta + O\left(\sqrt{V_1(\theta)+V_2(\theta)}\right) + O\left(V_1(\theta)+V_2(\theta)\right) + O(\epsilon).$$

The theorem statement follows by recalling that $V_1(\theta)+V_2(\theta) \leq 1$.

Let us turn to the complexity. The total number of subgradient descent steps is bounded by

$$KT_K = 16\frac{2\beta^2}{\epsilon^4}\frac{\log\left(\frac{2\sqrt{V_{\max}}}{\epsilon}\right)}{\log\left(1+\frac{\epsilon}{2\beta V_{\max}/\epsilon+\sqrt{V_{\max}}}\right)} = O\left(\epsilon^{-4}\right)$$

and the total number of samples needed to estimate the violation function is

$$nK = \frac{8(S(C_1+1)+SC_2)^2}{\epsilon^2}\log\left(\frac{4K}{\delta}\right)\frac{\log\left(\frac{2\sqrt{V_{\max}}}{\epsilon}\right)}{\log\left(1+\frac{\epsilon}{2\beta V_{\max}/\epsilon+\sqrt{V_{\max}}}\right)} = O\left(\epsilon^{-2}\log(1/\delta)\right).$$

□

B Discounted Cost Excess Loss Analysis

This section presents the necessary technical tools and the proof of Theorem 10. We begin by showing that if some vector ν is close to a feasible point of the LP, then it almost equals the expected frequencies of visits of the policy π_ν (when the system runs under the policy π_h with the initial distribution α), i.e.,

$$\nu_{\pi_\nu}(x,a) = \sum_{x'} \alpha(x') \sum_{t=1}^{\infty} \gamma^{t-1} P^{\pi_h}\left(x_t = x, a_t = a | x_1 = x'\right). \tag{42}$$

Lemma 13. *For any vector $\nu \in \mathbb{R}^{\mathcal{X}\mathcal{A}}$, let $\mathcal{N}$ be the set of points (x,a) where $\nu(x,a) \leq 0$ and $\mathcal{S} = \mathcal{N}^c$ and define the constants $\sum_{(x,a)\in\mathcal{N}} |\nu(x,a)| = \epsilon'$ and $\|(B-\gamma P)^\intercal \nu - \alpha\|_1 = \epsilon''$. Further assume that for each x, there exists an a such that $(x,a) \in \mathcal{S}$. Then, for the policy π_ν define by*

$$\pi_\nu(a|x) = \frac{[\nu(x,a)]_+}{\sum_{a'}[\nu(x,a')]_+}, \tag{43}$$

the expected frequencies of visits under the policy is close to ν:

$$\|\nu_{\pi_\nu} - \nu\|_1 \leq \frac{3\epsilon' + \epsilon''}{1-\gamma}.$$

Proof. First, we notice that,

$$\|\,[\nu]_+ - \nu\|_1 \leq \sum_{(x,a)\in\mathcal{N}} |\nu(x,a)|_1 = \epsilon'. \tag{44}$$

Let $\xi = (B-\gamma P)^\intercal \nu - \alpha \in \mathbb{R}^{\mathcal{X}}$ with $\|\xi\|_1 = \epsilon''$ according to the assumption. For any $x' \in [\mathcal{X}]$, we have,

$$\sum_{(x,a)\in\mathcal{S}} \nu(x,a)(B-\gamma P)_{(x,a),x'} - \alpha(x') = -\sum_{(x,a)\in\mathcal{N}} \nu(x,a)(B-\gamma P)_{(x,a),x'} + \xi(x').$$

Let $v_0 = (B - \gamma P)^\intercal h - \alpha$, we have

$$\begin{aligned}
\|v_0\|_1 &= \sum_{x'} \left| \sum_{(x,a)} h(x,a)(B-\gamma P)_{(x,a),x'} - \alpha(x') \right| \\
&= \sum_{x'} \left| \sum_{(x,a)\in\mathcal{S}} \nu(x,a)(B-\gamma P)_{(x,a),x'} - \alpha(x') \right| \\
&= \sum_{x'} \left| - \sum_{(x,a)\in\mathcal{N}} \nu(x,a)(B-\gamma P)_{(x,a),x'} + \xi(x') \right| \qquad (45)
\end{aligned}$$

with the upper bound

$$\begin{aligned}
\|v_0\|_1 &\le \sum_{x'} \left| - \sum_{(x,a)\in\mathcal{N}} \nu(x,a)(B-\gamma P)_{(x,a),x'} \right| + \|\xi\|_1 \\
&\le \sum_{(x,a)\in\mathcal{N}} \left(|\nu(x,a)| \sum_{x'} \left| (B-\gamma P)_{(x,a),x'} \right| \right) + \epsilon'' \\
&\le 2 \sum_{(x,a)\in\mathcal{N}} |\nu(x,a)| + \epsilon'' \\
&\le 2\epsilon' + \epsilon''. \qquad (46)
\end{aligned}$$

Let M^h be a $\mathcal{X} \times (\mathcal{X}\mathcal{A})$ matrix that encodes the policy π_ν, where $M^h_{(i,(i-1)\mathcal{A}+1)-(i,iA)} = \pi_\nu(\cdot|x_i)$. As a concrete example with state space $\{x_1, x_2\}$ and action space $\{a_1, a_2\}$, we have

$$M^h = \begin{pmatrix} \pi_\nu(a_1|x_1) & \pi_\nu(a_2|x_1) & 0 & 0 \\ 0 & 0 & \pi_\nu(a_1|x_2) & \pi_\nu(a_2|x_2) \end{pmatrix}.$$

By the definition of π_ν in (43), it is easy to check that $h^\intercal B M^h = h^\intercal$.

With M^h, the ν_{π_h} defined in (42) can be written as,

$$\nu_{\pi_h}^\intercal = \sum_{t=1}^{\infty} \gamma^{t-1} \alpha^\intercal M^h (P M^h)^{t-1} \qquad (47)$$

Now, we are ready to bound $\|\nu_{\pi_\nu} - \nu\|_1$. By the definition of v_0 (i.e., $v_0 = (B - \gamma P)^\intercal h - \alpha$), we have,

$$\alpha^\intercal M^h = h^\intercal B M^h - \gamma h^\intercal P M^h - v_0^\intercal M^h = h^\intercal - \gamma h^\intercal P M^h - v_0^\intercal M^h,$$

where the last equality is due to $h^\intercal BM^h = h^\intercal$. Therefore,

$$\alpha^\intercal M^h (PM)^{t-1} = h^\intercal (PM^h)^{t-1} - \gamma h^\intercal (PM^h)^t - v_0^\intercal M^h (PM)^{t-1},$$

By (47), we have,

$$\nu_{\pi_h}^\intercal = h^\intercal - \sum_{t=1}^{\infty} \gamma^{t-1} v_0^\intercal M_h (PM^h)^{t-1}. \tag{48}$$

Let $z_t = v_0^\intercal M_h (PM^h)^t$. By (46), we have

$$\|z_0\| = \|v_0^\intercal M_h\|_1 = \sum_{x,a} |v_0(x)\pi_\nu(a|x)| \leq \sum_x \left(|v_0(x)| \sum_a |\pi_\nu(a|x)| \right) = \|v_0\|_1 \leq 2\epsilon' + \epsilon''.$$

Further,

$$\begin{aligned} \|z_{t+1}\|_1 = \|z_t PM^h\|_1 &= \sum_{x,a} \sum_{x',a'} \left| z_t(x',a') P(x|x',a') \pi_\nu(a|x) \right| \\ &\leq \sum_{x,a} \left(\left| z_t(x',a') \right| \sum_{x',a'} \left| P_{\pi_\nu}(x,a|x',a') \right| \right) = \|z_t\|_1. \end{aligned}$$

By the induction, we know that $\|z_t\|_1 \leq 2\epsilon' + \epsilon''$ for all t. By (48),

$$\|\nu_{\pi_h} - h\|_1 \leq \sum_{t=1}^{\infty} \gamma^{t-1} \|z_{t-1}\|_1 \leq \frac{2\epsilon' + \epsilon''}{1-\gamma}. \tag{49}$$

Combining this with (44) and the triangle inequality,

$$\|\nu_{\pi_h} - \nu\|_1 \leq \frac{2\epsilon' + \epsilon''}{1-\gamma} + \epsilon' \leq \frac{3\epsilon' + \epsilon''}{1-\gamma}. \tag{50}$$

□

Next, we need the analog of Lemma 6 for the discounted case, which is again a direct application of Theorem 4.

Lemma 14. *Given some error tolerance $\epsilon > 0$ and desired maximum probability of error $\delta > 0$, running the stochastic subgradient method (shown in Figure 1) on $c^\gamma(\theta)$ with $T \geq 1/\epsilon^4$, $H = 1/\epsilon$, and constant learning rate $\eta = \frac{S}{\sqrt{T}} \left(\sqrt{d} + H(C_3 + C_4) \right)$ produces a $\widehat{\theta}_T$ such that, with probability*

at least $1-\delta$,

$$c^\gamma(\widehat{\theta}_T) - \min_{\theta\in\Theta} c^\gamma(\theta) \le S\frac{\sqrt{d}+H(C_3+C_4)}{\sqrt{T}} + \sqrt{\frac{1+4S^2T}{T^2}\left(2\log\frac{1}{\delta} + d\log\left(1+\frac{S^2T}{d}\right)\right)}\,. \tag{51}$$

Proof. We (once again) prove the lemma by showing that conditions of Theorem 4 are satisfied. First, the subgradient norms have the easy bound

$$\|g_t^\gamma\| \le \|\ell^\intercal\Phi\| + H\frac{\left\|\Phi_{(x_t,a_t),:}\right\|}{q_3(x_t,a_t)} + H\frac{\left\|(P-\gamma B)^\intercal_{:,x'_t}\Phi\right\|}{q_4(x'_t)} \le \sqrt{d} + H(C_3+C_4)\,.$$

Finally, we show that the subgradient estimate is unbiased:

$$\begin{aligned}
\mathbb{E}\left[g_t^\gamma(\theta)\right] &= \ell^\intercal\Phi - H\sum_{(x,a)} q_3(x,a)\frac{\Phi_{(x,a),:}}{q_3(x,a)}\mathbb{I}\{\mu_0(x,a)+\Phi_{(x,a),:}\theta<0\} \\
&\qquad\qquad + H\sum_{x'} q_4(x')\frac{(P-\gamma B)^\intercal_{:,x'}\Phi}{q_4(x')}\operatorname{sgn}((P-\gamma B)^\intercal_{:,x'}\Phi\theta) \\
&= \ell^\intercal\Phi - H\sum_{(x,a)}\Phi_{(x,a),:}\mathbb{I}\{\mu_0(x,a)+\Phi_{(x,a),:}\theta<0\} + H\sum_{x'}(P-\gamma B)^\intercal_{:,x'}\Phi\operatorname{sgn}((P-\gamma B)^\intercal_{:,x'}\Phi\theta) \\
&= \nabla_\theta c^\gamma(\theta)\,.
\end{aligned}$$

□

With this lemma in hand, the proof of Theorem 10] proceeds in much the same way as the proof of Theorem 2].

Proof of Theorem 10. Recall that the convex surrogate for the discounted cost is

$$c^\gamma(\theta) = \ell^\intercal\Phi\theta + H\|\left[\Phi\theta\right]_-\|_1 + H\|(B-\gamma P)^\intercal\Phi\theta - \alpha\|_1,$$

with the constraint set $\Theta = \{\theta : \|\theta\|_2 \le S\}$.

Now, obtain $\widehat{\theta}_T$ from the stochastic subgradient descent algorithm. By Lemma 14, the error bound must be less than

$$b_T = \frac{S}{\sqrt{T}}\left((\sqrt{d}+H(C_3+C_4)) + 2\sqrt{10\log\frac{1}{\delta}} + 2\sqrt{5d\log\left(1+\frac{S^2T}{d}\right)}\right) + O\left(\frac{1}{T}\right).$$

Then with high probability, we have for any $\theta\in\Theta$,

$$\ell^\intercal\Phi\widehat{\theta}_T + H\,V_3(\widehat{\theta}_T) + H\,V_4(\widehat{\theta}_T) \le \ell^\intercal\Phi\theta + H\,V_3(\theta) + H\,V_4(\theta) + b_T\,.$$

Since we can bound

$$\ell^\intercal \Phi\theta \le \|\ell\|_\infty \|\Phi\theta\|_1 \le \sqrt{d}\, CS,$$

rearranging Equation (B) yields

$$V_3(\widehat{\theta}_T) \le \frac{1}{H}\left(2\sqrt{d}\, CS + H\, V_3(\theta) + H\, V_4(\theta) + b_T\right) \stackrel{\text{def}}{=} \epsilon' \text{, and}$$
$$V_4(\widehat{\theta}_T) \le \frac{1}{H}\left(2\sqrt{d}\, CS + H\, V_3(\theta) + H\, V_4(\theta) + b_T\right) \stackrel{\text{def}}{=} \epsilon'' .$$

Using these bounds on $V_3(\widehat{\theta}_T)$ and $V_3(\widehat{\theta}_T)$ with Lemma 13 gives

$$\left|\ell^\intercal \nu_{\widehat{\theta}_T} - \ell^\intercal \Phi\widehat{\theta}_T\right| \le \|\nu_{\widehat{\theta}_T} - \Phi\widehat{\theta}_T\|_1 \le \frac{3\epsilon' + \epsilon''}{1-\gamma}.$$

Lemma 13, applied to ν_θ, implies that

$$|\ell^\intercal \nu_\theta - \ell^\intercal \Phi\theta| \le \|\nu_\theta - \Phi\theta\|_1 \le \frac{3V_3(\theta) + V_4(\theta)}{1-\gamma},$$

and so

$$\begin{aligned}
\ell^\intercal \nu_{\widehat{\theta}_T} &\le \ell^\intercal \Phi\widehat{\theta}_T + \frac{3\epsilon' + \epsilon''}{1-\gamma} \\
&\le \ell^\intercal \Phi\theta + H\, V_3(\theta) + H\, V_4(\theta) + b_T + \frac{3\epsilon' + \epsilon''}{1-\gamma} \\
&\le \ell^\intercal \nu_\theta + \frac{3V_3(\theta) + V_4(\theta)}{1-\gamma} + H\, V_3(\theta) + H\, V_4(\theta) + b_T + \frac{3\epsilon' + \epsilon''}{1-\gamma} .
\end{aligned}$$

First, we simplify

$$\begin{aligned}
\frac{3\epsilon' + \epsilon''}{1-\gamma} &= \frac{3}{H(1-\gamma)}\left(2\sqrt{d}CS + HV_3(\theta) + HV_4(\theta) + b_T\right) \\
&= \frac{3}{(1-\gamma)}\left(V_3(\theta) + V_4(\theta)\right) + \frac{3}{H(1-\gamma)} 2\sqrt{d}CS + \frac{4S(\sqrt{d} + C_3 + C_4)}{\sqrt{T}H(1-\gamma)} \\
&\quad + \frac{3S}{\sqrt{T}H(1-\gamma)} 2\sqrt{10\log\frac{1}{\delta}} + \frac{3S}{\sqrt{T}H(1-\gamma)} 2\sqrt{5d\log\left(1 + \frac{S^2T}{d}\right)} + O\left(\frac{1}{T^{3/2}(1-\gamma)H}\right) \\
&= \frac{3}{(1-\gamma)}\left(V_3(\theta) + V_4(\theta)\right) + \frac{6}{H(1-\gamma)}\sqrt{d}CS + O\left(\frac{\log(T)}{(1-\gamma)H\sqrt{T}}\right).
\end{aligned}$$

Plugging in this expression and b_T, we have

$$\begin{aligned}
\ell^\intercal \nu_{\widehat{\theta}_T} &\leq \ell^\intercal \nu_\theta + \left(\frac{6}{1-\gamma} + H\right)(V_3(\theta) + V_4(\theta)) + \frac{6\sqrt{d}CS}{H(1-\gamma)} + O\left(\frac{\log(T)}{(1-\gamma)H\sqrt{T}}\right) + b_T \\
&\leq \ell^\intercal \nu_\theta + \left(\frac{6}{1-\gamma} + H\right)(V_3(\theta) + V_4(\theta)) + \frac{6\sqrt{d}CS}{H(1-\gamma)} + \frac{S}{\sqrt{T}} H(C_3 + C_4) \\
&\quad + \frac{S}{\sqrt{T}}\left(C_3 + C_4 + \sqrt{d} + 2\sqrt{10\log\frac{1}{\delta}} + 2\sqrt{5d\log\left(1 + \frac{S^2 T}{d}\right)}\right) \\
&\quad + O\left(\frac{\log(T)}{(1-\gamma)H\sqrt{T}}\right) + O\left(\frac{1}{T}\right).
\end{aligned}$$

Thus, setting T such that

$$T \geq \frac{S^2}{\epsilon^2}\left(H(C_3 + C_4) + \sqrt{d} + 2\sqrt{10\log\frac{1}{\delta}} + 2\sqrt{5d\log\left(1 + \frac{S^2 T}{d}\right)}\right)^2$$

or, more compactly, $T = O\left(S^2 \log\left(\frac{1}{\delta}\right)\frac{H^2}{\epsilon^2}\right)$, yields

$$\ell^\intercal \nu_{\widehat{\theta}_T} \leq \ell^\intercal \nu_\theta + \left(\frac{6}{1-\gamma} + H\right)(V_3(\theta) + V_4(\theta)) + \frac{6\sqrt{d}CS}{H(1-\gamma)} + O(\epsilon)$$

where, as usual, the O hides log factors. This statement holds with probability at least $1-\delta$ and for any $\theta \in \Theta$.

□

C Analysis of the Discounted Cost Meta-Algorithm

It is important to note that the optimum H^* need never be smaller than $\beta/\sqrt{V_{\max}}$, where $V_{\max}$ is some bound on $V_3(\theta^*) + V_4(\theta^*)$. Even though we cannot compute this quantity, we may still restrict the domain of H to

$$H \geq \min_\theta 1/\sqrt{V_3(\theta) + V_4(\theta)} \geq \left(1 + \sqrt{d}CS(2+\gamma)\right)^{-\frac{1}{2}} \geq \left(4\sqrt{d}CS\right)^{-\frac{1}{2}}.$$

where the bound on $V_3(\theta) + V_4(\theta)$ is taken from (35).

For convenience, we will overload the notation from the average cost analysis. Define

$$c(H, \theta) \stackrel{\text{def}}{=} \ell^\intercal \Phi\theta + \left(H + \frac{6}{1-\gamma}\right)(V_3(\theta) + V_4(\theta)),$$

where $\theta^*_H \stackrel{\text{def}}{=} \arg\min_\theta c(H, \theta)$, and $F(H) = c(H, \theta^*_H) + \frac{\beta}{H}$. The *meta-algorithm for discounted cost*

takes as inputs a bound on the violation function $V_{\max}$, discount factor γ, an error tolerance ϵ, and desired probability tolerance δ. The algorithm then carefully chooses a grid $H_1, \ldots, H_K$, computes the corresponding $\widehat{\theta}_k$, then returns $\pi_{\widehat{\theta}_{\hat{k}}}$ where

$$\hat{k} \stackrel{\text{def}}{=} \arg\min_k \ell^\intercal \Phi \widehat{\theta}_k + \left(H_k + \frac{1}{1-\gamma}\right) \widehat{V}_k + \frac{\beta}{H_k}.$$

C.1 Estimating the Violation Functions

Given some θ, we can estimate the violation function $V_3(\theta) + V_4(\theta)$ in much the same way as the average cost case. For some n and samples $y_1, \ldots, y_n \sim q_3$ and $(x_1, a_1), \ldots, (x_n, a_n) \sim q_4$, define

$$\widehat{V}_n(\theta) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \frac{[\Phi_{(x_i,a_i),:}\theta]_-}{q_3(x,a)} + \frac{|(B - \gamma P)^\intercal_{:,y_i} \Phi\theta - \alpha|}{q_4(y_i)}. \tag{52}$$

Since $V_3(\theta) = \sum_{(x,a)} \left|[\Phi_{(x,a),:}\theta]_-\right|$ and $V_4(\theta) = \sum_{x'} \left|(B - \gamma P)^\intercal_{:,x'} \Phi\theta - \alpha\right|$, this estimate is clearly unbiased. Also, we earlier assumed the existence of constants

$$C_3 = \max_{(x,a) \in [\mathcal{X}] \times [\mathcal{A}]} \frac{\left\|\Phi_{(x,a),:}\right\|}{q_3(x,a)}, \qquad C_4 = \max_{x \in [\mathcal{X}]} \frac{\|(P - \gamma B)^\intercal_{:,x} \Phi\|}{q_4(x)},$$

and so we can bound

$$\frac{[\Phi_{(x_i,a_i),:}\theta]_-}{q_3(x,a)} + \frac{|(B - \gamma P)^\intercal_{:,y_i} \Phi\theta - \alpha|}{q_4(y_i)} \leq S(C_3 + 2C_4).$$

Therefore, we have concentration of $\widehat{V}$ around V. The analogous result to Lemma 7 (also using Hoeffding's inequality) is the following.

Lemma 15. *Given $\epsilon > 0$ and $\delta \in [0, 1]$, for any θ, the violation function estimate $\widehat{V}_n(\theta)$ has*

$$\left|\widehat{V}_n(\theta) - (V_3(\theta) + V_4(\theta))\right| \leq \epsilon$$

with probability at least $1 - \delta$ as long as we choose $n \geq \frac{(S(C_3+2C_4))^2}{2\epsilon^2} \log\left(\frac{2}{\delta}\right)$.

C.2 Defining the Grid

As before, let $\epsilon > 0$ be some desired error tolerance and $V_{\max}$ be some upper bound on $V_3(\theta) + V_4(\theta)$; we can always take $V_{\max} = 4\sqrt{d}CS$. As we shall see, the H_k sequence can be taken to be identical to the average cost case as long an appropriate β and $V_{\max}$ are used. Recall that H is chosen to approximately minimize $\left(H + \frac{1}{\gamma}\right) V(\theta) + \frac{\beta}{H} \leq \left(H + \frac{1}{\gamma}\right) V_{\max} + \frac{\beta}{H}$, and so limiting H to $H \leq \frac{\beta}{\sqrt{V_{\max}}}$ suffices in the discounted case as well.

Lemma 16. *Let $\epsilon > 0$ be some desired error tolerance and $V_{\max}$ be some upper bound on $V_3(\theta) + V_4(\theta)$; we can always take $V_{\max} = 3 + S(d+2)$. Consider the H_k sequence defined in Algorithm 2 by the base case $H_0 \stackrel{\text{def}}{=} \beta \left(\sqrt{V_{\max}}\right)^{-1}$, induction step $H_{k+1} \stackrel{\text{def}}{=} H_k + \epsilon \left(V_{\max} + \frac{\beta}{H_k^2}\right)^{-1}$, and terminal condition $K \stackrel{\text{def}}{=} \min\left\{i \in \mathbb{N} : H_i \geq \frac{2\beta}{\epsilon}\right\}$. The grid $H_0, \ldots, H_K$ has the property that*

$$\max_{H,H' \in [H_k, H_{k+1}]} \left|F(H) - F(H')\right| \leq \epsilon. \tag{53}$$

Additionally, we have $K = O(\log(1/\epsilon))$.

Proof. Our first goal is to bound $\max_{H,H' \in [H_i, H_{i+1}]} |F(H) - F(H')|$. We first note that $c(H, \theta_H^*)$, which is a function of H only, is increasing since

$$\begin{aligned}
c(H, \theta_H^*) &= \min_\theta \ell^\intercal \Phi \theta + \left(H + \frac{1}{1-\gamma}\right)(V_3(\theta) + V_4(\theta)) \\
&\leq \min_\theta \ell^\intercal \Phi \theta + \left(H + \frac{1}{1-\gamma} + \delta\right)(V_3(\theta) + V_4(\theta)) \\
&= c(H + \delta, \theta_{H+\delta}^*).
\end{aligned}$$

We also note that $c(H, \theta_H^*)$ is sublinear in H, and indeed

$$\begin{aligned}
c(H + \delta, \theta_{H+\delta}^*) &= \min_\theta \ell^\intercal \Phi \theta + \left(H + \frac{1}{1-\gamma} + \delta\right)(V_3(\theta) + V_4(\theta)) \\
&\leq \ell^\intercal \Phi \theta_H^* + \left(H + \frac{1}{1-\gamma} + \delta\right)(V_3(\theta_H^*) + V_4(\theta_H^*)) \\
&= c(H, \theta_H^*) + \delta(V_3(\theta_H^*) + V_4(\theta_H^*)) \\
&\leq c(H, \theta_H^*) + \delta V_{\max}.
\end{aligned}$$

The two observations imply that

$$\max_{H,H' \in [H_i, H_{i+1}]} \left|c(H', \theta_{H'}^*) - c(H, \theta_H^*)\right| \leq c(H_i, \theta_{H_i}^*) + V_{\max}\left(H_{i+1} - H_i\right),$$

and hence we may bound

$$\begin{aligned}
\max_{H,H' \in [H_i, H_{i+1}]} \left|F(H) - F(H')\right| &\leq \left|c(H_{i+1}, \theta_{H_{i+1}}^*) - c(H_i, \theta_{H_i}^*)\right| + \beta \max_{H_i \leq H \leq H_{i+1}} \left|\frac{1}{H} - \frac{1}{H'}\right| \\
&\leq (H_{i+1} - H_i) V_{\max} + \beta\left(\frac{1}{H_i} - \frac{1}{H_{i+1}}\right),
\end{aligned}$$

which is exactly the same bound as in the average cost case. Therefore, the same analysis shows

that

$$V_{\max}(H_{i+1} - H_i) + \beta\left(\frac{1}{H_i} - \frac{1}{H_{i+1}}\right) \leq \epsilon.$$

for all $i \geq 0$ and that we may bound

$$K \leq \frac{\log\left(\frac{2\sqrt{V_{\max}}}{\epsilon}\right)}{\log\left(1 + \frac{\epsilon}{2\beta V_{\max}/\epsilon + \sqrt{V_{\max}}}\right)},$$

leading to the conclusion that $K = O(\log(1/\epsilon))$.

□

Proof of Theorem 12. Running the discounted SGD Algorithm (Figure 1 with subgradient $g^\gamma(\theta)$) for H_k

$H_1, \ldots, H_K$ with $4T$ steps, where T is set as in Theorem 10, produces a sequence $\widehat{\theta}_1, \ldots, \widehat{\theta}_K$ such that

$$c(H_k, \widehat{\theta}_K) \leq c(H_k, \theta^*_K) + H_k V(\theta^*) + \frac{\beta}{H_k} + \frac{\epsilon}{4}$$

holds for all k simultaneously with probability at least $1 - \frac{\delta}{2}$, which is easily argued by noting that the probability of error for any single k is δ/K and applying the union bound.

Lemma 15, along with our choice of

$$n \geq \frac{(S(C_3 + 2C_4))^2}{2\epsilon^2} \log\left(\frac{4K}{\delta}\right)$$

guarantees that $\left|V_3(\widehat{\theta}_k) + V_4(\widehat{\theta}_k) - \widehat{V}_k\right| \leq \frac{\epsilon}{4}$ holds with probability at least $1 - \frac{\delta}{2K}$, and hence the statement holds for all $\widehat{V}_k$ with probability at most $1 - \frac{\delta}{2}$.

We now turn to bounding the suboptimality of the objective. Recalling that $\hat{k}$ is the minimizer of $\ell^\intercal \Phi \widehat{\theta}_k + \left(H_k + \frac{1}{1-\gamma}\right)\widehat{V}_k + \frac{\beta}{H_k}$, and using k^* as the minimizer of $c(H_k, \theta^*_k) + \frac{\beta}{H_k}$, we have

$$\begin{aligned}
\ell^\intercal \Phi \widehat{\theta}_{\hat{k}} + \left(H_{\hat{k}} + \frac{1}{1-\gamma}\right)\widehat{V}_{\hat{k}} + \frac{\beta}{H_{\hat{k}}} &= \min_k \ell^\intercal \Phi \widehat{\theta}_k + \left(H_k + \frac{1}{1-\gamma}\right)\widehat{V}_k + \frac{\beta}{H_k} \\
&\leq \ell^\intercal \Phi \widehat{\theta}_{k^*} + \left(H_{k^*} + \frac{1}{1-\gamma}\right)\widehat{V}_{k^*} + \frac{\beta}{H_{k^*}} \\
&\leq c(H_{k^*}, \widehat{\theta}_{k^*}) + \frac{\beta}{H_{k^*}} + \frac{\epsilon}{4} && \text{(Lemma 15)} \\
&\leq c(H_{k^*}, \theta^*_{k^*}) + \frac{\beta}{H_{k^*}} + \frac{\epsilon}{2} && \text{(Theorem 10)} \\
&= \min_k c(H_k, \theta^*_k) + \frac{\beta}{H_k} + \frac{\epsilon}{2} \\
&\leq \min_{H,\theta} c(H, \theta) + \frac{\beta}{H} + \epsilon && \text{(Lemma 16).}
\end{aligned}$$

The statement holds with probability at least $\frac{\delta}{2} + \frac{\delta}{2}$, where the first term is from estimating $\widehat{V}_k$ (Lemma 15) and the second term is from bounding the SGD error (Theorem 10). Hence, the Meta-algorithm minimizes the objective to within ϵ.

Next, we use Lemma 13 to bound the discrepancy between $\Phi\theta$ and ν_θ. Therefore, we need to bound $V_3(\widehat{\theta}_{\hat{k}})$ and $V_4(\widehat{\theta}_{\hat{k}})$. Since all quantities are non-negative, this implies that $\left|\frac{\beta}{H_{\hat{k}}} - \frac{\beta}{H^*}\right| \leq \epsilon$. Using the bounded suboptimality of $\widehat{\theta}_{\hat{k}}$ as an optimizer of $c(H_{\hat{k}}, \theta)$, we have

$$\begin{aligned}
\ell^\intercal \Phi \widehat{\theta}_{\hat{k}} + \left(\frac{1}{1-\gamma} + H_{\hat{k}}\right)\left(V_3(\widehat{\theta}_{\hat{k}}) + V_4(\widehat{\theta}_{\hat{k}})\right) &\leq \ell^\intercal \Phi \theta^*_{\hat{k}} + \left(\frac{1}{1-\gamma} + H_{\hat{k}}\right)\left(V_3(\theta^*_{\hat{k}}) + V_4(\theta^*_{\hat{k}})\right) + \frac{\epsilon}{2} \\
&\leq \ell^\intercal \Phi \theta^* + \left(\frac{1}{1-\gamma} + H^*\right)(V_3(\theta^*) + V_4(\theta^*)) + \epsilon \\
&= \ell^\intercal \Phi \theta^* + \frac{1}{1-\gamma}(V_3(\theta^*) + V_4(\theta^*)) \\
&\quad + \sqrt{V_3(\theta^*) + V_4(\theta^*)} + \epsilon.
\end{aligned}$$

Next, we crudely bound $\ell^\intercal \Phi \theta \leq \sqrt{d} C S$ and use $\left(\frac{1}{1-\gamma} + H_{\hat{k}}\right)^{-1} \leq \frac{1}{H_{\hat{k}}}$ to obtain

$$\begin{aligned}
V_3(\widehat{\theta}_{\hat{k}}) + V_4(\widehat{\theta}_{\hat{k}}) &\leq \frac{1}{H_{\hat{k}}}\left(2\sqrt{d}CS + \sqrt{V_3(\theta^*) + V_4(\theta^*)} + \frac{1}{(1-\gamma)}(V_3(\theta^*) + V_4(\theta^*)) + \epsilon\right) \\
&\leq \left(\frac{1}{H^*} + \beta\epsilon\right)\left(2\sqrt{d}CS + \sqrt{V_3(\theta^*) + V_4(\theta^*)} + \frac{1}{(1-\gamma)}(V_3(\theta^*) + V_4(\theta^*)) + \epsilon\right) \\
&\leq 2\sqrt{d}CS\sqrt{V_3(\theta^*) + V_4(\theta^*)} + (V_3(\theta^*) + V_4(\theta^*)) + \frac{(V_3(\theta^*) + V_4(\theta^*))^{\frac{3}{2}}}{(1-\gamma)} + O(\epsilon).
\end{aligned}$$

Then, applying Lemma 13, we have

$$\begin{aligned}
\ell^\intercal \Phi \mu_{\theta_{\hat{k}}} &\leq \ell^\intercal \Phi \widehat{\theta}_{\hat{k}} + \frac{3}{1-\gamma}\left(2\sqrt{d}CS\sqrt{V_3(\theta^*)+V_4(\theta^*)} + (V_3(\theta^*)+V_4(\theta^*)) + \frac{(V_3(\theta^*)+V_4(\theta^*))^{\frac{3}{2}}}{(1-\gamma)}\right) \\
&\quad + O\left(\frac{\epsilon}{1-\gamma}\right) \\
&\leq \ell^\intercal \Phi \theta^* + \frac{3}{1-\gamma}\left(2\sqrt{d}CS\sqrt{V_3(\theta^*)+V_4(\theta^*)} + (V_3(\theta^*)+V_4(\theta^*)) + \frac{(V_3(\theta^*)+V_4(\theta^*))^{\frac{3}{2}}}{(1-\gamma)}\right) \\
&\quad + \left(\frac{1}{1-\gamma} + H^*\right)(V_3(\theta^*)+V_4(\theta^*)) + O\left(\frac{\epsilon}{1-\gamma}\right) \\
&\leq \ell^\intercal \nu_{\theta^*} + \frac{3}{1-\gamma}\left(2\sqrt{d}CS\sqrt{V_3(\theta^*)+V_4(\theta^*)} + (V_3(\theta^*)+V_4(\theta^*)) + \frac{(V_3(\theta^*)+V_4(\theta^*))^{\frac{3}{2}}}{(1-\gamma)}\right) \\
&\quad + \left(\frac{1}{1-\gamma} + H^*\right)(V_3(\theta^*)+V_4(\theta^*)) + \epsilon + \frac{3}{1-\gamma}(V_3(\theta^*)+V_4(\theta^*)) + O\left(\frac{\epsilon}{1-\gamma}\right) \\
&\leq \ell^\intercal \nu_{\theta^*} + \left(1 + \frac{3}{1-\gamma}2\sqrt{d}CS\right)\sqrt{V_3(\theta^*)+V_4(\theta^*)} + \frac{3}{1-\gamma}\frac{(V_3(\theta^*)+V_4(\theta^*))^{\frac{3}{2}}}{(1-\gamma)} \\
&\quad + \frac{7}{1-\gamma}(V_3(\theta^*)+V_4(\theta^*)) + O\left(\frac{\epsilon}{1-\gamma}\right).
\end{aligned}$$

All in all, this simplifies to

$$\ell^\intercal \nu_{\theta_{\hat{k}}} \leq \min_\theta \ell^\intercal \nu_\theta + O\left(\sqrt{V_3(\theta)+V_4(\theta)}\right) + O\left((V_3(\theta)+V_4(\theta))^{\frac{3}{2}}\right) + O\left(\frac{\epsilon}{1-\gamma}\right).$$

Using our assumption that $(V_3(\theta)+V_4(\theta)) < 1$, we obtain the theorem statement.

We now turn towards bounding the subgradient steps and number of samples. Since the H_k are equal to the average cost case, we can still bound $K = O(\log(1/\epsilon))$. Theorem 10 requires we use

$$T_k = \frac{S^2}{\epsilon^2}\left(H_k(C_3+C_4) + \sqrt{d} + 2\sqrt{10\log\frac{1}{\delta}} + 2\sqrt{5d\log\left(1+\frac{S^2T}{d}\right)}\right)^2,$$

and so the total number of gradient descent steps can be bounded by

$$\sum_k T_k \leq KT_K = O\left(\frac{1}{\epsilon^4}\right),$$

with the same number of samples as in the average cost case. □

D Related Work

One of the approximate linear programming methods, proposed by Schweitzer and Seidmann [1985], was to project the primal LP into a subspace. These ideas have seen lots of recent work [de Farias and Van Roy, 2003a,b, Hauskrecht and Kveton, 2003, Guestrin et al., 2004, Petrik and Zilberstein, 2009, Desai et al., 2012]. As noted by Desai et al. [2012], the prior work on ALP either requires access to samples from a distribution that depends on optimal policy or assumes the ability to solve an LP with as many constraints as states.

The first theoretical analysis of ALP methods, by de Farias and Van Roy [2003a], analyzed the discounted primal LP (7) performance when only value functions of the form $J = \Phi w$, for some feature matrix Φ, are considered. Roughly, they show that the ALP solution w^* has the family of error bound indexed by a vector $u \in \mathbb{R}^{\mathcal{X}}$

$$\|J_* - \Psi w_*\| \leq \frac{2c^\intercal u}{1-\beta_u} \min_w \|J_* - \Psi w\|_{\infty,1/u} \tag{54}$$

where c is a "state-relevance" vector and $\beta_u = \gamma \max_{x,a} \sum_{x'} P_{(x,a),x'} u(x')/u(x)$ is a "goodness-of-fit" parameter that measures how well u represents a stationary distribution. Unfortunately, c and u are typically hard to choose (for example, a good choice of c would be the stationary distribution under w^*, which we do not know); but more importantly, the bound can be vacuous if Ψ does not model the optimal value function well and $\|J_* - \Psi w\|$ is always large. In particular, the problem we are considering in Definition 2 requires an additive bound with respect to the optimal parameter.

This result has some limitations. We need to specify c, but a good choice is usually not known a priori. The authors show that, if the ALP is solved iteratively using the $c = \mu_{\pi_{\Psi w_*},\nu}$ from the last iteration, then for an arbitrary probability distribution $\nu \in \Delta_{[\mathcal{X}]}$ and accompanying $\mu_{\pi,\nu} = (1-\gamma)\nu^\intercal (I - \gamma P^\pi)^{-1}$, we must have

$$\|J_{\pi_J} - J_*\|_{1,\nu} \leq \frac{1}{1-\gamma} \|J - J_*\|_{1,\mu_{\pi_J,\nu}} ,$$

where J_* is the discounted cost of the optimal policy. This suggests that we should choose $c = \mu_{\pi_{\Psi w_*},\nu}$, which is impossible as w_* is not known a priori.

A second limitation is that the ALP remains computationally expensive if the number of constraints is large and was addressed in de Farias and Van Roy [2004] by reducing the number of constraints by sampling them. The idea is to sample a relatively small number of constraints and solve the resulting LP. Let $\mathcal{N} \subset \mathbb{R}^d$ be a known set that contains w_* (solution of ALP). Let $\mu^V_{\pi,c}(x) = \mu_{\pi,c}(x)V(x)/(\mu^\intercal_{\pi,c}V)$ and define the distribution $\rho^V_{\pi,c}(x,a) = \mu^V_{\pi,c}(x)/\mathcal{A}$. Let $\delta \in (0,1)$ and $\epsilon \in (0,1)$. Let $\overline{\beta}_u = \gamma \max_x \sum_{x'} P_{(x,\pi_*(x)),x'} u(x')/u(x)$ and

$$D = \frac{(1+\overline{\beta}_V)\mu^\intercal_{\pi_*,c}V}{2c^\intercal J_*} \sup_{w\in\mathcal{N}} \|J_* - \Psi w\|_{\infty,1/V} , \qquad m \geq \frac{16\mathcal{A}D}{(1-\gamma)\epsilon}\left(d \log \frac{48\mathcal{A}D}{(1-\gamma)\epsilon} + \log\frac{2}{\delta}\right) .$$

Let $\mathcal{S}$ be a set of m random state-action pairs sampled under $\rho^V_{\pi_*,c}$. Let $\widehat{w}$ be a solution of the following sampled LP:

$$\begin{aligned}&\max_{w\in\mathbb{R}^d}\ c^\intercal \Psi w\,,\\ &\text{s.t.}\quad w\in\mathcal{N},\ \forall (x,a)\in\mathcal{S},\ \ell(x,a)+\gamma P_{(x,a),:}\Psi w\geq(\Psi w)(x)\ .\end{aligned}$$

de Farias and Van Roy [2004] prove that with probability at least $1-\delta$, we have

$$\|J_*-\Psi\widehat{w}\|_{1,c}\leq\|J_*-\Psi w_*\|_{1,c}+\epsilon\,\|J_*\|_{1,c}\ .$$

Unfortunately, $\mu_{\pi_*,c}$ (which was used in the definition of D) depends on the optimal policy, which is obviously unknown, which makes this method difficult to implement.

In the primal form (4), an extra constraint $h=\Psi w$ is added to obtain

$$\begin{aligned}&\max_{\lambda,w}\lambda\,,\\ &\text{s.t.}\quad B(\lambda e+\Psi w)\geq\ell+P\Psi w\ .\end{aligned}\tag{55}$$

Let λ_* be the average loss of the optimal policy and $(\widetilde{\lambda},\widetilde{w})$ be the solution of this LP. It turns out that the greedy policy with respect to $\widetilde{w}$ can be arbitrarily bad even if $\left|\lambda_*-\widetilde{\lambda}\right|$ was small [de Farias and Van Roy, 2003b]. de Farias and Van Roy [2003b] propose a two stage procedure, where the above LP is the first stage and the second stage is

$$\begin{aligned}&\max_{w}c^\intercal\Psi w\,,\\ &\text{s.t.}\quad B(\widetilde{\lambda}e+\Psi w)\leq\ell+P\Psi w\,,\end{aligned}\tag{56}$$

where c is a user specified weight vector. Let $\widehat{w}$ be the solution of the second stage. Let λ_w and μ_w be the average loss and the stationary distribution of the greedy policy with respect to Ψw. de Farias and Van Roy [2003b] prove that

$$\lambda_w-\lambda_*\leq\|h_*-\Psi w\|_{1,\mu_w}\ .$$

Further, it is shown that $\widehat{w}$ minimizes $\left\|h_{\widetilde{\lambda}}-\Psi w\right\|_{1,c}$ and that

$$\|h_*-\Psi\widehat{w}\|_{1,c}\leq\left\|h_{\widetilde{\lambda}}-\Psi\widehat{w}\right\|_{1,c}+(\lambda_*-\widetilde{\lambda})c^\intercal(I-P^{\pi_*})^{-1}e\,,$$

which implies that $\|h_*-\Psi\widehat{w}\|_{1,c}$ is small. To get that $\lambda_{\widehat{w}}-\lambda_*$ is small, we need to use $c=\mu_{\widehat{w}}$. Value of $\mu_{\widehat{w}}$ is obtained only after solving the optimization problem (56). To fix this problem, de Farias and Van Roy [2003b] propose to solve (56) iteratively, using $c=\mu_{\widehat{w}}$ from the solution of

the last round.

The above approach has two problems. First, it is still not clear if the average loss of the resultant policy is close to λ_* (or the best policy in the policy class). Second, iteratively solving (56) is computationally expensive. Similar results are also obtained by Desai et al. [2012] who also show that if we were able to sample from the stationary distribution of the optimal policy, then LP (55) can be solved efficiently.

Desai et al. [2012] study a smoothed version of ALP, in which slack variables are introduced that allow for some violation of the constraints. Let D' be a violation budget. The smoothed ALP (SALP) has the form of

$$\begin{aligned} &\max_{w,s} c^\intercal \Psi w\,, && \max_{w,s} c^\intercal \Psi w - \frac{2\mu_{\pi_*,c}^\intercal s}{1-\gamma}\,, \\ &\text{s.t.} \quad \Psi w \le L\Psi w + s,\ \mu_{\pi_*,c}^\intercal s \le D',\ s \ge \mathbf{0}, && \text{s.t.} \quad \Psi w \le L\Psi w + s,\ s \ge \mathbf{0}\,. \end{aligned}$$

The ALP on RHS is equivalent to LHS with a specific choice of D'. Let $\overline{U} = \{u \in \mathbb{R}^{\mathcal{X}} \ : \ u \ge \mathbf{1}\}$ be a set of weight vectors. Desai et al. [2012] prove that if w_* is a solution to above problem, then

$$\|J_* - \Psi w_*\|_{1,c} \le \inf_{w,u\in\overline{U}} \|J_* - \Psi w\|_{\infty,1/u} \left(c^\intercal u + \frac{2(\mu_{\pi_*,c}^\intercal u)(1+\beta_u)}{1-\gamma} \right)\,.$$

The above bound improves (54) as $\overline{U}$ is larger than U and RHS in the above bound is smaller than RHS of (54). Further, they prove that if η is a distribution and we choose $c = (1-\gamma)\eta^\intercal(I - \gamma P^{\pi_{\Psi w_*}})$, then

$$\left\|J_{\mu_{\Psi w_*}} - J_*\right\|_{1,\eta} \le \frac{1}{1-\gamma} \left(\inf_{w,u\in\overline{U}} \|J_* - \Psi w\|_{\infty,1/u} \left(c^\intercal u + \frac{2(\mu_{\pi_*,\nu}^\intercal u)(1+\beta_u)}{1-\gamma} \right) \right)\,.$$

Similar methods are also proposed by Petrik and Zilberstein [2009]. One problem with this result is that c is defined in terms of w_*, which itself depends on c. Also, the smoothed ALP formulation uses π_* which is not known. Desai et al. [2012] also propose a computationally efficient algorithm. Let $\mathcal{S}$ be a set of S random states drawn under distribution $\mu_{\pi_*,c}$. Let $\mathcal{N}' \subset \mathbb{R}^d$ be a known set that contains the solution of SALP. The algorithm solves the following LP:

$$\begin{aligned} &\max_{w,s} c^\intercal \Psi w - \frac{2}{(1-\gamma)S} \sum_{x\in\mathcal{S}} s(x)\,, \\ &\text{s.t.} \quad \forall x \in \mathcal{S},\ (\Psi w)(x) \le (L\Psi w)(x) + s(x),\ s \ge \mathbf{0},\ w \in \mathcal{N}'\,. \end{aligned}$$

Let $\widehat{w}$ be the solution of this problem. Desai et al. [2012] prove high probability bounds on the approximation error $\|J_* - \Psi\widehat{w}\|_{1,c}$. However, it is no longer clear if a performance bound on $\|J_* - J_{\pi_{\Psi\widehat{w}}}\|_{1,c}$ can be obtained from this approximation.

Next, we turn our attention to average cost ALP. Let ν be a distribution over states, $u : [\mathcal{X}] \to$

$[1,\infty)$, $\eta > 0$, $\gamma \in [0,1]$, $P_\gamma^\pi = \gamma P^\pi + (1-\gamma)\mathbf{1}\nu^\intercal$, and $L_\gamma h = \min_\pi(\ell_\pi + P_\gamma^\pi h)$. de Farias and Van Roy [2006] propose the following optimization problem:

$$\begin{aligned} &\min_{w,s_1,s_2} \; s_1 + \eta s_2 \,, \\ &\text{s.t.} \quad L_\gamma \Psi w - \Psi w + s_1 \mathbf{1} + s_2 u \geq \mathbf{0},\; s_2 \geq 0 \,. \end{aligned} \tag{57}$$

Let $(w_*, s_{1,*}, s_{2,*})$ be the solution of this problem. Define the mixing time of policy π by

$$\tau_\pi = \inf\left\{ \tau \;:\; \left| \frac{1}{t} \sum_{t'=0}^{t-1} \nu^\intercal (P^\pi)^{t'} \ell_\pi - \lambda_\pi \right| \leq \frac{\tau}{t},\, \forall t \right\} .$$

Let $\tau_* = \liminf_{\delta \to 0}\{\tau_\pi : \lambda_\pi \leq \lambda_* + \delta\}$. Let π_γ^* be the optimal policy when discount factor is γ. Let $\pi_{\gamma,w}$ be the greedy policy with respect to Ψw when discount factor is γ, $\mu_{\gamma,\pi}^\intercal = (1-\gamma)\sum_{t=0}^\infty \gamma^t \nu^\intercal (P^\pi)^t$ and $\mu_{\gamma,w} = \mu_{\gamma,\pi_{\gamma,w}}$. de Farias and Van Roy [2006] prove that if $\eta \geq (2-\gamma)\mu_{\gamma,\pi_\gamma^*}^\intercal u$,

$$\lambda_{w_*} - \lambda_* \leq \frac{(1+\beta)\eta \max(D'',1)}{1-\gamma} \min_w \left\| h_\gamma^* - \Psi w \right\|_{\infty,1/u} + (1-\gamma)(\tau_* + \tau_{\pi_{w_*}}) \,,$$

where $\beta = \max_\pi \|I - \gamma P^\pi\|_{\infty,1/u}$, $D'' = \mu_{\gamma,w_*}^\intercal V/(\nu^\intercal V)$ and $V = L_\gamma \Psi w_* - \Psi w_* + s_{1,*}\mathbf{1} + s_{2,*} u$. Similar results are obtained more recently by Veatch [2013].

An appropriate choice for vector ν is $\nu = \mu_{\gamma,w_*}$. Unfortunately, w_* depends on ν. We should also note that solving (57) can be computationally expensive. de Farias and Van Roy [2006] propose constraint sampling techniques similar to [de Farias and Van Roy, 2004], but no performance bounds are provided.