A. Related Work

de Farias and Van Roy (2003a) study the *discounted* version of the primal form (1). Let $c \in \mathbb{R}^X$ be a vector with positive components and $\gamma \in (0, 1)$ be a discount factor. Let $L : \mathbb{R}^X \to \mathbb{R}^X$ be the Bellman operator defined by $(LJ)(x) = \min_{a\in\mathcal{A}}(\ell(x, a) + \gamma \sum_{x'\in\mathcal{X}} P_{(x,a),x'} J(x'))$ for $x \in \mathcal{X}$. Let $\Psi \in \mathbb{R}^{X\times d}$ be a feature matrix. The exact and approximate LP problems are as follows:

$$\begin{aligned} &\max_{J\in\mathbb{R}^X} c^\top J\,, && \max_{w\in\mathbb{R}^d} c^\top \Psi w\,,\\ &\text{s.t.}\quad LJ \geq J\,, && \text{s.t.}\quad L\Psi w \geq \Psi w\,. \end{aligned}$$

which can also be written as

$$\begin{aligned} &\max_{J\in\mathbb{R}^X} c^\top J\,, && \max_{w\in\mathbb{R}^d} c^\top \Psi w\,,\\ &\text{s.t.}\quad \forall (x, a),\ \ell(x, a) + \gamma P_{(x,a),:} J \geq J(x)\,, && \text{s.t.}\quad \forall (x, a),\ \ell(x, a) + \gamma P_{(x,a),:} \Psi w \geq (\Psi w)(x)\,. \end{aligned} \tag{18}$$

The optimization problem on the RHS is an approximate LP with the choice of $J = \Psi w$. Let $J_\pi(x) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \ell(x_t, \pi(x_t)) | x_0 = x\right]$ be value of policy π, J_* be the solution of LHS, and $\pi_J(x) = \text{argmin}_{a\in\mathcal{A}}(\ell(x, a) + \gamma P_{(x,a),:} J)$ be the greedy policy with respect to J. Let $\nu \in \Delta_{\mathcal{X}}$ be a probability distribution and define $\mu_{\pi,\nu} = (1-\gamma)\nu^\top (I - \gamma P^\pi)^{-1}$. de Farias and Van Roy (2003a) prove that for any J satisfying the constraints of the LHS of (18),

$$\left\| J_{\pi_J} - J_* \right\|_{1,\nu} \leq \frac{1}{1-\gamma} \left\| J - J_* \right\|_{1,\mu_{\pi_J,\nu}}\ . \tag{19}$$

Define $\beta_u = \gamma \max_{x,a} \sum_{x'} P_{(x,a),x'} u(x')/u(x)$. Let $U = \{u \in \mathbb{R}^X : u \geq \mathbf{1}, u \in \text{span}(\Psi), \beta_u < 1\}$. Let w_* be the solution of ALP. de Farias and Van Roy (2003a) show that for any $u \in U$,

$$\left\| J_* - \Psi w_* \right\|_{1,c} \leq \frac{2c^\top u}{1-\beta_u} \min_w \left\| J_* - \Psi w \right\|_{\infty,1/u}\ . \tag{20}$$

This result has a number of limitations. First, solving ALP can be computationally expensive as the number of constraints is large. Second, it assumes that the feasible set of ALP is non-empty. Finally, Inequality (19) implies that $c = \mu_{\pi_{\Psi w_*},\nu}$ is an appropriate choice to obtain performance bounds. However, w_* itself is function of c and is not known before solving ALP.

de Farias and Van Roy (2004) propose a computationally efficient algorithm that is based on a constraint sampling technique. 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 $\mu^V_{\pi,c}(x) = \mu_{\pi,c}(x)V(x)/(\mu^\top_{\pi,c}V)$ and define the distribution $\rho^V_{\pi,c}(x, a) = \mu^V_{\pi,c}(x)/A$. 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

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

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:

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

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

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

This result has a number of limitations. First, vector $\mu_{\pi_*,c}$ (that is used in the definition of D) depends on the optimal policy, but an optimal policy is what we want to compute in the first place. Second, we can no longer use Inequality (19) to obtain a performance bound (a bound on $\| J_{\pi_{\Psi\widehat{w}}} - J_* \|_{1,c}$), as $\Psi\widehat{w}$ does not necessarily satisfy all constraints of ALP.

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

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

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

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

The above bound improves (20) as $\overline{U}$ is larger than U and RHS in the above bound is smaller than RHS of (20). Further, they prove that if η is a distribution and we choose $c = (1-\gamma)\eta^\top(I - \gamma P^{\pi_{\Psi w_*}})$, then

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

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:

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

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 bound.

Next, we turn our attention to average cost ALP, which is a more challenging problem and is also the focus of this paper. Let ν be a distribution over states, $u : \mathcal{X} \to [1,\infty)$, $\eta > 0$, $\gamma \in [0,1]$, $P_\gamma^\pi = \gamma P^\pi + (1-\gamma)\mathbf{1}\nu^\top$, and $L_\gamma h = \min_\pi(\ell_\pi + P_\gamma^\pi h)$. de Farias and Van Roy (2006) propose the following optimization problem:

$$
\begin{aligned}
&\min_{w,s_1,s_2}\ s_1 + \eta s_2\,, \\
&\text{s.t.}\quad L_\gamma \Psi w - \Psi w + s_1\mathbf{1} + s_2 u \ge \mathbf{0},\ s_2 \ge 0\,.
\end{aligned}
\tag{21}
$$

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

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

Let $\tau_* = \liminf_{\delta\to 0}\{\tau_\pi : \lambda_\pi \le \lambda_* + \delta\}$. 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, $\mu_{\gamma,\pi}^\top = (1-\gamma)\sum_{t=0}^\infty \gamma^t \nu^\top (P^\pi)^t$ and $\mu_{\gamma,w} = \mu_{\gamma,\pi_{\gamma,w}}$. de Farias and Van Roy (2006) prove that if $\eta \ge (2-\gamma)\mu_{\gamma,\pi_\gamma^*}^\top u$,

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

where $\beta = \max_\pi \|I - \gamma P^\pi\|_{\infty,1/u}$, $D'' = \mu_{\gamma,w_*}^\top V/(\nu^\top V)$ 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 (21) 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.

Wang et al. (2008) study ALP (3) and show that there is a dual form for standard value function based algorithms, including on-policy and off-policy updating and policy improvement. They also study the convergence of these methods, but no performance bounds are shown.

B. Proofs of Section 2

Proof of Theorem 3. Let $z_* = \mathrm{argmin}_{z\in\mathcal{Z}} \sum_{t=1}^T f_t(z)$ 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

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

Thus,

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

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 $1-\delta$,

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

Thus,

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

□

Proof of Lemma 5. We prove the lemma by showing that conditions of Theorem 3 are satisfied. We begin by calculating the subgradient and bounding its norm independently of the number of states. 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,

$$\begin{aligned}\nabla_\theta c(\theta) &= \ell^\top \Phi + H \sum_{(x,a)} \nabla_\theta \left|[\mu_0(x,a) + \Phi_{(x,a),:}\theta]_-\right| + H \sum_{x'} \nabla_\theta \left|(P-B)_{:,x'}^\top \Phi\theta\right| \\ &= \ell^\top \Phi - H \sum_{(x,a)} \Phi_{(x,a),:}\mathbb{I}_{\{\mu_0(x,a) + \Phi_{(x,a),:}\theta < 0\}} + H \sum_{x'} (P-B)_{:,x'}^\top \Phi s((P-B)_{:,x'}^\top \Phi\theta)\,,\end{aligned} \tag{22}$$

where $s(z) = \mathbb{I}_{\{z>0\}} - \mathbb{I}_{\{z<0\}}$ is the sign function. Let ± denote the plus or minus sign (the exact sign does not matter here). Let $G = \|\nabla_\theta c(\theta)\|$. We have that

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

Thus,

$$G \leq \sqrt{\sum_{i=1}^{d} (\ell^\top \Phi_{:,i})^2} + H\sqrt{d} + H\sqrt{\sum_{i=1}^{d} \left(\sum_{(x,a)} \left(\pm \sum_{x'} (P-B)_{(x,a),x'} \right) \Phi_{(x,a),i} \right)^2}$$

$$\leq \sqrt{d} + H\sqrt{d} + H\sqrt{\sum_{i=1}^{d} \left(2 \sum_{(x,a)} \left| \Phi_{(x,a),i} \right| \right)^2} = \sqrt{d}(1+3H)\,,$$

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

Next, we show that norm of the subgradient estimate is bounded by G':

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

Finally, we show that the subgradient estimate is unbiased:

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

The result then follows from Theorem 3 and Remark 4.

□

C. Sampling Constraints

In this section we describe our second algorithm for average cost MDP problems. The main idea is to use the results on polytope constraint sampling (de Farias and Van Roy, 2004; Calafiore and Campi, 2005; Campi and Garatti, 2008) to reduce the dual LP to a size we can solve exactly. Using classical uniform convergence results (Vapnik and Chervonenkis, 1971), de Farias and Van Roy (2004) show that if we sample $k = O(d/\epsilon)$ affine constraints in $\mathbb{R}^d$, then with high probability, any point that satisfies all k sampled constraints also satisfies most of the original set of constraints: a proportion at least $1-\epsilon$ under the sampling distribution.

Let $\mathcal{L}$ be a family of affine constraints indexed by i: constraint i is satisfied at point $w \in \mathbb{R}^d$ if $a_i^\top w + b_i \geq 0$. Let $\mathcal{I}$ be the family of constraints by selecting k random constraints in $\mathcal{L}$ with respect to measure q.

Theorem 7 (de Farias and Van Roy (2004))**.** *Assume there exists a vector that satisfies all constraints in $\mathcal{L}$. For any δ and ϵ, if we take $m \geq \frac{4}{\epsilon}\left(d \log \frac{12}{\epsilon} + \log \frac{2}{\delta}\right)$, then, with probability $1-\delta$, a set $\mathcal{I}$ of m i.i.d. random variables drawn from $\mathcal{L}$ with respect to distribution q satisfies*

$$\sup_{\{w : \forall i \in \mathcal{I}, a_i^\top w + b_i \geq 0\}} q(\{j : a_j^\top w + b_j < 0\}) \leq \epsilon\,.$$

Our algorithm takes the following inputs: a positive constant S, a stationary distribution μ_0, a set $\Theta = \{\theta \ : \ \theta^\top \Phi^\top \mathbf{1} = 1 - \mu_0^\top \mathbf{1},\ \|\theta\| \leq S\}$, a distribution q_1 over the state-action space, a distribution q_2 over the state space, and constraint violation functions $v_1 : \mathcal{X} \times \mathcal{A} \to [-1, 0]$ and $v_2 : \mathcal{X} \to [0, 1]$. We will consider two families of constraints:

$$\mathcal{L}_1 = \{\mu_0(x,a) + \Phi_{(x,a),:}\theta \geq v_1(x,a) \mid (x,a) \in \mathcal{X} \times \mathcal{A}\}\,,$$

$$\mathcal{L}_2 = \left\{(P-B)^\top_{:,x}(\mu_0 + \Phi\theta) \leq v_2(x) \mid x \in \mathcal{X}\right\} \bigcup \left\{(P-B)^\top_{:,x}(\mu_0 + \Phi\theta) \geq -v_2(x) \mid x \in \mathcal{X}\right\}\,.$$

Input: Constant $S > 0$, stationary distribution μ_0, distributions q_1 and q_2, constraint violation functions v_1 and v_2, number of samples k_1 and k_2.
For $i = 1, 2$, let $\mathcal{I}_i$ be k_i constraints sampled from $\mathcal{L}_i$ under distribution q_i.
Let $\mathcal{I}$ be the set of vectors that satisfy all constraints in $\mathcal{I}_1$ and $\mathcal{I}_2$.
Let $\widetilde{\theta}$ be the solution to LP:

$$\begin{aligned} &\min_{\theta \in \Theta} \ell^\top (\mu_0 + \Phi\theta) \,, \\ &\text{s.t.} \quad \theta \in \mathcal{I},\, \theta \in \Theta \,. \end{aligned} \tag{24}$$

Return policy $\mu_{\widetilde{\theta}}$.

Figure 4. The Constraint Sampling Method for Markov Decision Processes

Let θ_* be the solution of

$$\begin{aligned} &\min_{\theta \in \Theta} \ell^\top (\mu_0 + \Phi\theta) \,, \\ &\text{s.t.} \quad \theta \in \mathcal{L}_1,\, \theta \in \mathcal{L}_2,\, \theta \in \Theta \,. \end{aligned} \tag{23}$$

The constraint sampling algorithm is shown in Figure 4. We refer to (24) as the sampled ALP, while we refer to (3) as the full ALP.

C.1. Analysis

We require Assumption A1 as well as:

Assumption A2 ***(Feasibility)*** There exists a vector that satisfies all constraints $\mathcal{L}_1$ and $\mathcal{L}_2$.

Validity of this assumption depends on the choice of functions v_1 and v_2. Larger functions ensure that this assumption is satisfied, but as we show, this leads to larger error.

The next two lemmas apply theorem 7 to constraints $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively.

Lemma 8. *Let $\delta_1 \in (0, 1)$ and $\epsilon_1 \in (0, 1)$. If we choose $k_1 = \frac{4}{\epsilon_1}\left(d \log \frac{12}{\epsilon_1} + \log \frac{2}{\delta_1}\right)$, then with probability at least $1 - \delta_1$, $\sum_{(x,a)} \left| [\mu_0(x, a) + \Phi_{(x,a),:}\widetilde{\theta}]_- \right| \le S C_1 \epsilon_1 + \|v_1\|_1$.*

Proof. Applying theorem 7, we have that w.p. $1 - \delta_1$, $q_1(\mu_0(x, a) + \Phi_{(x,a),:}\widetilde{\theta} \ge v_1(x, a)) \ge 1 - \epsilon_1$, and thus

$$\sum_{(x,a)} q_1(x, a) \mathbb{I}_{\{\mu_0(x,a) + \Phi_{(x,a),:}\widetilde{\theta} < v_1(x,a)\}} \le \epsilon_1 \ .$$

Let $L = \sum_{(x,a)} \left| [\mu_0(x,a) + \Phi_{(x,a),:}\widetilde{\theta}]_- \right|$. With probability $1-\delta_1$,

$$\begin{aligned}
L &= \sum_{(x,a)} \left| [\mu_0(x,a) + \Phi_{(x,a),:}\widetilde{\theta}]_- \right| \mathbb{I}_{\{\mu_0(x,a)+\Phi_{(x,a),:}\widetilde{\theta} \le v_1(x,a)\}} \\
&\quad + \sum_{(x,a)} \left| [\mu_0(x,a) + \Phi_{(x,a),:}\widetilde{\theta}]_- \right| \mathbb{I}_{\{\mu_0(x,a)+\Phi_{(x,a),:}\widetilde{\theta} > v_1(x,a)\}} \\
&\le \sum_{(x,a)} \left| \Phi_{(x,a),:}\widetilde{\theta} \right| \mathbb{I}_{\{\mu_0(x,a)+\Phi_{(x,a),:}\widetilde{\theta} \le v_1(x,a)\}} + \left\| v_1 \right\|_1 \\
&\le \sum_{(x,a)} \left\| \Phi_{(x,a),:} \right\| \left\| \widetilde{\theta} \right\| \mathbb{I}_{\{\mu_0(x,a)+\Phi_{(x,a),:}\widetilde{\theta} \le v_1(x,a)\}} + \left\| v_1 \right\|_1 \\
&\le \sum_{(x,a)} SC_1 q_1(x,a) \mathbb{I}_{\{\mu_0(x,a)+\Phi_{(x,a),:}\widetilde{\theta} \le v_1(x,a)\}} + \left\| v_1 \right\|_1 \\
&\le SC_1 \epsilon_1 + \left\| v_1 \right\|_1 \ .
\end{aligned}$$

□

Lemma 9. *Let $\delta_2 \in (0,1)$ and $\epsilon_2 \in (0,1)$. If we choose $k_2 = \frac{4}{\epsilon_2}\left(d \log \frac{12}{\epsilon_2} + \log \frac{2}{\delta_2}\right)$, then with probability at least $1-\delta_2$, $\left\| (P-B)^\top \Phi \widetilde{\theta} \right\|_1 \le SC_2 \epsilon_2 + \left\| v_2 \right\|_1$.*

Proof. Applying theorem 7, we have that $q_2\left(\left| (P-B)^\top_{:,x} \Phi \widetilde{\theta} \right| \le v_2(x) \right) \ge 1 - \epsilon_2$. This yields

$$\sum_x q_2(x) \mathbb{I}_{\{\left| (P-B)^\top_{:,x} \Phi \widetilde{\theta} \right| \ge v_2(x)\}} \le \epsilon_2 \ . \tag{25}$$

Let $L' = \sum_x \left| (P-B)^\top_{:,x} \Phi \widetilde{\theta} \right|$. Thus, with probability $1-\delta_2$,

$$\begin{aligned}
L' &= \sum_x \left| (P-B)^\top_{:,x} \Phi \widetilde{\theta} \right| \mathbb{I}_{\{\left| (P-B)^\top_{:,x} \Phi \widetilde{\theta} \right| > v_2(x)\}} \\
&\quad + \sum_x \left| (P-B)^\top_{:,x} \Phi \widetilde{\theta} \right| \mathbb{I}_{\{\left| (P-B)^\top_{:,x} \Phi \widetilde{\theta} \right| \le v_2(x)\}} \\
&\le \sum_x \left\| (P-B)^\top_{:,x} \Phi \right\| \left\| \widetilde{\theta} \right\| \mathbb{I}_{\{\left| (P-B)^\top_{:,x} \Phi \widetilde{\theta} \right| > v_2(x)\}} + \left\| v_2 \right\|_1 \\
&\le \sum_x SC_2 q_2(x) \mathbb{I}_{\{\left| (P-B)^\top_{:,x} \Phi \widetilde{\theta} \right| > v_2(x)\}} + \left\| v_2 \right\|_1 \\
&\le SC_2 \epsilon_2 + \left\| v_2 \right\|_1 \ ,
\end{aligned}$$

where the last step follows from (25).

□

We are ready to prove the main result of this section. Let $\widetilde{\theta}$ denote the solution of the sampled ALP, θ_* denote the solution of the full ALP (23), and $\mu_{\widetilde{\theta}}$ be the stationary distribution of the solution policy. Our goal is to compare $\ell^\top \mu_{\widetilde{\theta}}$ and $\ell^\top \mu_{\theta_*}$.

Theorem 10. *Let $\epsilon \in (0,1)$ and $\delta \in (0,1)$. Let $\epsilon' = SC_1 \epsilon + \left\| v_1 \right\|_1$ and $\epsilon'' = SC_2 \epsilon + \left\| v_2 \right\|_1$. If we sample constraints with $k_1 = \frac{4}{\epsilon}\left(d \log \frac{12}{\epsilon} + \log \frac{4}{\delta}\right)$ and $k_2 = \frac{4}{\epsilon}\left(d \log \frac{12}{\epsilon} + \log \frac{4}{\delta}\right)$, then, with probability $1-\delta$,*

$$\begin{aligned}
\ell^\top \mu_{\widetilde{\theta}} &\le \ell^\top \mu_{\theta_*} + \tau(\mu_{\widetilde{\theta}}) \log(1/\epsilon')(2\epsilon' + \epsilon'') + 3\epsilon' \\
&\quad + \tau(\mu_*) \log(1/\left\| v_1 \right\|)(2 \left\| v_1 \right\| + \left\| v_2 \right\|) + 3 \left\| v_1 \right\| \ .
\end{aligned}$$

Proof. Let $\delta_1 = \delta_2 = \delta/2$. By Lemmas 8 and 9, w.p. $1-\delta$, $\sum_{(x,a)} \left| [\mu_0(x,a) + \Phi_{(x,a),:}\widetilde{\theta}]_- \right| \le \epsilon'$ and $\left\| (P-B)^\top (\mu_0 + \Phi\widetilde{\theta}) \right\|_1 \le \epsilon''$. Then by Lemma 2,

$$\left| \ell^\top \mu_{\widetilde{\theta}} - \ell^\top (\mu_0 + \Phi\widetilde{\theta}) \right| \le \tau(\mu_{\widetilde{\theta}}) \log(1/\epsilon')(2\epsilon' + \epsilon'') + 3\epsilon' \ .$$

We also have that $\ell^\top(\mu_0 + \Phi\widetilde{\theta}) \le \ell^\top(\mu_0 + \Phi\theta_*)$. Thus,

$$\begin{aligned} \ell^\top \mu_{\widetilde{\theta}} &\le \ell^\top(\mu_0 + \Phi\theta_*) + \tau(\mu_{\widetilde{\theta}}) \log(1/\epsilon')(2\epsilon' + \epsilon'') + 3\epsilon' \\ &\le \ell^\top \mu_{\theta_*} + \tau(\mu_{\widetilde{\theta}}) \log(1/\epsilon')(2\epsilon' + \epsilon'') + 3\epsilon' \\ &\quad + \tau(\mu_{\theta_*}) \log(1/\left\|v_1\right\|)(2\left\|v_1\right\| + \left\|v_2\right\|) + 3\left\|v_1\right\| , \end{aligned}$$

where the last step follows from Lemma 2. □