# Supplementary Material

## Index of Notation

For the convenience of the reader, we have collected the most frequently used symbols and their meanings in the following table:

| | |
|---|---|
| $\mathbb{R}, \mathbb{R}_+$ | real numbers; non-negative real numbers. |
| $\mathbb{R}^d, \mathbb{R}^{m\times n}$ | $d$-dimensional vectors; matrices of size $m \times n$. |
| $\boldsymbol{e}_i, \boldsymbol{0}, \boldsymbol{1}$ | standard basis vector: $e_{i,i} = 1$ and $e_{i,j} = 0$ for $i \neq j$; constant zero or one vectors. |
| $\boldsymbol{a} \oplus \boldsymbol{b}$ | concatenation of vectors: if $\boldsymbol{a} \in \mathbb{R}^n$ and $\boldsymbol{b} \in \mathbb{R}^m$, then $\boldsymbol{a} \oplus \boldsymbol{b} \in \mathbb{R}^{n+m}$. |
| $\mathcal{S}, \mathcal{A}, \mathcal{S}_*$ | sets of states, actions, and core states (Section 1.1 and Assumption 2). |
| $S, A, m$ | number of states $\lvert\mathcal{S}\rvert$, actions $\lvert\mathcal{A}\rvert$, and core states $\lvert\mathcal{S}_*\rvert$, respectively. |
| $s_0, s, s', a$ | planning state (Section 3) and other states $\in \mathcal{S}$; actions $\in \mathcal{A}$. |
| $\boldsymbol{P}, \boldsymbol{E}$ | row-stochastic matrices in $\mathbb{R}_+^{SA\times S}$ (Sections 1.1 and 1.2); $\boldsymbol{E}_{sa} = \boldsymbol{e}_s \in \mathbb{R}^S$. |
| $\boldsymbol{r}, \hat{r}$ | expected rewards $\in \mathbb{R}^{SA}$ (Section 1.1); random reward $\in [-1, 1]$ (Assumption 3). |
| $\gamma \in [0, 1)$ | discount factor (Section 1.1). |
| $v, v^*, v_\pi$ | value functions $S \longrightarrow \mathbb{R}$. |
| $\Delta_{\mathcal{S}}, \Delta_{\mathcal{A}}$ | sets of probability distributions over states and actions. |
| $\boldsymbol{\mu}, \boldsymbol{\pi}, \boldsymbol{\pi}(s)$ | probability distributions in $\Delta_{\mathcal{S}}$ and $\Delta_{\mathcal{A}}$, respectively; policy in $\mathcal{S} \longrightarrow \Delta_{\mathcal{A}}$. |
| $\boldsymbol{\Phi}, \boldsymbol{\varphi}_s, \boldsymbol{\varphi}_0$ | feature matrix $\in \mathbb{R}^{S\times d}$; state features $\in \mathbb{R}^d$; features of planning state $\boldsymbol{\varphi}_{s_0}$. |
| $\varepsilon_{\text{approx}}$ | approximation error of $\boldsymbol{\Phi}$: $\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \lVert \boldsymbol{v}^* - \boldsymbol{\Phi\theta} \rVert_\infty$. |
| $\boldsymbol{W}, \boldsymbol{W}_*$ | constraint matrices in $\{0,1\}^{(1+m)A\times SA}$ and $\{0,1\}^{mA\times SA}$ (Theorems 2 and 5). |
| $\boldsymbol{\lambda}, \boldsymbol{\lambda}_*, \boldsymbol{\theta}$ | dual variables $\in \mathbb{R}_+^{(1+m)A}$ and $\in \mathbb{R}_+^{mA}$, respectively; primal variables $\in \mathbb{R}^d$. |
| $\Lambda, \Lambda_\gamma, \mathcal{B}$ | dual spaces $\subset \mathbb{R}_+^{(1+m)A}$; primal space $\subset \mathbb{R}^d$ (Theorem 2 and Lemmas 8 and 11). |
| $\lVert \cdot \rVert_*$ | dual norm of $\lVert \cdot \rVert$: defined by $\lVert \boldsymbol{u} \rVert_* = \sup_{\lVert \boldsymbol{x} \rVert = 1} \langle \boldsymbol{u}, \boldsymbol{x} \rangle$. |

## A Proofs

### A.1 Approximation Error for the Linearly Relaxed Approximate LP

We start by recalling and improving the approximation error bounds for the Linearly Relaxed Approximate Linear Program (LRALP$_\mu$) of Lakshminarayanan et al. [27].

**Theorem 5.** *Suppose Assumptions 1 and 2 hold. Define the matrix* $\boldsymbol{W}_* \in \{0,1\}^{mA\times SA}$ *with rows* $[\boldsymbol{W}_*]_{sa} = \boldsymbol{e}_{sa} \in \mathbb{R}^{SA}$ *(*$s \in \mathcal{S}_*, a \in \mathcal{A}$*). For any (possibly unnormalized) initial distribution* $\boldsymbol{\mu} \in \mathbb{R}_+^S$*,*

$$V_{\text{LRALP}}(\boldsymbol{\mu}) \coloneqq \min \{ \boldsymbol{\mu}^\mathsf{T} \boldsymbol{\Phi\theta} \mid \boldsymbol{\theta} \in \mathbb{R}^d,\ \boldsymbol{W}_* \boldsymbol{r} + \boldsymbol{W}_*(\gamma \boldsymbol{P} - \boldsymbol{E})\boldsymbol{\Phi\theta} \leq \boldsymbol{0} \}. \qquad (\text{LRALP}_\mu)$$

*The value of* (LRALP$_\mu$) *is close to the optimal value of that initial distribution:*

$$\lvert V_{\text{LRALP}}(\boldsymbol{\mu}) - \boldsymbol{\mu}^\mathsf{T} \boldsymbol{v}^* \rvert \leq \frac{10 \lVert \boldsymbol{\mu} \rVert_1 \varepsilon_{\text{approx}}}{1-\gamma}.$$

This result follows from Lakshminarayanan et al. [27, Theorem IV.1], which we will not reproduce here for brevity. The error bound there is $2\lVert\boldsymbol{\mu}\rVert_1 (3\varepsilon_{\text{approx}} + \lVert J^*_{\text{ALP}} - J^*_{\text{LRA}} \rVert_\infty)/(1-\gamma)$, defining

$$J^*_{\text{ALP}}(s) \coloneqq \min \{ \boldsymbol{\varphi}_s^\mathsf{T} \boldsymbol{\theta} \mid \boldsymbol{\theta} \in \mathbb{R}^d,\ \boldsymbol{\Phi\theta} \geq \boldsymbol{v}^* \},$$

$$J^*_{\text{LRA}}(s) \coloneqq \min \{ \boldsymbol{\varphi}_s^\mathsf{T} \boldsymbol{\theta} \mid \boldsymbol{\theta} \in \mathbb{R}^d,\ \boldsymbol{W}_* \boldsymbol{E\Phi\theta} \geq \boldsymbol{W}_* \boldsymbol{E} \boldsymbol{v}^* \}.$$

It only remains for us to bound $\lVert J^*_{\text{ALP}} - J^*_{\text{LRA}} \rVert_\infty$, improving upon Theorem IV.2 [27]:

**Lemma 6.** *Under the conditions of Theorem 5,* $\lVert J^*_{\text{ALP}} - J^*_{\text{LRA}} \rVert_\infty \leq 2\varepsilon_{\text{approx}}$.

*Proof.* By Assumption 1, the optimal value function is well-approximated by the feature representation; $\boldsymbol{v}^* = \boldsymbol{\Phi\theta} + \boldsymbol{\delta}$ for some $\boldsymbol{\theta} \in \mathbb{R}^d$ and $\boldsymbol{\delta} \in \mathbb{R}^S$ with $\lVert\boldsymbol{\delta}\rVert_\infty \leq \varepsilon_{\text{approx}}$. By Assumption 2, $\boldsymbol{\Phi} = \boldsymbol{Z\Phi}_*$, so $\boldsymbol{v}^* = \boldsymbol{Z\Phi}_* \boldsymbol{\theta} + \boldsymbol{\delta}$. We use these facts after writing the linear program defining $J^*_{\text{ALP}}(s)$ in its dual form:

$$\begin{aligned} J^*_{\text{ALP}}(s) &= \max \{ \boldsymbol{\mu}^\mathsf{T} \boldsymbol{v}^* \mid \boldsymbol{\mu} \in \mathbb{R}_+^S,\ \boldsymbol{\mu}^\mathsf{T} \boldsymbol{\Phi} = \boldsymbol{\varphi}_s^\mathsf{T} \} \\ &= \max \{ \boldsymbol{\mu}^\mathsf{T} (\boldsymbol{Z\Phi}_* \boldsymbol{\theta} + \boldsymbol{\delta}) \mid \boldsymbol{\mu} \in \mathbb{R}_+^S,\ \boldsymbol{\mu}^\mathsf{T} \boldsymbol{Z\Phi}_* = \boldsymbol{\varphi}_s^\mathsf{T} \} \end{aligned}$$

By Assumption 1, there is some $\boldsymbol{\eta} \in \mathbb{R}^d$ such that $\boldsymbol{\Phi}\boldsymbol{\eta} = \mathbf{1}$. If $\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Phi} = \boldsymbol{\varphi}_s^{\mathsf{T}}$, then $\|\boldsymbol{\mu}\|_1 = \boldsymbol{\mu}^{\mathsf{T}}\mathbf{1} = \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Phi}\boldsymbol{\eta} = \boldsymbol{\varphi}_s^{\mathsf{T}}\boldsymbol{\eta} = 1$, which means that $\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\delta} \le \|\boldsymbol{\delta}\|_\infty$. Replacing $\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\delta}$ with $\|\boldsymbol{\delta}\|_\infty$ in the objective increases its value; we move the resulting constant term out of the maximization:

$$\le \|\boldsymbol{\delta}\|_\infty + \max \{ \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{\Phi}_*\boldsymbol{\theta} \mid \boldsymbol{\mu} \in \mathbb{R}^S_+,\ \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{Z}\boldsymbol{\Phi}_* = \boldsymbol{\varphi}_s^{\mathsf{T}} \}$$

The objective and constraints of this maximization problem depend on $\boldsymbol{\mu}$ only through $\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{Z}$. Thus we can replace $\boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{Z}$ with $\boldsymbol{\mu}_* \in \mathbb{R}^m_+$, which can only expand the feasible set of the maximization and increase its value:

$$\le \|\boldsymbol{\delta}\|_\infty + \max \{ \boldsymbol{\mu}_*^{\mathsf{T}}\boldsymbol{\Phi}_*\boldsymbol{\theta} \mid \boldsymbol{\mu}_* \in \mathbb{R}^m_+,\ \boldsymbol{\mu}_*^{\mathsf{T}}\boldsymbol{\Phi}_* = \boldsymbol{\varphi}_s^{\mathsf{T}} \}$$

The matrix $\boldsymbol{U} \in \{0,1\}^{m \times S}$ with rows $[\boldsymbol{U}_s]_{s \in \mathcal{S}_*} = \boldsymbol{e}_s$ can be used to "select" the core state features from $\boldsymbol{\Phi}$, giving $\boldsymbol{\Phi}_* = \boldsymbol{U}\boldsymbol{\Phi}$:

$$= \|\boldsymbol{\delta}\|_\infty + \max \{ \boldsymbol{\mu}_*^{\mathsf{T}}\boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{\theta} \mid \boldsymbol{\mu}_* \in \mathbb{R}^m_+,\ \boldsymbol{\mu}_*^{\mathsf{T}}\boldsymbol{U}\boldsymbol{\Phi} = \boldsymbol{\varphi}_s^{\mathsf{T}} \}$$

By a similar argument as before, we see that $\|\boldsymbol{\mu}_*^{\mathsf{T}}\boldsymbol{U}\|_1 = 1$. We add $\boldsymbol{\mu}_*^{\mathsf{T}}\boldsymbol{U}\boldsymbol{\delta} + \|\boldsymbol{\delta}\|_\infty \ge 0$ to the objective (increasing its value), then move the constant out:

$$\begin{aligned}
&\le 2\|\boldsymbol{\delta}\|_\infty + \max \{ \boldsymbol{\mu}_*^{\mathsf{T}}\boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{\theta} + \boldsymbol{\mu}_*^{\mathsf{T}}\boldsymbol{U}\boldsymbol{\delta} \mid \boldsymbol{\mu}_* \in \mathbb{R}^m_+,\ \boldsymbol{\mu}_*^{\mathsf{T}}\boldsymbol{U}\boldsymbol{\Phi} = \boldsymbol{\varphi}_s^{\mathsf{T}} \} \\
&= 2\|\boldsymbol{\delta}\|_\infty + \max \{ \boldsymbol{\mu}_*^{\mathsf{T}}\boldsymbol{U}\boldsymbol{v}^* \mid \boldsymbol{\mu}_* \in \mathbb{R}^m_+,\ \boldsymbol{\mu}_*^{\mathsf{T}}\boldsymbol{U}\boldsymbol{\Phi} = \boldsymbol{\varphi}_s^{\mathsf{T}} \} \\
&= 2\|\boldsymbol{\delta}\|_\infty + \min \{ \boldsymbol{\varphi}_s^{\mathsf{T}}\boldsymbol{\theta} \mid \boldsymbol{\theta} \in \mathbb{R}^d,\ \boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{\theta} \ge \boldsymbol{U}\boldsymbol{v}^* \},
\end{aligned}$$

where the last step is obtained by writing the dual of the linear program in the previous step. Now observe that the constraint $\boldsymbol{U}\boldsymbol{\Phi}\boldsymbol{\theta} \ge \boldsymbol{U}\boldsymbol{v}^*$ is equivalent to the constraint $\boldsymbol{W}_*\boldsymbol{E}\boldsymbol{\Phi}\boldsymbol{\theta} \ge \boldsymbol{W}_*\boldsymbol{E}\boldsymbol{v}^*$ in the definition of $J^*_{\mathrm{LRA}}$ — both of them require that $\boldsymbol{\varphi}_s\boldsymbol{\theta} \ge v^*_s$ for $s \in \mathcal{S}_*$. Thus we have shown that $J^*_{\mathrm{ALP}}(s) - J^*_{\mathrm{LRA}}(s) \le 2\varepsilon_{\mathrm{approx}}$ for all $s \in \mathcal{S}$. We also know that $J^*_{\mathrm{ALP}}(s) \ge J^*_{\mathrm{LRA}}(s)$, since $J^*_{\mathrm{LRA}}(s)$ is a relaxation of $J^*_{\mathrm{ALP}}(s)$. It follows that $\|J^*_{\mathrm{ALP}} - J^*_{\mathrm{LRA}}\|_\infty \le 2\varepsilon_{\mathrm{approx}}$. □

### A.2 Proof of Theorem 2 — Approximation Error for CoreLP

**Theorem 2** (CoreLP)**.** *Suppose Assumptions 1 and 2 hold, $s_0 \in \mathcal{S}$, $\boldsymbol{\varphi}_0 := \boldsymbol{\varphi}_{s_0}$, $\boldsymbol{W} \in \{0,1\}^{(1+m)A \times SA}$ has rows $[\boldsymbol{W}_{s_i a}]_{s_i \in \mathcal{S}_+, a \in \mathcal{A}} = \boldsymbol{e}_{s_i a}$, and $\Lambda := \{\boldsymbol{\lambda} \in \mathbb{R}^{(1+m)A}_+ \mid \sum_{a \in \mathcal{A}} \lambda_{s_0 a} = 1\}$. Define*

$$V^\dagger = \max \left\{ \boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{r} \;\middle|\; \boldsymbol{\lambda} \in \Lambda,\ \boldsymbol{\varphi}_0^{\mathsf{T}} + \boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{W}(\gamma\boldsymbol{P} - \boldsymbol{E})\boldsymbol{\Phi} = \mathbf{0} \right\}. \qquad \text{(CoreLP)}$$

*Let $\boldsymbol{\lambda}^\dagger \in \Lambda$ be a maximizer of (CoreLP) and let $\boldsymbol{\pi}^\dagger \in \Delta_{\mathcal{A}}$ be given by $\pi^\dagger(a) = \lambda^\dagger_{s_0 a}$. Then*

$$|V^\dagger - v^*(s_0)| \le \frac{10\gamma\varepsilon_{\mathrm{approx}}}{1-\gamma}, \qquad v^*(s_0) - \sum_{a \in \mathcal{A}} \pi^\dagger(a)\, q^*(s_0, a) \le \frac{20\gamma\varepsilon_{\mathrm{approx}}}{1-\gamma}.$$

By the definition of $\Lambda \subset \mathbb{R}^{(1+m)A}_+$, we can decompose its elements as $\boldsymbol{\lambda} = \boldsymbol{\pi} \oplus \boldsymbol{\lambda}_*$, with $\boldsymbol{\pi} \in \Delta_{\mathcal{A}}$ as in the statement of the theorem and $\boldsymbol{\lambda}_* \in \mathbb{R}^{mA}_+$ defined by $\lambda_{*,sa} = \lambda_{sa}$ for $s \in \mathcal{S}_*, a \in \mathcal{A}$ — in other words, $\Lambda \cong \Delta_{\mathcal{A}} \times \mathbb{R}^{mA}_+$. The main idea of the proof is that when $\boldsymbol{\lambda}$ is a solution of (CoreLP), then $\boldsymbol{\lambda}_*$ is a solution for the dual form of (LRALP$_\mu$) from Theorem 5. To make this connection between the two problems more precise, let us write the saddle-point forms of (LRALP$_\mu$) and (CoreLP):

$$V_{\mathrm{LRALP}}(\boldsymbol{\mu}) = \max_{\boldsymbol{\lambda}_* \in \mathbb{R}^{mA}_+} \min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left[ g_{\boldsymbol{\mu}}(\boldsymbol{\lambda}_*, \boldsymbol{\theta}) := \boldsymbol{\lambda}_*^{\mathsf{T}}\boldsymbol{W}_*\boldsymbol{r} + \boldsymbol{\mu}^{\mathsf{T}}\boldsymbol{\Phi}\boldsymbol{\theta} + \boldsymbol{\lambda}_*^{\mathsf{T}}\boldsymbol{W}_*(\gamma\boldsymbol{P} - \boldsymbol{E})\boldsymbol{\Phi}\boldsymbol{\theta} \right] \qquad \text{(Saddle LRALP}_\mu\text{)}$$

$$V^\dagger = \max_{\boldsymbol{\lambda} \in \Lambda} \min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left[ f(\boldsymbol{\lambda}, \boldsymbol{\theta}) := \boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{W}\boldsymbol{r} + \boldsymbol{e}_{s_0}^{\mathsf{T}}\boldsymbol{\Phi}\boldsymbol{\theta} + \boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{W}(\gamma\boldsymbol{P} - \boldsymbol{E})\boldsymbol{\Phi}\boldsymbol{\theta} \right] \qquad \text{(Saddle CoreLP)}$$

**Lemma 7** (Corresponding (LRALP$_\mu$) and (CoreLP) solutions)**.** *Let $\boldsymbol{\lambda} \in \Lambda \subset \mathbb{R}^{(1+m)A}_+$ be arbitrary and decompose it as $\boldsymbol{\lambda} = \boldsymbol{\pi} \oplus \boldsymbol{\lambda}_*$, where $\boldsymbol{\pi} \in \Delta_{\mathcal{A}}$ and $\boldsymbol{\lambda}_* \in \mathbb{R}^{mA}_+$. Define the distribution $\boldsymbol{\mu}_{\boldsymbol{\pi}} \in \Delta_{\mathcal{S}}$ as*

$$\boldsymbol{\mu}_{\boldsymbol{\pi}}^{\mathsf{T}} := \sum_{a \in \mathcal{A}} \pi(a)\, \boldsymbol{P}_{s_0 a}, \qquad \text{where } \pi(a) := \lambda_{s_0 a} \text{ for } a \in \mathcal{A}.$$

*Then, for any $\boldsymbol{\theta} \in \mathbb{R}^d$, and $g_{\boldsymbol{\mu}}(\boldsymbol{\lambda}_*, \boldsymbol{\theta})$ and $f(\boldsymbol{\lambda}, \boldsymbol{\theta})$ as in (Saddle LRALP$_\mu$) and (Saddle CoreLP),*

$$f(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \sum_{a \in \mathcal{A}} \pi(a)\, r(s_0, a) + g_{\gamma\boldsymbol{\mu}_{\boldsymbol{\pi}}}(\boldsymbol{\lambda}_*, \boldsymbol{\theta}), \qquad \text{where } \lambda_{*,sa} = \lambda_{sa} \text{ for } s \in \mathcal{S}_*, a \in \mathcal{A}.$$

$\boldsymbol{P}_{s_0 a}$ is the next-state distribution for action $a$ at state $s_0$ — thus the distribution $\boldsymbol{\mu}_{\boldsymbol{\pi}} \in \Delta_{\mathcal{S}}$ defined here is the expected next-state distribution when an action $a \sim \boldsymbol{\pi}$ is taken at state $s_0$. This lemma therefore connects solutions of (CoreLP) with (LRALP$_\mu$) when $\boldsymbol{\mu} = \gamma \boldsymbol{\mu}_{\boldsymbol{\pi}}$ is the discounted next-state distribution for action $a \sim \boldsymbol{\pi}$.

*Proof of Lemma 7.* Recall that $\boldsymbol{W}$ and $\boldsymbol{W}_*$ (defined in Theorems 2 and 5) are related — the rows of $\boldsymbol{W}_*$ correspond to state-action pairs in $\mathcal{S}_* \times \mathcal{A}$, to which $\boldsymbol{W}$ adds $A$ more rows corresponding to the actions at the current planning state $s_0$. Thus

$$\boldsymbol{\lambda}^\mathsf{T} \boldsymbol{W} = \sum_{a \in \mathcal{A}} \pi_a \boldsymbol{e}_{s_0 a}^\mathsf{T} + \boldsymbol{\lambda}_*^\mathsf{T} \boldsymbol{W}_*, \tag{8}$$

which upon multiplying by $\boldsymbol{r}$ gives

$$\boldsymbol{\lambda}^\mathsf{T} \boldsymbol{W} \boldsymbol{r} = \sum_{a \in \mathcal{A}} \pi_a r_{s_0 a} + \boldsymbol{\lambda}_*^\mathsf{T} \boldsymbol{W}_* \boldsymbol{r}. \tag{9}$$

Using (8) again,

$$\begin{aligned} \boldsymbol{e}_{s_0}^\mathsf{T} + \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{W}(\gamma \boldsymbol{P} - \boldsymbol{E}) &= \boldsymbol{e}_{s_0}^\mathsf{T} + \sum_{a \in \mathcal{A}} \pi_a (\gamma \boldsymbol{P}_{s_0 a} - \boldsymbol{E}_{s_0 a}) + \boldsymbol{\lambda}_*^\mathsf{T} \boldsymbol{W}_* (\gamma \boldsymbol{P} - \boldsymbol{E}), \\ &= \Big[\boldsymbol{e}_{s_0}^\mathsf{T} - \sum_{a \in \mathcal{A}} \pi_a \boldsymbol{E}_{s_0 a}\Big] + \gamma \Big[\sum_{a \in \mathcal{A}} \pi_a \boldsymbol{P}_{s_0 a}\Big] + \boldsymbol{\lambda}_*^\mathsf{T} \boldsymbol{W}_* (\gamma \boldsymbol{P} - \boldsymbol{E}). \end{aligned}$$

The first term is zero because $\boldsymbol{E}_{s_0 a} = \boldsymbol{e}_{s_0}$ for all $a \in \mathcal{A}$, and the second term becomes $\gamma \boldsymbol{\mu}_{\boldsymbol{\pi}}$ when we substitute the definition of $\boldsymbol{\mu}_{\boldsymbol{\pi}}$. We then multiply both sides by $\boldsymbol{\Phi\theta}$:

$$\boldsymbol{e}_{s_0}^\mathsf{T} \boldsymbol{\Phi\theta} + \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{W}(\gamma \boldsymbol{P} - \boldsymbol{E}) \boldsymbol{\Phi\theta} = \gamma \boldsymbol{\mu}_{\boldsymbol{\pi}}^\mathsf{T} \boldsymbol{\Phi\theta} + \boldsymbol{\lambda}_*^\mathsf{T} \boldsymbol{W}_* (\gamma \boldsymbol{P} - \boldsymbol{E}) \boldsymbol{\Phi\theta}.$$

Adding this to (9) gives

$$\boldsymbol{\lambda}^\mathsf{T} \boldsymbol{W} \boldsymbol{r} + \boldsymbol{e}_{s_0}^\mathsf{T} \boldsymbol{\Phi\theta} + \boldsymbol{\lambda}^\mathsf{T} \boldsymbol{W}(\gamma \boldsymbol{P} - \boldsymbol{E}) \boldsymbol{\Phi\theta} = \sum_{a \in \mathcal{A}} \pi_a r_{s_0 a} + \boldsymbol{\lambda}_*^\mathsf{T} \boldsymbol{W}_* \boldsymbol{r} + \gamma \boldsymbol{\mu}_{\boldsymbol{\pi}}^\mathsf{T} \boldsymbol{\Phi\theta} + \boldsymbol{\lambda}_*^\mathsf{T} \boldsymbol{W}_* (\gamma \boldsymbol{P} - \boldsymbol{E}) \boldsymbol{\Phi\theta},$$

where we substitute the definitions of $f(\boldsymbol{\lambda}, \boldsymbol{\theta})$ and $g_{\gamma \boldsymbol{\mu}_{\boldsymbol{\pi}}}(\boldsymbol{\lambda}_*, \boldsymbol{\theta})$ to get the desired result:

$$f(\boldsymbol{\lambda}, \boldsymbol{\theta}) = \sum_{a \in \mathcal{A}} \pi_a r_{s_0 a} + g_{\gamma \boldsymbol{\mu}_{\boldsymbol{\pi}}}(\boldsymbol{\lambda}_*, \boldsymbol{\theta}).$$

□

*Proof of Theorem 2.* Using the decomposition $\Lambda \cong \Delta_{\mathcal{A}} \times \mathbb{R}_+^{mA}$ in (Saddle CoreLP):

$$\begin{aligned} V^\dagger &= \max_{\boldsymbol{\pi} \in \Delta_{\mathcal{A}}} \max_{\boldsymbol{\lambda}_* \in \mathbb{R}_+^{mA}} \min_{\boldsymbol{\theta} \in \mathbb{R}^d} f(\boldsymbol{\pi} \oplus \boldsymbol{\lambda}_*, \boldsymbol{\theta}) && \text{(where } \boldsymbol{\pi} \oplus \boldsymbol{\lambda}_* = \boldsymbol{\lambda} \in \Lambda) \\ &= \max_{\boldsymbol{\pi} \in \Delta_{\mathcal{A}}} \max_{\boldsymbol{\lambda}_* \in \mathbb{R}_+^{mA}} \min_{\boldsymbol{\theta} \in \mathbb{R}^d} \Big[\sum_{a \in \mathcal{A}} \pi_a r_{s_0 a} + g_{\gamma \boldsymbol{\mu}_{\boldsymbol{\pi}}}(\boldsymbol{\lambda}_*, \boldsymbol{\theta})\Big] && \text{(using Lemma 7)} \\ &= \max_{\boldsymbol{\pi} \in \Delta_{\mathcal{A}}} \Big[\sum_{a \in \mathcal{A}} \pi_a r_{s_0 a} + \max_{\boldsymbol{\lambda}_* \in \mathbb{R}_+^{mA}} \min_{\boldsymbol{\theta} \in \mathbb{R}^d} g_{\gamma \boldsymbol{\mu}_{\boldsymbol{\pi}}}(\boldsymbol{\lambda}_*, \boldsymbol{\theta})\Big] \\ &= \max_{\boldsymbol{\pi} \in \Delta_{\mathcal{A}}} \Big[\, q^\dagger(\boldsymbol{\pi}) := \sum_{a \in \mathcal{A}} \pi_a r_{s_0 a} + V_{\text{LRALP}}(\gamma \boldsymbol{\mu}_{\boldsymbol{\pi}}) \,\Big]. && \text{(from (Saddle LRALP}_\mu)) \end{aligned}$$

We now turn our attention to bounding $V^\dagger$. From Theorem 5, we know that $|V_{\text{LRALP}}(\gamma \boldsymbol{\mu}_{\boldsymbol{\pi}}) - \gamma \boldsymbol{\mu}_{\boldsymbol{\pi}}^\mathsf{T} \boldsymbol{v}^*| \leq 10\gamma\varepsilon_{\text{approx}}/(1-\gamma)$ for any distribution over states $\boldsymbol{\mu}_{\boldsymbol{\pi}} \in \Delta_{\mathcal{S}}$. Through a slight abuse of notation, we define $q^*(s_0, \boldsymbol{\pi}) := \sum_a \pi_a r_{s_0 a} + \gamma \boldsymbol{\mu}_{\boldsymbol{\pi}}^\mathsf{T} \boldsymbol{v}^*$ as a generalization of the standard $q^*(s, a)$ value function to action distributions. Note that we will only need $q^*(s_0, \cdot)$, for which this abuse is 'sensible'. Then for all $\boldsymbol{\pi} \in \Delta_{\mathcal{A}}$,

$$|q^\dagger(\boldsymbol{\pi}) - q^*(s_0, \boldsymbol{\pi})| = |V_{\text{LRALP}}(\gamma \boldsymbol{\mu}_{\boldsymbol{\pi}}) - \gamma \boldsymbol{\mu}_{\boldsymbol{\pi}}^\mathsf{T} \boldsymbol{v}^*| \leq \frac{10\gamma\varepsilon_{\text{approx}}}{1-\gamma}. \tag{10}$$

We also know that $v^*(s_0) = \max_{\boldsymbol{\pi} \in \Delta_{\mathcal{A}}} q^*(s_0, \boldsymbol{\pi})$ (the equality happens with $\boldsymbol{\pi}^* = \boldsymbol{e}_{a^*}$ for an optimal action $a^*$). Hence,

$$|V^\dagger - v^*(s_0)| = |\max_{\boldsymbol{\pi} \in \Delta_{\mathcal{A}}} q^\dagger(\boldsymbol{\pi}) - \max_{\boldsymbol{\pi} \in \Delta_{\mathcal{A}}} q^*(s_0, \boldsymbol{\pi})| \leq \max_{\boldsymbol{\pi} \in \Delta_{\mathcal{A}}} |q^\dagger(\boldsymbol{\pi}) - q^*(s_0, \boldsymbol{\pi})| \leq \frac{10\gamma\varepsilon_{\text{approx}}}{1-\gamma},$$

where the last inequality follows from (10).

For the second part of the result, let $\boldsymbol{\lambda}^\dagger$ be a maximizer of (CoreLP) and $\boldsymbol{\pi}^\dagger$ be the action-distribution component (as before) so that $V^\dagger = q^\dagger(\boldsymbol{\pi}^\dagger)$. Then, using again (10), combined with the last inequality,

$$\begin{aligned}\sum_{a\in\mathcal{A}} \pi^\dagger(a)\, q^*(s_0,a) \equiv q^*(s_0,\boldsymbol{\pi}^\dagger) &\geq q^\dagger(\boldsymbol{\pi}^\dagger) - \frac{10\gamma\varepsilon_{\text{approx}}}{1-\gamma}\\ &= V^\dagger - \frac{10\gamma\varepsilon_{\text{approx}}}{1-\gamma}\\ &\geq v^*(s_0) - \frac{20\gamma\varepsilon_{\text{approx}}}{1-\gamma}.\end{aligned}$$

Reordering gives the desired result, namely that $v^*(s_0) - \sum_a \pi^\dagger(a)\, q^*(s_0,a) \leq 20\gamma\varepsilon_{\text{approx}}/(1-\gamma)$. □

### A.3 Proof of Theorem 3 — Error Bounds for the CoreStoMP Algorithm

**Theorem 3** (CoreStoMP)**.** *Suppose Assumptions 1, 2, and 3 hold, and define*

$$B := \frac{(9/8)\sqrt{m}}{1-\gamma}, \qquad\qquad C := \frac{(9/4)\sqrt{m(1+2\log A + 2\gamma\log m)}}{(1-\gamma)^2}.$$

*Let $\hat{\boldsymbol{\lambda}}$ be the result of running Algorithm 1 for $T$ iterations with the parameter $B$ and the step size $\eta = C^{-1}\sqrt{2/7T}$, which requires $2T(1+(1+m)A)$ simulator queries. Define $\hat{\boldsymbol{\pi}} \in \Delta_{\mathcal{A}}$ by $\hat{\pi}(a) = \hat{\lambda}_{s_0 a}$ (as in Theorem 2) and $a \sim \hat{\boldsymbol{\pi}}$. Then*

$$v^*(s_0) - \mathbb{E}[q^*(s_0,a)] \leq \frac{32\varepsilon_{\text{approx}}}{1-\gamma} + \frac{21}{2(1-\gamma)^2}\sqrt{\frac{3m(1+2\log A + 2\gamma\log m)}{T}}.$$

The proof of this theorem has two main ingredients: First, in Lemma 8, we show that approximate solutions of (Saddle CoreLP) can be used to recover near-optimal action distributions for the planning state $s_0$ — the approximation quality is measured by the *duality gap*. Second, in Lemma 11, we bound the expected duality gap of the Stochastic Mirror-Prox algorithm when specialized to our setting.

**Lemma 8** (Approximate (Saddle CoreLP) solutions)**.** *Suppose $\mathcal{B} \subset \mathbb{R}^d$ and $C_{\mathcal{B}} \geq 0$ are chosen such that, for any distribution over states $\boldsymbol{\mu} \in \Delta_{\mathcal{S}}$, there is some $\boldsymbol{\theta} \in \mathcal{B}$ that is feasible for (LRALP$_{\mu}$) and at most $C_{\mathcal{B}}$-suboptimal. Define*

$$\Lambda_\gamma := \{\boldsymbol{\lambda} \in \Lambda \mid \|\boldsymbol{\lambda}\|_1 = 1/(1-\gamma)\}, \tag{11}$$

*a subset of the set $\Lambda \subset \mathbb{R}_+^{(1+m)A}$ from Theorem 2. Define the $\mathcal{B}$-*bounded duality gap *of an approximate solution of (Saddle CoreLP) as*

$$\delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\theta}}) := \max_{\boldsymbol{\lambda}\in\Lambda_\gamma} f(\boldsymbol{\lambda},\hat{\boldsymbol{\theta}}) - \inf_{\boldsymbol{\theta}\in\mathcal{B}} f(\hat{\boldsymbol{\lambda}},\boldsymbol{\theta}), \qquad \textit{where } \hat{\boldsymbol{\lambda}} \in \Lambda \textit{ and } \hat{\boldsymbol{\theta}} \in \mathbb{R}^d. \tag{12}$$

*For any $\hat{\boldsymbol{\lambda}} \in \Lambda$ and $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$, let $\hat{\boldsymbol{\pi}}$ be the action distribution component of $\hat{\boldsymbol{\lambda}}$, as in Theorem 2. Then*

$$v^*(s_0) - \sum_{a\in\mathcal{A}} \hat{\pi}(a)\, q^*(s_0,a) \leq \frac{20\gamma\varepsilon_{\text{approx}}}{1-\gamma} + \gamma C_{\mathcal{B}} + \delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\theta}}).$$

This lemma generalizes the second result of Theorem 2 in two ways: First, the Stochastic Mirror-Prox algorithm does not produce exact solutions of (Saddle CoreLP); the optimization error is measured by the duality gap — here we see the effect of a non-zero duality gap on the resulting action distribution. Second, the primal variables $\boldsymbol{\theta}$ in (Saddle CoreLP) have the unbounded domain $\mathbb{R}^d$, whereas the Stochastic Mirror-Prox algorithm requires the optimization domain to have a bounded *diameter*; see Proof of Lemma 11 — Stochastic Mirror-Prox. This lemma shows that restricting $\boldsymbol{\theta}$ to a large-enough bounded set $\mathcal{B}$ only incurs an additional $C_{\mathcal{B}}$ error. Indeed, the second issue is related to the first — an *unbounded* form of the duality gap would be infinite for *any* approximate solution, making it useless as a measure of optimization accuracy; the $\mathcal{B}$-bounded duality gap therefore addresses both these issues:

*Claim* 9. For any $\hat{\boldsymbol{\lambda}} \in \Lambda$, $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$, $\boldsymbol{\theta} \in \mathcal{B} \subset \mathbb{R}^d$, and $\delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\theta}})$ being the $\mathcal{B}$-bounded duality gap (12),

$$V^\dagger \leq f(\hat{\boldsymbol{\lambda}},\boldsymbol{\theta}) + \delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}},\hat{\boldsymbol{\theta}}).$$

*Proof.* Let $\boldsymbol{\lambda}^* \in \Lambda \subset \mathbb{R}_+^{(1+m)A}$ be a maximizer of (CoreLP) — this exists because the optimization is bounded (Theorem 2). Then

$$\begin{aligned}\mathbf{0} &= \boldsymbol{\varphi}_0^\mathsf{T} + \boldsymbol{\lambda}^{*\mathsf{T}}\boldsymbol{W}(\gamma\boldsymbol{P}-\boldsymbol{E})\boldsymbol{\Phi} && \text{(since } \boldsymbol{\lambda}^* \text{ is feasible for (CoreLP))} \quad (13)\\ &= \boldsymbol{\varphi}_0^\mathsf{T}\hat{\boldsymbol{\theta}} + \boldsymbol{\lambda}^{*\mathsf{T}}\boldsymbol{W}(\gamma\boldsymbol{P}-\boldsymbol{E})\boldsymbol{\Phi}\hat{\boldsymbol{\theta}}. && \text{(multiplying by } \hat{\boldsymbol{\theta}} \in \mathbb{R}^d)\end{aligned}$$

Since, $\boldsymbol{\lambda}^*$ is a maximizer of (CoreLP), $V^\dagger = \boldsymbol{\lambda}^{*\mathsf{T}}\boldsymbol{W}\boldsymbol{r}$:

$$\begin{aligned}V^\dagger &= \boldsymbol{\lambda}^{*\mathsf{T}}\boldsymbol{W}\boldsymbol{r} + \boldsymbol{\varphi}_0^\mathsf{T}\hat{\boldsymbol{\theta}} + \boldsymbol{\lambda}^{*\mathsf{T}}\boldsymbol{W}(\gamma\boldsymbol{P}-\boldsymbol{E})\boldsymbol{\Phi}\hat{\boldsymbol{\theta}} && \text{(adding } V^\dagger \text{ on l.h.s. and } \boldsymbol{\lambda}^{*\mathsf{T}}\boldsymbol{W}\boldsymbol{r} \text{ on r.h.s.)}\\ &= f(\boldsymbol{\lambda}^*, \hat{\boldsymbol{\theta}}). && \text{(definition of } f \text{ from (Saddle CoreLP))} \quad (14)\end{aligned}$$

Assumption 1 tells us that $\boldsymbol{\Phi}\boldsymbol{\eta} = \mathbf{1}$ for some $\boldsymbol{\eta} \in \mathbb{R}^d$ — multiplying (13) by $\boldsymbol{\eta}$, we see that $\boldsymbol{\lambda}^*$ must satisfy $1 + \gamma\|\boldsymbol{\lambda}^*\|_1 = \|\boldsymbol{\lambda}^*\|_1$, as does any other feasible solution of (CoreLP). In particular, this means that $\|\boldsymbol{\lambda}^*\|_1 = 1/(1-\gamma)$ and so $\boldsymbol{\lambda}^* \in \Lambda_\gamma$. Using the definition of $\delta_{\mathcal{B}}$ from (12),

$$\begin{aligned}\delta(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}}) &\geq f(\boldsymbol{\lambda}^*, \hat{\boldsymbol{\theta}}) - f(\hat{\boldsymbol{\lambda}}, \boldsymbol{\theta}) && \text{(since } \boldsymbol{\lambda}^* \in \Lambda_\gamma \text{ and } \boldsymbol{\theta} \in \mathcal{B})\\ &= V^\dagger - f(\hat{\boldsymbol{\lambda}}, \boldsymbol{\theta}). && \text{(using (14))}\end{aligned}$$

□

*Claim* 10. For any $\hat{\boldsymbol{\lambda}}_* \in \mathbb{R}_+^{mA}$ and distribution over states $\boldsymbol{\mu} \in \Delta_{\mathcal{S}}$, suppose $\boldsymbol{\theta} \in \mathbb{R}^d$ is feasible for (LRALP$_\mu$) and at most $C_{\mathcal{B}}$-suboptimal. Then, with $g_\mu$ being the objective function of (Saddle LRALP$_\mu$),

$$V_{\text{LRALP}}(\gamma\boldsymbol{\mu}) \geq g_{\gamma\mu}(\hat{\boldsymbol{\lambda}}_*, \boldsymbol{\theta}) - \gamma C_{\mathcal{B}}.$$

*Proof.* Since $\boldsymbol{\theta}$ is feasible for (LRALP$_\mu$), $\boldsymbol{W}_*\boldsymbol{r} + \boldsymbol{W}_*(\gamma\boldsymbol{P}-\boldsymbol{E})\boldsymbol{\Phi}\boldsymbol{\theta} \leq \mathbf{0}$. Multiplying both sides of this inequality by $\hat{\boldsymbol{\lambda}}_*^\mathsf{T} \geq \mathbf{0}$,

$$\begin{aligned}\hat{\boldsymbol{\lambda}}_*^\mathsf{T}\boldsymbol{W}_*\boldsymbol{r} + \hat{\boldsymbol{\lambda}}_*^\mathsf{T}\boldsymbol{W}_*(\gamma\boldsymbol{P}-\boldsymbol{E})\boldsymbol{\Phi}\boldsymbol{\theta} &\leq 0\\ \hat{\boldsymbol{\lambda}}_*^\mathsf{T}\boldsymbol{W}_*\boldsymbol{r} + \gamma\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi}\boldsymbol{\theta} + \hat{\boldsymbol{\lambda}}_*^\mathsf{T}\boldsymbol{W}_*(\gamma\boldsymbol{P}-\boldsymbol{E})\boldsymbol{\Phi}\boldsymbol{\theta} &\leq \gamma\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi}\boldsymbol{\theta} && \text{(adding } \gamma\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi}\boldsymbol{\theta} \text{ to both sides)}\\ g_{\gamma\mu}(\hat{\boldsymbol{\lambda}}_*, \boldsymbol{\theta}) &\leq \gamma\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi}\boldsymbol{\theta}. && \text{(definition of } g \text{ from (Saddle LRALP}_\mu))\end{aligned}$$

Note that the choice of $\boldsymbol{\mu}$ does not affect the constraints of (LRALP$_\mu$), only its objective function — thus $\boldsymbol{\theta}$ is feasible for the problem defining $V_{\text{LRALP}}(\gamma\boldsymbol{\mu})$ and is $\gamma C_{\mathcal{B}}$ suboptimal: $\gamma\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi}\boldsymbol{\theta} \leq V_{\text{LRALP}}(\gamma\boldsymbol{\mu}) + \gamma C_{\mathcal{B}}$. Substituting this into the last inequality and rearranging gives the desired result. □

*Proof of Lemma 8.* As in Lemma 7, we write $\hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\pi}} \oplus \hat{\boldsymbol{\lambda}}_*$ with $\hat{\boldsymbol{\pi}} \in \Delta_{\mathcal{A}}$ and $\hat{\boldsymbol{\lambda}}_* \in \mathbb{R}_+^{mA}$ and define $\boldsymbol{\mu}_{\hat{\pi}} = \sum_a \hat{\pi}_a \boldsymbol{P}_{s_0 a}$. By our assumption, there is some $\boldsymbol{\theta} \in \mathcal{B}$ that is feasible and at most $C_{\mathcal{B}}$-suboptimal for (LRALP$_\mu$) with $\boldsymbol{\mu} = \boldsymbol{\mu}_{\hat{\pi}}$ — this allows us to apply Claims 9 and 10 below:

$$\begin{aligned}v^*(s_0) &\leq V^\dagger + \frac{10\gamma\varepsilon_{\text{approx}}}{1-\gamma} && \text{(Theorem 2)}\\ &\leq f(\hat{\boldsymbol{\lambda}}, \boldsymbol{\theta}) + \delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}}) + \frac{10\gamma\varepsilon_{\text{approx}}}{1-\gamma} && \text{(Claim 9)}\\ &= \sum_{a\in\mathcal{A}} \hat{\pi}_a r_{s_0 a} + g_{\gamma\mu_{\hat{\pi}}}(\hat{\boldsymbol{\lambda}}_*, \boldsymbol{\theta}) + \delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}}) + \frac{10\gamma\varepsilon_{\text{approx}}}{1-\gamma} && \text{(Lemma 7)}\\ &\leq \sum_{a\in\mathcal{A}} \hat{\pi}_a r_{s_0 a} + V_{\text{LRALP}}(\gamma\boldsymbol{\mu}_{\hat{\pi}}) + \gamma C_{\mathcal{B}} + \delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}}) + \frac{10\gamma\varepsilon_{\text{approx}}}{1-\gamma} && \text{(Claim 10)}\\ &\leq \sum_{a\in\mathcal{A}} \hat{\pi}_a r_{s_0 a} + \gamma\boldsymbol{\mu}_{\hat{\pi}}^\mathsf{T}\boldsymbol{v}^* + \gamma C_{\mathcal{B}} + \delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}}) + \frac{20\gamma\varepsilon_{\text{approx}}}{1-\gamma} && \text{(Theorem 5)}\\ &= \sum_{a\in\mathcal{A}} \hat{\pi}_a q^*(s_0, a) + \gamma C_{\mathcal{B}} + \delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}}) + \frac{20\gamma\varepsilon_{\text{approx}}}{1-\gamma},\end{aligned}$$

where the last step used $\sum_a \hat{\pi}_a r_{s_0 a} + \gamma\boldsymbol{\mu}_{\hat{\pi}}^\mathsf{T}\boldsymbol{v}^* = \sum_a \hat{\pi}_a(r_{s_0 a} + \gamma\boldsymbol{P}_{s_0 a}\boldsymbol{v}^*) = \sum_a \hat{\pi}_a q^*(s_0, a)$. Rearranging the inequality completes the proof:

$$v^*(s_0) - \sum_{a\in\mathcal{A}} \hat{\pi}_a q^*(s_0, a) \leq \frac{20\gamma\varepsilon_{\text{approx}}}{1-\gamma} + \gamma C_{\mathcal{B}} + \delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}}).$$

□

The following lemma bounds the expected duality gap of the Stochastic Mirror-Prox algorithm when applied to our setting. We defer the proof to Appendix A.4.

**Lemma 11** (Stochastic Mirror-Prox)**.** *Using the constants $B$ and $C$ from Theorem 3, define*

$$\mathcal{B} := \{\boldsymbol{\theta} \in \mathbb{R}^d \mid \|\boldsymbol{\Phi}_* \boldsymbol{\theta}\|_2 \le B\}, \tag{15}$$

*and let $\delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})$ be the $\mathcal{B}$-bounded duality gap (12). Then the results of running Algorithm 1 for $T$ iterations satisfy*

$$\varepsilon_{\text{opt}} := \mathbb{E}[\delta(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})] \le \frac{14C}{\sqrt{3T}}.$$

*Proof of Theorem 3.* First, observe that $v^*(s_0) - q^*(s_0, a) \le 2/(1-\gamma)$ for any action $a$, since all the rewards lie in $[-1, 1]$ by Assumption 3. Thus, if $\varepsilon_{\text{approx}} > 1/16$ then $32\varepsilon_{\text{approx}}/(1-\gamma) > 2/(1-\gamma)$ and the result is trivially true. From now on, we will assume that $\varepsilon_{\text{approx}} \le 1/16$.

To prove our result, we will combine Lemmas 8 and 11, for which we need to show that the set $\mathcal{B}$ defined in (15) satisfies the requirements of Lemma 8. Specifically, we need to show that $\mathcal{B}$ contains a feasible solution of the linear program (LRALP$_\mu$) with sub-optimality bounded by a constant $C_{\mathcal{B}}$. Note that the constraints of (LRALP$_\mu$) do not depend on $\boldsymbol{\mu}$, only the objective function, so the choice of $\boldsymbol{\mu}$ does not affect feasibility.

Since (LRALP$_\mu$) is a relaxation of the ALP (see Section 1.2), any feasible solution of the ALP is also feasible for (LRALP$_\mu$). De Farias and Van Roy [14, Theorem 2] show that the ALP has a feasible solution $\boldsymbol{\Phi\theta}$ that is close to $\boldsymbol{v}^*$ — more precisely

$$\|\boldsymbol{\Phi\theta} - \boldsymbol{v}^*\|_\infty \le \frac{2\varepsilon_{\text{approx}}}{1-\gamma}. \tag{16}$$

Since $\|\boldsymbol{v}^*\|_\infty \le 1/(1-\gamma)$, we must have $\|\boldsymbol{\Phi\theta}\|_\infty \le (1 + 2\varepsilon_{\text{approx}})/(1-\gamma)$. It follows that

$$\|\boldsymbol{\Phi}_*\boldsymbol{\theta}\|_2 \le \sqrt{m}\|\boldsymbol{\Phi}_*\boldsymbol{\theta}\|_\infty = \sqrt{m}\|\boldsymbol{\Phi\theta}\|_\infty \le \frac{(1+2\varepsilon_{\text{approx}})\sqrt{m}}{1-\gamma} \le \frac{(9/8)\sqrt{m}}{1-\gamma} = B,$$

where the first inequality is a property of the 2-norm, the next equality is thanks to Assumption 2, and the last inequality is because we assumed $\varepsilon_{\text{approx}} \le 1/16$ — this shows that $\boldsymbol{\theta} \in \mathcal{B}$. We also have

$$\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi\theta} - \boldsymbol{\mu}^\mathsf{T}\boldsymbol{v}^* \le \|\boldsymbol{\mu}\|_1\|\boldsymbol{\Phi\theta} - \boldsymbol{v}^*\|_\infty \le \frac{2\varepsilon_{\text{approx}}}{1-\gamma} \qquad \text{(using (16) and } \|\boldsymbol{\mu}\|_1 = 1)$$

$$|\boldsymbol{\mu}^\mathsf{T}\boldsymbol{v}^* - V_{\text{LRALP}}(\boldsymbol{\mu})| \le \frac{10\varepsilon_{\text{approx}}}{1-\gamma} \qquad \text{(using Theorem 5 and } \|\boldsymbol{\mu}\|_1 = 1)$$

We get the value of $C_{\mathcal{B}}$ by putting these two bounds together:

$$\boldsymbol{\mu}^\mathsf{T}\boldsymbol{\Phi\theta} - V_{\text{LRALP}}(\boldsymbol{\mu}) \le \frac{12\varepsilon_{\text{approx}}}{1-\gamma} =: C_{\mathcal{B}}.$$

Lemma 8 applied to $\mathcal{B}$ with this value of $C_{\mathcal{B}}$ gives

$$v^*(s_0) - \sum_{a \in \mathcal{A}} \hat{\pi}(a)\, q^*(s_0, a) \le \frac{32\gamma\varepsilon_{\text{approx}}}{1-\gamma} + \delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}}).$$

Taking expectations on both sides and substituting the value of $\varepsilon_{\text{opt}} = \mathbb{E}[\delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})]$ from Lemma 11,

$$v^*(s_0) - \mathbb{E}[q^*(s_0, a)] \le \frac{32\gamma\varepsilon_{\text{approx}}}{1-\gamma} + \frac{14C}{\sqrt{3T}}.$$

We drop the $\gamma$ factor from the leading term and plug in the value of $C$ from the statement of Theorem 3 to finish the proof. □

### A.4 Proof of Lemma 11 — Stochastic Mirror-Prox

**Lemma 11** (Stochastic Mirror-Prox)**.** *Using the constants $B$ and $C$ from Theorem 3, define*

$$\mathcal{B} := \{\boldsymbol{\theta} \in \mathbb{R}^d \mid \|\boldsymbol{\Phi}_* \boldsymbol{\theta}\|_2 \le B\}, \tag{15}$$

*and let $\delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})$ be the $\mathcal{B}$-bounded duality gap (12). Then the results of running Algorithm 1 for $T$ iterations satisfy*

$$\varepsilon_{\text{opt}} := \mathbb{E}[\delta(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})] \le \frac{14C}{\sqrt{3T}}.$$

Throughout this section, we will use the definitions of $\Lambda_\gamma$ and $\mathcal{B}$ from (11) and (15), respectively. We will also define the composite space $Z := \Lambda_\gamma \times \mathcal{B}$. We will use the norm $\|\boldsymbol{\lambda}\|_1$ for $\boldsymbol{\lambda} \in \Lambda_\gamma$, whose dual norm is $\|\cdot\|_\infty$. For $\boldsymbol{\theta} \in \mathcal{B}$ we will use the norm $\|\boldsymbol{\theta}\| \equiv \|\boldsymbol{\Phi\theta}\|_2$ — the corresponding dual norm enjoys the convenient bound $\|\boldsymbol{\Phi}_*{}^{\mathsf{T}}\boldsymbol{u}\|_* = \sup_{\|\boldsymbol{\theta}\|\leq 1} \boldsymbol{u}^{\mathsf{T}}\boldsymbol{\Phi}_*\boldsymbol{\theta} \leq \|\boldsymbol{u}\|_2 \leq \|\boldsymbol{u}\|_1$ for any vector $\boldsymbol{u}$.[4] The last inequality is due a general property of $p$-norms: $\|\boldsymbol{u}\|_p \leq \|\boldsymbol{u}\|_q$ whenever $\infty \geq p \geq q \geq 1$.

### A.4.1 Lipschitz Constants

Our first step will be to bound the Lipschitz constants associated with $f$, the objective function of (CoreLP). In other words, we are looking for bounds on $\|f_\lambda(\boldsymbol{\theta})\|_\infty$ and $\|f_\theta(\boldsymbol{\lambda})\|_*$.

First, for any $\boldsymbol{\theta} \in \mathcal{B}$ we have

$$\begin{aligned}
\|f_\lambda(\boldsymbol{\theta})\|_\infty &= \|\boldsymbol{Wr} + \boldsymbol{W}(\gamma\boldsymbol{P} - \boldsymbol{E})\boldsymbol{\Phi\theta}\|_\infty \\
&\leq \|\boldsymbol{Wr}\|_\infty + \|\boldsymbol{W}(\gamma\boldsymbol{P} - \boldsymbol{E})\boldsymbol{\Phi\theta}\|_\infty && \text{(by the triangle inequality)} \\
&\leq 1 + \max_i \|\boldsymbol{W}_i(\gamma\boldsymbol{P} - \boldsymbol{E})\|_1 \|\boldsymbol{\Phi\theta}\|_\infty. && \text{(by definition of } \|\cdot\|_\infty \text{ and Hölder's inequality)}
\end{aligned}$$

Now, by the property of norms, $\|\boldsymbol{\Phi\theta}\|_\infty \leq \|\boldsymbol{\Phi\theta}\|_2 \leq B$. Secondly, $\boldsymbol{W}_i\boldsymbol{P}$ and $\boldsymbol{W}_i\boldsymbol{E}$ are probability distributions, so $\|\boldsymbol{W}_i(\gamma\boldsymbol{P} - \boldsymbol{E})\|_1 \leq \gamma\|\boldsymbol{W}_i\boldsymbol{P}\|_1 + \|\boldsymbol{W}_i\boldsymbol{E}\|_1 = 1 + \gamma$ and

$$\|f_\lambda(\boldsymbol{\theta})\|_\infty \leq 1 + (1+\gamma)B \leq 2B. \qquad \text{(since } B \geq 1/(1-\gamma)\text{)}$$

For the other gradient, we use the bound on dual norms mentioned above:

$$\begin{aligned}
\|f_\theta(\boldsymbol{\lambda})\|_* &= \|(\boldsymbol{\varphi}_{s_0}^{\mathsf{T}} + \boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{W}(\gamma\boldsymbol{P} - \boldsymbol{E}))\boldsymbol{\Phi}\|_* \\
&\leq \|\boldsymbol{e}_{s_0}^{\mathsf{T}}\|_1 + \|\boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{W}(\gamma\boldsymbol{P} - \boldsymbol{E})\|_1 \\
&\leq 1 + \frac{1+\gamma}{1-\gamma} = \frac{2}{1-\gamma},
\end{aligned}$$

where the last inequality uses the fact that $(1-\gamma)\boldsymbol{\lambda}$ is a probability distribution, as are the rows of $\boldsymbol{W}$, $\boldsymbol{P}$, and $\boldsymbol{E}$.

### A.4.2 Gradient Estimator Variance

Next, we will bound the variance in the stochastic estimators $\hat{f}_\lambda(\boldsymbol{\theta})$ and $\hat{f}_\theta(\boldsymbol{\lambda})$ defined in (6) and (7), respectively, compared to the true gradients $f_\lambda(\boldsymbol{\theta})$ and $f_\theta(\boldsymbol{\lambda})$ defined in (4) and (5), respectively.

First, we bound $\mathbb{E}[\|\hat{f}_\lambda(\boldsymbol{\theta}) - f_\lambda(\boldsymbol{\theta})\|_\infty^2]$ for any $\boldsymbol{\theta} \in \mathcal{B}$ by bounding its components. For any state $s \in \mathcal{S}_+$, action $a \in \mathcal{A}$, and reward $\hat{r}$, we have

$$\begin{aligned}
&|[\hat{f}_\lambda(\boldsymbol{\theta})]_{sa} - [f_\lambda(\boldsymbol{\theta})]_{sa}| \\
&\quad = |(\hat{r} + \gamma\boldsymbol{\varphi}_{s'}^{\mathsf{T}}\boldsymbol{\theta} - \boldsymbol{\varphi}_s^{\mathsf{T}}\boldsymbol{\theta}) - (r_{sa} + \gamma\boldsymbol{P}_{sa}\boldsymbol{\Phi\theta} - \boldsymbol{\varphi}_s^{\mathsf{T}}\boldsymbol{\theta})| && \text{(for some random } s' \sim \boldsymbol{P}_{sa}\text{)} \\
&\quad \leq |\hat{r} - r_{sa}| + \gamma|\boldsymbol{\varphi}_{s'}^{\mathsf{T}}\boldsymbol{\theta} - \boldsymbol{P}_{sa}\boldsymbol{\Phi\theta}| && \text{(by the triangle inequality)} \\
&\quad \leq 2 + \gamma|(\boldsymbol{e}_{s'}^{\mathsf{T}} - \boldsymbol{P}_{sa})\boldsymbol{\Phi\theta}| && \text{(bounded rewards)} \\
&\quad \leq 2 + \gamma\|\boldsymbol{e}_{s'}^{\mathsf{T}} - \boldsymbol{P}_{sa}\|_1\|\boldsymbol{\Phi\theta}\|_\infty && \text{(using Hölder's inequality)} \\
&\quad \leq 2 + 2\gamma B && \text{(since } \|\boldsymbol{\Phi\theta}\|_\infty \leq \|\boldsymbol{\Phi\theta}\|_2 \leq B\text{)} \\
&\quad \leq 2B. && \text{(since } B \geq 1/(1-\gamma)\text{)}
\end{aligned}$$

It follows that $\|\hat{f}_\lambda(\boldsymbol{\theta}) - f_\lambda(\boldsymbol{\theta})\|_\infty^2 \leq (2B)^2$, and the same bound must hold for the expectation.

We will now bound the other gradient, using the following property of Euclidean norms: for any vector-valued random variable $\boldsymbol{h}$ with mean $\bar{\boldsymbol{h}}$, $\mathbb{E}[\|\boldsymbol{h} - \bar{\boldsymbol{h}}\|^2] = \mathbb{E}[\|\boldsymbol{h}\|^2] - \|\bar{\boldsymbol{h}}\|^2 \leq \mathbb{E}[\|\boldsymbol{h}\|^2]$. Then, for a random choice of state $s \in \mathcal{S}_+$, action $a \in \mathcal{A}$, and next state $s' \sim \boldsymbol{P}_{sa}$:

$$\begin{aligned}
\mathbb{E}[\|\hat{f}_\theta(\boldsymbol{\lambda}) - f_\theta(\boldsymbol{\lambda})\|_*^2] &= \mathbb{E}\big[\big\|(\boldsymbol{\varphi}_0^{\mathsf{T}} + \|\boldsymbol{\lambda}\|_1(\gamma\boldsymbol{\varphi}_{s'} - \boldsymbol{\varphi}_s)) - (\boldsymbol{\varphi}_0^{\mathsf{T}} + \boldsymbol{\lambda}^{\mathsf{T}}\boldsymbol{W}(\gamma\boldsymbol{P} - \boldsymbol{E})\boldsymbol{\Phi})\big\|_*^2\big] \\
&= \left(\frac{1}{1-\gamma}\right)^2 \mathbb{E}\big[\big\|(\gamma\boldsymbol{\varphi}_{s'} - \boldsymbol{\varphi}_s) - (\boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|_1)^{\mathsf{T}}\boldsymbol{W}(\gamma\boldsymbol{P} - \boldsymbol{E})\boldsymbol{\Phi}\big\|_*^2\big]
\end{aligned}$$

[4] More generally, $\|\boldsymbol{\xi}\|_* = \inf\{\|\boldsymbol{\eta}\|_2 \mid \boldsymbol{\xi}^{\mathsf{T}} = \boldsymbol{\eta}^{\mathsf{T}}\boldsymbol{\Phi}_*\}$, which is non-zero when $\boldsymbol{\xi} \neq \boldsymbol{0}$ and $\boldsymbol{\Phi}_*$ has full column rank.

Now, since $(s, a) \sim \boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|_1$ and $s' \sim \boldsymbol{P}_{sa}$, we have $\mathbb{E}[\gamma\boldsymbol{\varphi}_{s'} - \boldsymbol{\varphi}_s] = (\boldsymbol{\lambda}/\|\boldsymbol{\lambda}\|_1)^{\mathsf{T}}\boldsymbol{W}(\gamma\boldsymbol{P} - \boldsymbol{E})\boldsymbol{\Phi}$, so by the above property of variance for vector-valued random variables,

$$\begin{aligned}
&\leq \left(\frac{1}{1-\gamma}\right)^2 \mathbb{E}[\|\gamma\boldsymbol{\varphi}_{s'} - \boldsymbol{\varphi}_s\|_*^2] \\
&= \left(\frac{1}{1-\gamma}\right)^2 \mathbb{E}[\|(\gamma\boldsymbol{e}_{s'} - \boldsymbol{e}_s)\boldsymbol{\Phi}\|_*^2] \\
&\leq \left(\frac{1+\gamma}{1-\gamma}\right)^2 \leq \left(\frac{2}{1-\gamma}\right)^2.
\end{aligned}$$

### A.4.3 Distance-Generating Functions

The Stochastic Mirror-Prox algorithm requires strongly convex *distance-generating functions* for $\Lambda_\gamma$ and $\mathcal{B}$ with respect to their respective norms. A function $\omega : \mathcal{X} \to \mathbb{R}$ (with domain $\mathcal{X} \subset \mathbb{R}^n$) is said to be $\sigma$-*strongly convex* (where $\sigma > 0$ is called the modulus of convexity) with respect to a norm $\|\cdot\|$ on $\mathcal{X}$ if any of the following conditions hold for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$

(i) For all $\alpha \in [0, 1]$, $\alpha\omega(\boldsymbol{x}) + (1-\alpha)\omega(\boldsymbol{y}) \geq \omega(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) + \sigma\alpha(1-\alpha)\|x - y\|^2/2$.

(ii) $\omega$ is convex and $\omega(\boldsymbol{x}) \geq \omega(\boldsymbol{y}) + \langle\nabla\omega(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle + \sigma\|\boldsymbol{x} - \boldsymbol{y}\|^2/2$.

(iii) $\mathcal{X}$ is convex and $\langle\nabla\omega(\boldsymbol{x}) - \nabla\omega(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle \geq \sigma\|\boldsymbol{x} - \boldsymbol{y}\|^2$.

Condition (i) is the definition of strong convexity; note that it reduces to convexity when $\sigma = 0$. Conditions (ii) and (iii) are equivalent to the definition under appropriate differentiability conditions on $\omega$ that hold in our setting and when $\boldsymbol{x}, \boldsymbol{y}$ are in the interior of $\mathcal{X}$; see Yu [44] for details. Juditsky et al. [22] uses "strongly convex" to mean that a function is 1-strongly convex according to condition (iii).

Define the *divergence function:*

$$D_\omega(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \omega(\boldsymbol{x}) - \omega(\boldsymbol{y}) - \langle\nabla\omega(\boldsymbol{y}), \boldsymbol{x} - \boldsymbol{y}\rangle.$$

When $\omega$ is convex, one can see that $D_\omega$ is always non-negative, and by condition (ii) the $\sigma$-strong convexity of $\omega$ is equivalent to $D_\omega(\boldsymbol{x}, \boldsymbol{y}) \geq \sigma\|\boldsymbol{x} - \boldsymbol{y}\|^2/2$. We will use this equivalence to establish the strong convexity of our distance-generating functions below. An important operation related to distance-generating functions is the *proximal projection* onto $\mathcal{X}$ with respect to $\omega$:

$$\Pi_\omega(\boldsymbol{x}, \boldsymbol{\xi}) \coloneqq \operatorname*{arg\,min}_{\boldsymbol{y}\in\mathcal{X}} D_\omega(\boldsymbol{y}, \boldsymbol{x}) + \langle\boldsymbol{\xi}, \boldsymbol{y}\rangle, \quad \boldsymbol{\xi} \in \mathbb{R}^n.$$

The *center* of $\mathcal{X}$ with respect to $\omega$ is defined as $\boldsymbol{x}_0 \coloneqq \arg\min_{\boldsymbol{x}\in\mathcal{X}} \omega(\boldsymbol{x})$ and the *diameter* of $\mathcal{X}$ is $\Omega_\mathcal{X} \coloneqq \sup_{\boldsymbol{y}\in\mathcal{X}} \sqrt{2D_\omega(\boldsymbol{y}, \boldsymbol{x}_0)}$.

A strongly convex distance-generating function can be thought of as a generalization of the squared norm $\|\cdot\|^2$ — the corresponding divergence generalizes the squared distance function $\|\boldsymbol{x} - \boldsymbol{y}\|^2$; unlike the squared distance, however, the divergence may not be symmetric. Indeed, when $\|\cdot\|$ is an Euclidean norm, and *only* for such norms, the function $\omega(\boldsymbol{x}) = \|\boldsymbol{x}\|^2/2$ is 1-strongly convex [44, Proposition 2]. In this special case, the divergence is $D_\omega(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2/2$ and the proximal projection is simply the Euclidean projection: $\Pi_\omega(\boldsymbol{x}, \boldsymbol{\xi}) = \arg\min_{\boldsymbol{y}\in\mathcal{X}} \|\boldsymbol{x} + \boldsymbol{\xi} - \boldsymbol{y}\|$.

Thus, since the domain $\mathcal{B}$ of the primal variables $\boldsymbol{\theta}$ is equipped with the Euclidean norm $\|\boldsymbol{\theta}\| \equiv \|\boldsymbol{\Phi}\boldsymbol{\theta}\|_2$, we will use the 1-strongly convex distance-generating function $\omega_\mathcal{B}(\boldsymbol{\theta}) = \|\boldsymbol{\theta}\|^2/2$. Since $\mathcal{B}$ is the Euclidean ball under this norm, the center of $\mathcal{B}$ is $\boldsymbol{\theta}_0 = \boldsymbol{0}$ and its "diameter" (actually the radius, in this case) is $\Omega_\mathcal{B} = B$; $\boldsymbol{\theta}_0$ is used as the initial value of $\boldsymbol{\theta}$ in Algorithm 1. The proximal projection is

$$\Pi_\mathcal{B}(\boldsymbol{\theta}, \boldsymbol{\xi}) = \operatorname*{arg\,min}_{\boldsymbol{\theta}'\in\mathcal{B}} \|\boldsymbol{\theta} + \boldsymbol{\xi} - \boldsymbol{\theta}'\| = \frac{\boldsymbol{\theta} + \boldsymbol{\xi}}{\max\{1, \|\boldsymbol{\theta} + \boldsymbol{\xi}\|/B\}}.$$

For the dual variables $\boldsymbol{\lambda} \in \Lambda_\gamma$, our distance-generating function is a modification of the *unnormalized negentropy* $h(\boldsymbol{\lambda}) = \sum_i \lambda_i(\log\lambda_i - 1)$. It is well-known that this function is 1-strongly convex on the set $\{\boldsymbol{\lambda} \geq \boldsymbol{0} \mid \|\boldsymbol{\lambda}\|_1 \leq 1\}$ [e.g., 44, Theorem 5]. To achieve 1-strong convexity on $\Lambda_\gamma$ (where $\|\boldsymbol{\lambda}\|_1 = 1/(1-\gamma) > 1$), we use a modified form of this function:

$$h_\gamma(\boldsymbol{\lambda}) \coloneqq \frac{h((1-\gamma)\boldsymbol{\lambda})}{(1-\gamma)^2}.$$

It follows that $\nabla h_\gamma(\boldsymbol{\lambda}) = [\nabla h((1-\gamma)\boldsymbol{\lambda})]/(1-\gamma)$. Thus, defining $D_\Lambda(\boldsymbol{\lambda}', \boldsymbol{\lambda}) \coloneqq D_{h_\gamma}(\boldsymbol{\lambda}', \boldsymbol{\lambda})$, we have

$$\begin{aligned}
D_\Lambda(\boldsymbol{\lambda}', \boldsymbol{\lambda}) &= \left(\frac{1}{1-\gamma}\right)\left[\frac{h((1-\gamma)\boldsymbol{\lambda}')}{1-\gamma} - \frac{h((1-\gamma)\boldsymbol{\lambda})}{1-\gamma} - \langle \nabla h((1-\gamma)\boldsymbol{\lambda}), \boldsymbol{\lambda}' - \boldsymbol{\lambda}\rangle\right] \\
&= \left(\frac{1}{1-\gamma}\right)\left[\frac{h((1-\gamma)\boldsymbol{\lambda}')}{1-\gamma} - \frac{h((1-\gamma)\boldsymbol{\lambda})}{1-\gamma} - \frac{\langle \nabla h((1-\gamma)\boldsymbol{\lambda}), (1-\gamma)\boldsymbol{\lambda}' - (1-\gamma)\boldsymbol{\lambda}\rangle}{1-\gamma}\right]
\end{aligned}$$

Now, since $(1-\gamma)\boldsymbol{\lambda}, (1-\gamma)\boldsymbol{\lambda}' \in \Delta_{\mathcal{S}_+\times\mathcal{A}}$ and $h$ is strongly convex on this set, we have

$$D_\Lambda(\boldsymbol{\lambda}', \boldsymbol{\lambda}) \geq \left(\frac{1}{1-\gamma}\right)^2 \frac{\|(1-\gamma)\boldsymbol{\lambda} - (1-\gamma)\boldsymbol{\lambda}'\|_1^2}{2} = \frac{\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_1^2}{2}.$$

Thus we have shown that $h_\gamma$ is 1-strongly convex on $\Lambda_\gamma$. By the properties of the negentropy function, we can verify that $h_\gamma(\boldsymbol{\lambda})$ is minimized for $\boldsymbol{\lambda}_0 = \mathbf{1}_{\mathcal{A}}/A \oplus \gamma\mathbf{1}_{mA}/(1-\gamma)mA$, i.e. the uniform distribution over actions concatenated with the scaled uniform distribution over state-action pairs in $\mathcal{S}_* \times \mathcal{A}$ — this value is used as the initializer in Algorithm 1. Conversely, $h_\gamma(\boldsymbol{\lambda})$ is maximized for $\bar{\boldsymbol{\lambda}} = \boldsymbol{e}_a \oplus \gamma\boldsymbol{e}_{s'a'}/(1-\gamma)$, i.e. when $\bar{\boldsymbol{\lambda}}$ is concentrated on $(s_0, a)$ for some $a \in \mathcal{A}$ and some $(s', a') \in \mathcal{S}_* \times \mathcal{A}$. We can verify through a short calculation that

$$\begin{aligned}
D_\Lambda(\bar{\boldsymbol{\lambda}}, \boldsymbol{\lambda}_0) &= \frac{\ell}{(1-\gamma)^2}, &\qquad \text{where } \ell \coloneqq \log A + \gamma \log m \\
\Omega_\Lambda &= \sqrt{2D_\Lambda(\bar{\boldsymbol{\lambda}}, \boldsymbol{\lambda}_0)} = \frac{\sqrt{2\ell}}{1-\gamma}.
\end{aligned}$$

Finally, the proximal projection onto $\Lambda_\gamma$ with respect to $h_\gamma$ is

$$\Pi_\Lambda(\boldsymbol{\lambda}, \boldsymbol{\rho}) = \frac{\tilde{\boldsymbol{\lambda}}_0}{\|\tilde{\boldsymbol{\lambda}}_0\|_1} \oplus \frac{\gamma\tilde{\boldsymbol{\lambda}}_*}{(1-\gamma)\|\tilde{\boldsymbol{\lambda}}_*\|_1}, \qquad \text{where } \tilde{\boldsymbol{\lambda}} \coloneqq \exp(\log\boldsymbol{\lambda} + \boldsymbol{\rho}),$$

where $\tilde{\boldsymbol{\lambda}}_{s_0} \coloneqq [\tilde{\lambda}_{s_0a}]_{a\in\mathcal{A}}$ and $\tilde{\boldsymbol{\lambda}}_* \coloneqq [\tilde{\lambda}_{sa}]_{s\in\mathcal{S}_*, a\in\mathcal{A}}$, so that $\tilde{\boldsymbol{\lambda}} = \tilde{\boldsymbol{\lambda}}_{s_0} \oplus \tilde{\boldsymbol{\lambda}}_*$.

### A.4.4 The Composite Space

We will now gather together the preceding results and use them to construct a norm and distance-generating function on the composite optimization domain $Z = \Lambda_\gamma \times \mathcal{B} \subset \mathbb{R}^{(1+m)A} \oplus \mathbb{R}^d$. We closely follow the construction of Juditsky et al. [22, §4.2]. First, we define the squared norm:

$$\|\boldsymbol{\lambda} \oplus \boldsymbol{\theta}\|^2 \coloneqq \Omega_\Lambda^{-2}\|\boldsymbol{\lambda}\|_1^2 + \Omega_\mathcal{B}^{-2}\|\boldsymbol{\theta}\|^2 = \left(\frac{1-\gamma}{\sqrt{2\ell}}\|\boldsymbol{\lambda}\|_1\right)^2 + \left(\frac{1}{B}\|\boldsymbol{\theta}\|\right)^2.$$

The corresponding squared dual norm is

$$\|\boldsymbol{\rho} \oplus \boldsymbol{\xi}\|_*^2 \coloneqq \Omega_\Lambda^2\|\boldsymbol{\lambda}\|_\infty^2 + \Omega_\mathcal{B}^2\|\boldsymbol{\theta}\|_*^2 = \left(\frac{\sqrt{2\ell}}{1-\gamma}\|\boldsymbol{\lambda}\|_\infty\right)^2 + (B\|\boldsymbol{\theta}\|)^2.$$

Define the operator $F(\boldsymbol{\lambda}, \boldsymbol{\theta}) \coloneqq -f_\lambda(\boldsymbol{\theta}) \oplus f_\theta(\boldsymbol{\lambda})$. Its Lipschitz constant with respect to this norm is

$$\begin{aligned}
\|F(\boldsymbol{\lambda}, \boldsymbol{\theta})\|_* &= \sqrt{\frac{2\ell\|f_\lambda(\boldsymbol{\theta})\|_\infty^2}{(1-\gamma)^2} + B^2\|f_\theta(\boldsymbol{\lambda})\|_*^2} \\
&\leq \sqrt{\frac{8\ell B^2}{(1-\gamma)^2} + \left(\frac{2B}{1-\gamma}\right)^2} \\
&= \frac{2B\sqrt{1+2\ell}}{1-\gamma} = C.
\end{aligned}$$

Similarly, define the estimator $\hat{F}(\boldsymbol{\lambda}, \boldsymbol{\theta}) \coloneqq -\hat{f}_\lambda(\boldsymbol{\theta}) \oplus \hat{f}_\theta(\boldsymbol{\lambda})$. Its variance enjoys

$$\mathbb{E}[\|\hat{F}(\boldsymbol{\lambda}, \boldsymbol{\theta}) - F(\boldsymbol{\lambda}, \boldsymbol{\theta})\|_*^2] \leq C^2.$$

We do not repeat the calculation because it is identical to the previous one, since the variances of our estimators have the same bounds as their squared Lipschitz constants.

Finally, we construct the composite distance-generating function:

$$\omega_Z(\boldsymbol{\lambda}, \boldsymbol{\theta}) \coloneqq \frac{(1-\gamma)^2 h_\gamma(\boldsymbol{\lambda})}{2\ell} + \frac{\|\boldsymbol{\theta}\|^2}{2B^2}.$$

We can verify that this function is 1-strongly convex on $Z$ with respect to the norm defined above, and that the diameter of $Z$ under this function is $\Omega \leq \sqrt{2}$.

*Proof of Lemma 11.* We apply the result of Juditsky et al. [22, Corollary 1] to our setting, where the Lipschitz constant of $F$ is $C$, the variance of $\hat{F}$ is $C^2$, and the diameter of $Z$ is $\Omega \leq \sqrt{2}$. Then the result tells us that a suitable learning rate is $\eta = C^{-1}\sqrt{2/7T}$, as specified in Theorem 3, and the resulting bound on the expected duality gap after $T$ iterations is

$$\varepsilon_{\text{opt}} \coloneqq \mathbb{E}[\delta_{\mathcal{B}}(\hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{\theta}})] \leq \frac{14C}{\sqrt{3T}}. \qquad \square$$