Appendix

A Review of RL Setups

We provide an extended review of different formulations of RL for interested readers. First, let us recall the problem setup. Let $\mathcal{S}$ and $\mathcal{A}$ be state and action spaces, and let $\pi(a|s)$ denote a policy. For $\gamma \in [0,1)$, we are interested in solving a γ-discounted infinite-horizon RL problem:

$$\max_\pi V^\pi(p), \qquad \text{s.t.} \quad V^\pi(p) \coloneqq (1-\gamma)\mathbb{E}_{s_0\sim p}\mathbb{E}_{\xi\sim\rho^\pi(s_0)}\left[\sum_{t=0}^\infty \gamma^t r(s_t,a_t)\right] \tag{24}$$

where $V^\pi(p)$ is the discounted average return, $r : \mathcal{S}\times\mathcal{A} \to [0,1]$ is the reward function, $\rho^\pi(s_0)$ denotes the distribution of trajectory $\xi = s_0, a_0, s_1, \dots$ generated by running π from state s_0 in a Markov decision process (MDP), and p is a fixed but unknown initial state distribution.

A.1 Coordinate-wise Formulations

RL in terms of stationary state distribution Let $d_t^\pi(s)$ denote the state distribution at time t given by running π starting from p. We define its γ-weighted mixture as

$$d^\pi(s) \coloneqq (1-\gamma)\sum_{t=0}^\infty \gamma^t d_t^\pi(s) \tag{25}$$

We can view d^π in (25) as a form of stationary state distribution of π, because it is a valid probability distribution of state and satisfies the stationarity property below,

$$d^\pi(s') = (1-\gamma)p(s') + \gamma\mathbb{E}_{s\sim d^\pi}\mathbb{E}_{a\sim\pi|s}[\mathcal{P}(s'|s,a)] \tag{2}$$

where $\mathcal{P}(s'|s,a)$ is the transition probability of the MDP. The definition in (25) generalizes the concept of stationary distribution of MDP; as $\gamma \to 1$, d^π is known as the limiting average state distribution, which is the same as the stationary distribution of the MDP under π, if one exists. Moreover, with the property in (2), d^π summarizes the Markov structure of RL, and allows us to write (24) simply as

$$\max_\pi V^\pi(p), \qquad \text{s.t.} \quad V^\pi(p) = \mathbb{E}_{s\sim d^\pi}\mathbb{E}_{a\sim\pi|s}\left[r(s,a)\right] \tag{26}$$

after commuting the order of expectation and summation. That is, an RL problem aims to maximize the expected reward under the stationary state-action distribution generated by the policy π.

RL in terms of value function We can also write (24) in terms of value function. Recall

$$V^\pi(s) \coloneqq (1-\gamma)\mathbb{E}_{\xi\sim\rho^\pi(s_0)|s_0=s}\left[\sum_{t=0}^\infty \gamma^t r(s_t,a_t)\right] \tag{1}$$

is the value function of π. By definition, V^π (like d^π) satisfies a stationarity property

$$V^\pi(s) = \mathbb{E}_{a\sim\pi|s}\left[(1-\gamma)r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}|s,a}\left[V^\pi(s')\right]\right] \tag{5}$$

which can be viewed as a dual equivalent of (2). Because r is in $[0,1]$, (5) implies V^π lies in $[0,1]$.

The value function V^* (a shorthand of V_{π^*}) of the optimal policy π^* of the RL problem satisfies the so-called Bellman equation (Bellman, 1954): $V^*(s) = \max_{a\in\mathcal{A}}(1-\gamma)r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}|s,a}\left[V^*(s')\right]$, where the optimal policy π^* can be recovered as the arg max. Equivalently, by the definition of max, the Bellman equation amounts to finding the smallest V such that $V(s) \geq (1-\gamma)r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}|s,a}\left[V(s')\right]$, $\forall s\in\mathcal{S}, a\in\mathcal{A}$. In other words, the RL problem in (24) can be written as

$$\min_V \mathbb{E}_{s\sim p}[V(s)] \qquad \text{s.t.} \quad V(s) \geq (1-\gamma)r(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}|s,a}\left[V(s')\right], \qquad \forall s\in\mathcal{S}, a\in\mathcal{A} \tag{27}$$

A.2 Linear Programming Formulations

We now connect the above two alternate expressions through the classical LP setup of RL (Manne et al., 1959; Denardo and Fox, 1968).

LP in terms of value function The classic LP formulation[6] is simply a restatement of (27):

$$\min_{\mathbf{v}} \quad \mathbf{p}^\top \mathbf{v} \qquad \text{s.t.} \quad (1-\gamma)\mathbf{r} + \gamma \mathbf{P}\mathbf{v} \le \mathbf{E}\mathbf{v} \tag{4}$$

where $\mathbf{p} \in \mathbb{R}^{|\mathcal{S}|}$, $\mathbf{v} \in \mathbb{R}^{|\mathcal{S}|}$, and $\mathbf{r} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ are the vector forms of p, V, r, respectively, $\mathbf{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$ is the transition probability[7], and $\mathbf{E} = \mathbf{I} \otimes \mathbf{1} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$ (we use $|\cdot|$ to denote the cardinality of a set, $\otimes$ the Kronecker product, $\mathbf{I} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$ is the identity, and $\mathbf{1} \in \mathbb{R}^{|\mathcal{A}|}$ a vector of ones). It is easy to verify that for all $\mathbf{p} > 0$, the solution to (4) is the same and equal to $\mathbf{v}^*$ (the vector form of V^*).

LP in terms of stationary state-action distribution Define the Lagrangian function

$$\mathcal{L}(\mathbf{v}, \mathbf{f}) \coloneqq \mathbf{p}^\top \mathbf{v} + \mathbf{f}^\top ((1-\gamma)\mathbf{r} + \gamma \mathbf{P}\mathbf{v} - \mathbf{E}\mathbf{v}) \tag{28}$$

where $\mathbf{f} \ge \mathbf{0} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is the Lagrangian multiplier. By Lagrangian duality, the dual problem of (4) is given as $\max_{\mathbf{f} \ge \mathbf{0}} \min_{\mathbf{v}} \mathcal{L}(\mathbf{v}, \mathbf{f})$. Or after substituting the optimality condition of $\mathbf{v}$ and define $\boldsymbol{\mu} \coloneqq (1-\gamma)\mathbf{f}$, we can write the dual problem as another LP problem

$$\max_{\boldsymbol{\mu} \ge \mathbf{0}} \quad \mathbf{r}^\top \boldsymbol{\mu} \qquad \text{s.t.} \quad (1-\gamma)\mathbf{p} + \gamma \mathbf{P}^\top \boldsymbol{\mu} = \mathbf{E}^\top \boldsymbol{\mu} \tag{3}$$

Note that this problem like (4) is normalized: we have $\|\boldsymbol{\mu}\|_1 = 1$ because $\|\mathbf{p}\|_1 = 1$, and

$$\|\boldsymbol{\mu}\|_1 = \mathbf{1}^\top \mathbf{E}^\top \boldsymbol{\mu} = (1-\gamma)\mathbf{1}^\top \mathbf{p} + \gamma \mathbf{1}^\top \mathbf{P}^\top \boldsymbol{\mu} = (1-\gamma)\|\mathbf{p}\|_1 + \gamma \|\boldsymbol{\mu}\|_1$$

where we use the facts that $\boldsymbol{\mu} \ge \mathbf{0}$ and $\mathbf{P}$ is a stochastic transition matrix. This means that $\boldsymbol{\mu}$ is a valid state-action distribution, from which we see that the equality constraint in (3) is simply a vector form (2). Therefore, (3) is the same as (26) if we define the policy π as the conditional distribution based on $\boldsymbol{\mu}$.

B Missing Proofs of Section 3

B.1 Proof of Lemma 1

Lemma 1. *For any $x = (\mathbf{v}, \boldsymbol{\mu})$, if $x' \in \mathcal{X}$ satisfies* (2) *and* (5) *(i.e. $\mathbf{v}'$ and $\boldsymbol{\mu}'$ are the value function and state-action distribution of policy $\pi_{\boldsymbol{\mu}'}$), $r_{ep}(x; x') = -\boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}'}$.*

Proof. First note that $F(x, x) = 0$. Then as x' satisfies stationarity, we can use Lemma 2 below and write

$$\begin{aligned}
r_{ep}(x; x') &= F(x, x) - F(x, x') \\
&= -F(x, x') \\
&= -(\mathbf{p}^\top \mathbf{v}' - \mathbf{p}^\top \mathbf{v}) - \boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}'} + \boldsymbol{\mu}'^\top \mathbf{a}_{\mathbf{v}} && (\because \text{Definition of } F \text{ in (14)}) \\
&= -\boldsymbol{\mu}' \mathbf{a}_{\mathbf{v}} - \boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}'} + \boldsymbol{\mu}'^\top \mathbf{a}_{\mathbf{v}} && (\because \text{Lemma 2}) \\
&= -\boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}'}
\end{aligned}$$

□

B.2 Proof of Lemma 2

Lemma 2. *Let $\mathbf{v}^\pi$ and $\boldsymbol{\mu}^\pi$ denote the value and state-action distribution of some policy π. Then for* any *function $\mathbf{v}'$, it holds that $\mathbf{p}^\top (\mathbf{v}^\pi - \mathbf{v}') = (\boldsymbol{\mu}^\pi)^\top \mathbf{a}_{\mathbf{v}'}$. In particular, it implies $V^\pi(p) - V^{\pi'}(p) = (\boldsymbol{\mu}^\pi)^\top \mathbf{a}_{\mathbf{v}^{\pi'}}$.*

Proof. This is the well-known performance difference lemma. The proof is based on the stationary properties in (2) and (5), which can be stated in vector form as

$$(\boldsymbol{\mu}^\pi)^\top \mathbf{E}\mathbf{v}^\pi = (\boldsymbol{\mu}^\pi)^\top ((1-\gamma)\mathbf{r} + \gamma \mathbf{P}\mathbf{v}^\pi) \qquad \text{and} \qquad (1-\gamma)\mathbf{p} + \gamma \mathbf{P}^\top \boldsymbol{\mu}^\pi = \mathbf{E}^\top \boldsymbol{\mu}^\pi$$

[6] Our setup in (4) differs from the classic one in the $(1-\gamma)$ factor in the constraint to normalize the problem.

[7] We arrange the coordinates in a way such that along the $|\mathcal{S}||\mathcal{A}|$ indices are contiguous in actions.

The proof is a simple application of these two properties.

$$\begin{aligned}\mathbf{p}^\top(\mathbf{v}^\pi - \mathbf{v}') &= \frac{1}{1-\gamma}(\mathbf{E}^\top\boldsymbol{\mu}^\pi - \gamma\mathbf{P}^\top\boldsymbol{\mu}^\pi)^\top(\mathbf{v}^\pi - \mathbf{v}') \\ &= \frac{1}{1-\gamma}(\boldsymbol{\mu}^\pi)^\top((\mathbf{E}-\gamma\mathbf{P})\mathbf{v}^\pi - (\mathbf{E}-\gamma\mathbf{P})\mathbf{v}') \\ &= \frac{1}{1-\gamma}(\boldsymbol{\mu}^\pi)^\top((1-\gamma)\mathbf{r} - (\mathbf{E}-\gamma\mathbf{P})\mathbf{v}') = (\boldsymbol{\mu}^\pi)^\top\mathbf{a}_{\mathbf{v}'}\end{aligned}$$

where we use the stationarity property of $\boldsymbol{\mu}^\pi$ in the first equality and that $\mathbf{v}^\pi$ in the third equality. □

B.3 Proof of Proposition 2

Proposition 2. *For any $x = (\mathbf{v}, \boldsymbol{\mu}) \in \mathcal{X}$, if $\mathbf{E}^\top\boldsymbol{\mu} \geq (1-\gamma)\mathbf{p}$, $r_{ep}(x; x^*) \geq (1-\gamma)\min_s p(s)\|\mathbf{v}^* - \mathbf{v}^{\pi_\mu}\|_\infty$.*

Proof. This proof mainly follows the steps in Wang (2017a) but written in our notation. First Lemma 1 shows $r_{ep}(x; x^*) = -\boldsymbol{\mu}^\top\mathbf{a}_{\mathbf{v}^*}$. We then lower bound $-\boldsymbol{\mu}^\top\mathbf{a}_{\mathbf{v}^*}$ by reversing the proof of the performance difference lemma (Lemma 2).

$$\begin{aligned}\boldsymbol{\mu}^\top\mathbf{a}_{\mathbf{v}^*} &= \frac{1}{1-\gamma}\boldsymbol{\mu}^\top((1-\gamma)\mathbf{r} - (\mathbf{E}-\gamma\mathbf{P})\mathbf{v}^*) && (\because \text{Definition of } \mathbf{a}_{\mathbf{v}^*}) \\ &= \frac{1}{1-\gamma}\boldsymbol{\mu}^\top((\mathbf{E}-\gamma\mathbf{P})\mathbf{v}^{\pi_\mu} - (\mathbf{E}-\gamma\mathbf{P})\mathbf{v}^*) && (\because \text{Stationarity of } \mathbf{v}^{\pi_\mu}) \\ &= \frac{1}{1-\gamma}\boldsymbol{\mu}^\top(\mathbf{E}-\gamma\mathbf{P})(\mathbf{v}^{\pi_\mu} - \mathbf{v}^*) \\ &= \frac{1}{1-\gamma}\mathbf{d}^\top(\mathbf{I}-\gamma\mathbf{P}_{\pi_\mu})(\mathbf{v}^{\pi_\mu} - \mathbf{v}^*)\end{aligned}$$

where we define $\mathbf{d} \coloneqq \mathbf{E}^\top\boldsymbol{\mu}$ and $\mathbf{P}_{\pi_\mu}$ as the state-transition of running policy $\pi_{\boldsymbol{\mu}}$.

We wish to further upper bound this quantity. To proceed, we appeal to the Bellman equation of the optimal value function $\mathbf{v}^*$ and the stationarity of $\mathbf{v}^{\pi_\mu}$:

$$\mathbf{v}^* \geq (1-\gamma)\mathbf{r}_{\pi_\mu} + \gamma\mathbf{P}_{\pi_\mu}\mathbf{v}^* \qquad \text{and} \qquad \mathbf{v}^{\pi_\mu} = (1-\gamma)\mathbf{r}_{\pi_\mu} + \gamma\mathbf{P}_{\pi_\mu}\mathbf{v}^{\pi_\mu},$$

which together imply that $(\mathbf{I}-\gamma\mathbf{P}_{\pi_\mu})(\mathbf{v}^{\pi_\mu} - \mathbf{v}^*) \leq 0$. We will also use the stationarity of $\mathbf{d}^{\pi_\mu}$ (the average state distribution of $\pi_{\boldsymbol{\mu}}$): $\mathbf{d}^{\pi_\mu} = (1-\gamma)\mathbf{p} + \gamma\mathbf{P}_{\boldsymbol{\pi}_{\boldsymbol{\mu}}}^\top\mathbf{d}^{\pi_\mu}$.

Since $\mathbf{d} \geq (1-\gamma)\mathbf{p}$ in the assumption, we can then write

$$\begin{aligned}\boldsymbol{\mu}^\top\mathbf{a}_{\mathbf{v}^*} &= \frac{1}{1-\gamma}\mathbf{d}^\top(\mathbf{I}-\gamma\mathbf{P}_{\pi_\mu})(\mathbf{v}^{\pi_\mu} - \mathbf{v}^*) \\ &\leq \mathbf{p}^\top(\mathbf{I}-\gamma\mathbf{P}_{\pi_\mu})(\mathbf{v}^{\pi_\mu} - \mathbf{v}^*) \\ &\leq -\min_s p(s)\|(\mathbf{I}-\gamma\mathbf{P}_{\pi_\mu})(\mathbf{v}^{\pi_\mu} - \mathbf{v}^*)\|_\infty \\ &\leq -\min_s p(s)(1-\gamma)\|\mathbf{v}^{\pi_\mu} - \mathbf{v}^*\|_\infty.\end{aligned}$$

Finally, flipping the sign of the inequality concludes the proof. □

B.4 Proof of Proposition 3

Proposition 3. *There is a class of MDPs such that, for some $x \in \mathcal{X}$, Proposition 2 is an equality.*

Proof. We show this equality holds for a class of MDPs. For simplicity, let us first consider an MDP with three states 1, 2, 3 and for each state there are three actions ($left$, $right$, $stay$). They correspond to intuitive, deterministic transition dynamics

$$\mathcal{P}(\max\{s-1, 1\}|s, left) = 1, \quad \mathcal{P}(\min\{s+1, 3\}|s, right) = 1, \quad \mathcal{P}(s|s, stay) = 1.$$

We set the reward as $r(s, right) = 1$ for $s = 1, 2, 3$ and zero otherwise. It is easy to see that the optimal policy is $\pi^*(right|s) = 1$, which has value function $\mathbf{v}^* = [1, 1, 1]^\top$.

Now consider $x = (\mathbf{v}, \boldsymbol{\mu}) \in \mathcal{X}$. To define $\boldsymbol{\mu}$, let $\mu(s, a) = d(s)\pi_{\boldsymbol{\mu}}(a|s)$. We set

$$\pi_{\boldsymbol{\mu}}(right|1) = 1, \quad \pi_{\boldsymbol{\mu}}(stay|2) = 1, \qquad \pi_{\boldsymbol{\mu}}(right|3) = 1$$

That is, $\pi_{\boldsymbol{\mu}}$ is equal to π^* except when $s = 2$. One can verify the value function of this policy is $\mathbf{v}^{\pi_\mu} = [(1-\gamma), 0, 1]^\top$.

As far as d is concerned ($\mathbf{d} = \mathbf{E}^\top \boldsymbol{\mu}$), suppose the initial distribution is uniform, i.e. $\mathbf{p} = [1/3, 1/3, 1/3]^\top$, we choose d as $\mathbf{d} = (1-\gamma)\mathbf{p} + \gamma[1, 0, 0]^\top$, which satisfies the assumption in Proposition 2. Therefore, we have $\boldsymbol{\mu} \in \mathcal{M}'$ and we will let $\mathbf{v}$ be some arbitrary point in $\mathcal{V}$.

Now we show for this choice $x = (\mathbf{v}, \boldsymbol{\mu}) \in \mathcal{V} \times \mathcal{M}'$, the equality in Proposition 2 holds. By Lemma 1, we know $r_{ep}(x; x') = -\boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}^*}$. Recall the advantage is defined as $\mathbf{a}_{\mathbf{v}^*} = \mathbf{r} + \frac{1}{1-\gamma}(\gamma\mathbf{P} - \mathbf{E})\mathbf{v}^*$. Let $A_{V^*}(s, a)$ denote the functional form of $\mathbf{a}_{\mathbf{v}^*}$ and define the expected advantage:

$$A_{V^*}(s, \pi_{\boldsymbol{\mu}}) \coloneqq \mathbb{E}_{a\sim\pi_{\boldsymbol{\mu}}}[A_{V^*}(s, a)].$$

We can verify it has the following values:

$$A_{V^*}(1, \pi_{\boldsymbol{\mu}}) = 0, \quad A_{V^*}(2, \pi_{\boldsymbol{\mu}}) = -1, \quad A_{V^*}(3, \pi_{\boldsymbol{\mu}}) = 0.$$

Thus, the above construction yields

$$r_{ep}(x; x^*) = -\boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}^*} = \frac{(1-\gamma)}{3} = (1-\gamma)\min_s p(s)\|\mathbf{v}^* - \mathbf{v}^{\pi_\mu}\|_\infty$$

One can easily generalize this 3-state MDP to an $|\mathcal{S}|$-state MDP where states are partitioned into three groups. □

C Missing Proofs of Section 4

C.1 Proof of Proposition 4

Proposition 4. *For $x = (\mathbf{v}, \boldsymbol{\mu}) \in \mathcal{X}$, define $y_x^* \coloneqq (\mathbf{v}^{\pi_\mu}, \boldsymbol{\mu}^*) \in \mathcal{X}$. It holds $r_{ep}(x; y_x^*) = V^*(p) - V^{\pi_\mu}(p)$.*

Proof. First we generalize Lemma 1.

Lemma 3. *Let $x = (\mathbf{v}, \boldsymbol{\mu})$ be arbitrary. Consider $\tilde{x}' = (\mathbf{v}' + \mathbf{u}', \boldsymbol{\mu}')$, where $\mathbf{v}'$ and $\boldsymbol{\mu}'$ are the value function and state-action distribution of policy $\pi_{\boldsymbol{\mu}'}$, and $\mathbf{u}'$ is arbitrary. It holds that $r_{ep}(x; \tilde{x}') = -\boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}'} - \mathbf{b}_{\boldsymbol{\mu}}^\top \mathbf{u}'$.*

To proceed, we write $y_x^* = (\mathbf{v}^* + (\mathbf{v}^{\pi_\mu} - \mathbf{v}^*), \boldsymbol{\mu}^*)$ and use Lemma 3, which gives $r_{ep}(x; y_x^*) = -\boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}^*} - \mathbf{b}_{\boldsymbol{\mu}}^\top(\mathbf{v}^{\pi_\mu} - \mathbf{v}^*)$. To relate this equality to the policy performance gap, we also need the following equality.

Lemma 4. *For $\boldsymbol{\mu} \in \mathcal{M}$, it holds that $-\boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}^*} = V^*(p) - V^{\pi_\mu}(p) + \mathbf{b}_{\boldsymbol{\mu}}^\top(\mathbf{v}^{\pi_\mu} - \mathbf{v}^*)$.*

Together they imply the desired equality $r_{ep}(x; y_x^*) = V^*(p) - V^{\pi_\mu}(p)$. □

C.1.1 Proof of Lemma 3

Lemma 3. *Let $x = (\mathbf{v}, \boldsymbol{\mu})$ be arbitrary. Consider $\tilde{x}' = (\mathbf{v}' + \mathbf{u}', \boldsymbol{\mu}')$, where $\mathbf{v}'$ and $\boldsymbol{\mu}'$ are the value function and state-action distribution of policy $\pi_{\boldsymbol{\mu}'}$, and $\mathbf{u}'$ is arbitrary. It holds that $r_{ep}(x; \tilde{x}') = -\boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}'} - \mathbf{b}_{\boldsymbol{\mu}}^\top \mathbf{u}'$.*

Proof. Let $x' = (\mathbf{v}', \boldsymbol{\mu}')$. As shorthand, define $\mathbf{f}' \coloneqq \mathbf{v}' + \mathbf{u}'$, and $\mathbf{L} \coloneqq \frac{1}{1-\gamma}(\gamma\mathbf{P} - \mathbf{E})$ (i.e. we can write $\mathbf{a}_{\mathbf{f}} = \mathbf{r} + \mathbf{L}\mathbf{f}$). Because $r_{ep}(x; x') = -F(x, x') = -(\mathbf{p}^\top \mathbf{v}' + \boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}'} - \mathbf{p}^\top \mathbf{v} - \boldsymbol{\mu}'^\top \mathbf{a}_{\mathbf{v}})$, we can write

$$\begin{aligned}
r_{ep}(x; \tilde{x}') &= -\mathbf{p}^\top \mathbf{f}' - \boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{f}'} + \mathbf{p}^\top \mathbf{v} + \boldsymbol{\mu}'^\top \mathbf{a}_{\mathbf{v}} \\
&= \left(-\mathbf{p}^\top \mathbf{v}' - \boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}'} + \mathbf{p}^\top \mathbf{v} + \boldsymbol{\mu}'^\top \mathbf{a}_{\mathbf{v}}\right) - \mathbf{p}^\top \mathbf{u}' - \boldsymbol{\mu}^\top \mathbf{L}\mathbf{u}' \\
&= r_{ep}(x; x') - \mathbf{p}^\top \mathbf{u}' - \boldsymbol{\mu}^\top \mathbf{L}\mathbf{u}' \\
&= r_{ep}(x; x') - \mathbf{b}_{\boldsymbol{\mu}}^\top \mathbf{u}'
\end{aligned}$$

Finally, by Lemma 1, we have also $r_{ep}(x; x') = -\boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}'}$ and therefore the final equality. □

C.1.2 Proof of Lemma 4

Lemma 4. *For $\boldsymbol{\mu} \in \mathcal{M}$, it holds that $-\boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}^*} = V^*(p) - V^{\pi_\mu}(p) + \mathbf{b}_{\boldsymbol{\mu}}^\top(\mathbf{v}^{\pi_\mu} - \mathbf{v}^*)$.*

Proof. Following the setup in Lemma 3, we prove the statement by the rearrangement below:

$$\begin{aligned}
-\boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}'} &= -(\boldsymbol{\mu}^{\pi_\mu})^\top \mathbf{a}_{\mathbf{v}'} + (\boldsymbol{\mu}^{\pi_\mu})^\top \mathbf{a}_{\mathbf{v}'} - \boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}'} \\
&= V^{\pi'}(p) - V^{\pi_\mu}(p) + (\boldsymbol{\mu}^{\pi_\mu} - \boldsymbol{\mu})^\top \mathbf{a}_{\mathbf{v}'} \\
&= \Big(V^{\pi'}(p) - V^{\pi_\mu}(p)\Big) + (\boldsymbol{\mu}^{\pi_\mu} - \boldsymbol{\mu})^\top \mathbf{r} + (\boldsymbol{\mu}^{\pi_\mu} - \boldsymbol{\mu})^\top \mathbf{L}\mathbf{v}'
\end{aligned}$$

where the second equality is due to the performance difference lemma, i.e. Lemma 2, and the last equality uses the definition $\mathbf{a}_{\mathbf{v}'} = \mathbf{r} + \mathbf{L}\mathbf{v}'$. For the second term above, let $\mathbf{r}_{\pi_\mu}$ and $\mathbf{P}_{\pi_\mu}$ denote the expected reward and transition under $\pi_{\boldsymbol{\mu}}$. Because $\boldsymbol{\mu} \in \mathcal{M}$, we can rewrite it as

$$\begin{aligned}
(\boldsymbol{\mu}^{\pi_\mu} - \boldsymbol{\mu})^\top \mathbf{r} &= (\mathbf{E}^\top \boldsymbol{\mu}^{\pi_\mu} - \mathbf{E}^\top \boldsymbol{\mu}) \mathbf{r}_{\pi_\mu} \\
&= ((1-\gamma)\mathbf{p} + \gamma \mathbf{P}^\top \boldsymbol{\mu}^{\pi_\mu} - \mathbf{E}^\top \boldsymbol{\mu}) \mathbf{r}_{\pi_\mu} \\
&= (1-\gamma)\mathbf{b}_{\boldsymbol{\mu}}^\top \mathbf{r}_{\pi_\mu} + \gamma (\boldsymbol{\mu}^{\pi_\mu} - \boldsymbol{\mu})^\top \mathbf{P}\mathbf{r}_{\pi_\mu} \\
&= (1-\gamma)\mathbf{b}_{\boldsymbol{\mu}}^\top \Big(\mathbf{r}_{\pi_\mu} + \gamma \mathbf{P}_{\pi_\mu}\mathbf{r}_{\pi_\mu} + \gamma^2 \mathbf{P}_{\pi_\mu}^2 \mathbf{r}_{\pi_\mu} + \dots\Big) \\
&= \mathbf{b}_{\boldsymbol{\mu}}^\top \mathbf{v}^{\pi_\mu}
\end{aligned}$$

where the second equality uses the stationarity of $\boldsymbol{\mu}^{\boldsymbol{\pi}_{\boldsymbol{\mu}}}$ given by (2). For the third term, it can be written

$$(\boldsymbol{\mu}^{\pi_\mu} - \boldsymbol{\mu})^\top \mathbf{L}\mathbf{v}' = (-\mathbf{p} - \mathbf{L}^\top \boldsymbol{\mu})^\top \mathbf{v}' = -\mathbf{b}_{\boldsymbol{\mu}}^\top \mathbf{v}'$$

where the first equality uses stationarity, i.e. $\mathbf{b}_{\boldsymbol{\mu}^{\pi_\mu}} = \mathbf{p} + \mathbf{L}^\top \boldsymbol{\mu}^{\pi_\mu} = 0$. Finally combining the three steps, we have

$$-\boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}'} = V^{\pi'}(p) - V^{\pi_\mu}(p) + \mathbf{b}_{\boldsymbol{\mu}}(\mathbf{v}^{\pi_\mu} - \mathbf{v}')$$

□

C.2 Proof of Corollary 1

Corollary 1. *Let $X_N = \{x_n \in \mathcal{X}_\theta\}_{n=1}^N$ be* any *sequence. Let $\hat{\pi}_N$ be the policy given either by the average or the best decision in X_N. It holds that*

$$V^{\hat{\pi}_N}(p) \geq V^*(p) - \frac{\text{Regret}_N(\Theta)}{N} - \epsilon_{\Theta,N}$$

where $\epsilon_{\Theta,N} = \min_{x_\theta \in \mathcal{X}_\theta} r_{ep}(\hat{x}_N; y_N^*) - r_{ep}(\hat{x}_N; x_\theta)$ measures the expressiveness of X_θ, and $\text{Regret}_N(\Theta) \coloneqq \sum_{n=1}^N l_n(x_n) - \min_{x \in \mathcal{X}_\Theta} \sum_{n=1}^N l_n(x)$.

Proof. This can be proved by a simple rearrangement

$$V^*(p) - V^{\hat{\pi}_N}(p) = r_{ep}(\hat{x}_N; y_N^*) = \epsilon_{\Theta,N} + \max_{x_\theta \in \mathcal{X}_\theta} r_{ep}(\hat{x}_N; x_\theta) \leq \epsilon_{\Theta,N} + \frac{\text{Regret}_N(\Theta)}{N}$$

where the first equality is Proposition 4 and the last inequality is due to the skew-symmetry of F, similar to the proof of Theorem 1. □

C.3 Proof of Proposition 5

Proposition 5. *Let $\hat{x}_N = (\hat{\mathbf{v}}_N, \hat{\boldsymbol{\mu}}_N)$. Under the setup in Corollary 1,* regardless *of the parameterization, it is true that $\epsilon_{\Theta,N}$ is no larger than*

$$\begin{aligned}
&\min_{(\mathbf{v}_\theta, \boldsymbol{\mu}_\theta) \in \mathcal{X}_\Theta} \frac{\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}^*\|_1}{1-\gamma} + \min_{\mathbf{w}: \mathbf{w} \geq 1} \|\mathbf{b}_{\hat{\boldsymbol{\mu}}_N}\|_{1,\mathbf{w}} \|\mathbf{v}_\theta - \mathbf{v}^{\hat{\pi}_N}\|_{\infty,1/\mathbf{w}} \\
&\leq \min_{(\mathbf{v}_\theta, \boldsymbol{\mu}_\theta) \in \mathcal{X}_\Theta} \frac{1}{1-\gamma} \Big(\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}^*\|_1 + 2\|\mathbf{v}_\theta - \mathbf{v}^{\hat{\pi}_N}\|_\infty\Big).
\end{aligned}$$

where the norms are defined as $\|\mathbf{x}\|_{1,\mathbf{w}} = \sum_i w_i |x_i|$ and $\|\mathbf{x}\|_{\infty,1/\mathbf{w}} = \max_i w_i^{-1} |x_i|$.

Proof. For shorthand, let us set $x = (\mathbf{v}, \boldsymbol{\mu}) = \hat{x}_N$ and write also $\pi_{\boldsymbol{\mu}} = \hat{\pi}_N$ as the associated policy. Let $y_x^* = (\mathbf{v}^{\pi_\mu}, \boldsymbol{\mu}^*)$ and similarly let $x_\theta = (\mathbf{v}_\theta, \boldsymbol{\mu}_\theta) \in \mathcal{X}_\Theta$. With $r_{ep}(x; x') = -F(x, x')$ and (14), we can write

$$\begin{aligned} r_{ep}(x; y_x^*) - r_{ep}(x; x_\theta) &= \left(-\mathbf{p}^\top \mathbf{v}^{\pi_\mu} - \boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}^{\pi_\mu}} + \mathbf{p}^\top \mathbf{v} + \boldsymbol{\mu}^{*\top} \mathbf{a}_{\mathbf{v}}\right) - \left(-\mathbf{p}^\top \mathbf{v}_\theta - \boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}_\theta} + \mathbf{p}^\top \mathbf{v} + \boldsymbol{\mu}_\theta^\top \mathbf{a}_{\mathbf{v}}\right) \\ &= \mathbf{p}^\top (\mathbf{v}_\theta - \mathbf{v}^{\pi_\mu}) + (\boldsymbol{\mu}^* - \boldsymbol{\mu}_\theta)^\top \mathbf{a}_{\mathbf{v}} + \boldsymbol{\mu}^\top (\mathbf{a}_{\mathbf{v}_\theta} - \mathbf{a}_{\mathbf{v}^{\pi_\mu}}) \\ &= \mathbf{b}_{\boldsymbol{\mu}}^\top (\mathbf{v}_\theta - \mathbf{v}^{\pi_\mu}) + (\boldsymbol{\mu}^* - \boldsymbol{\mu}_\theta)^\top \mathbf{a}_{\mathbf{v}} \end{aligned}$$

Next we quantize the size of $\mathbf{a}_{\mathbf{v}}$ and $\mathbf{b}_{\boldsymbol{\mu}}$.

Lemma 5. *For $(\mathbf{v}, \boldsymbol{\mu}) \in \mathcal{X}$, $\|\mathbf{a}_{\mathbf{v}}\|_\infty \le \frac{1}{1-\gamma}$ and $\|\mathbf{b}_{\boldsymbol{\mu}}\|_1 \le \frac{2}{1-\gamma}$.*

Proof of Lemma 5. Let Δ denote the set of distributions

$$\begin{aligned} \|\mathbf{a}_{\mathbf{v}}\|_\infty &= \frac{1}{1-\gamma} \|(1-\gamma)\mathbf{r} + \gamma \mathbf{P}\mathbf{v} - \mathbf{E}\mathbf{v}\|_\infty \le \frac{1}{1-\gamma} \max_{a,b \in [0,1]} |a - b| \le \frac{1}{1-\gamma} \\ \|\mathbf{b}_{\boldsymbol{\mu}}\|_1 &= \frac{1}{1-\gamma} \|(1-\gamma)\mathbf{p} + \gamma \mathbf{P}^\top \boldsymbol{\mu} - \mathbf{E}^\top \boldsymbol{\mu}\|_1 \le \frac{1}{1-\gamma} \max_{\mathbf{q}, \mathbf{q}' \in \Delta} \|\mathbf{q} - \mathbf{q}'\|_1 \le \frac{2}{1-\gamma} \end{aligned}$$

□

Therefore, we have preliminary upper bounds:

$$\begin{aligned} (\boldsymbol{\mu}^* - \boldsymbol{\mu}_\theta)^\top \mathbf{a}_{\mathbf{v}} &\le \|\mathbf{a}_{\mathbf{v}}\|_\infty \|\boldsymbol{\mu}^* - \boldsymbol{\mu}_\theta\|_1 \le \frac{1}{1-\gamma} \|\boldsymbol{\mu}^* - \boldsymbol{\mu}_\theta\|_1 \\ \mathbf{b}_{\boldsymbol{\mu}}^\top (\mathbf{v}_\theta - \mathbf{v}^{\pi_\mu}) &\le \|\mathbf{b}_{\boldsymbol{\mu}}\|_1 \|\mathbf{v}_\theta - \mathbf{v}^{\pi_\mu}\|_\infty \le \frac{2}{1-\gamma} \|\mathbf{v}_\theta - \mathbf{v}^{\pi_\mu}\|_\infty \end{aligned}$$

However, the second inequality above can be very conservative, especially when $\mathbf{b}_{\boldsymbol{\mu}} \approx 0$ which can be likely when it is close to the end of policy optimization. To this end, we introduce a free vector $\mathbf{w} \ge 1$. Define norms $\|\mathbf{v}\|_{\infty,1/\mathbf{w}} = \max_i \frac{|v_i|}{w_i}$ and $\|\boldsymbol{\delta}\|_{1,\mathbf{w}} = \sum_i w_i |\delta_i|$. Then we can instead have an upper bound

$$\mathbf{b}_{\boldsymbol{\mu}}^\top (\mathbf{v}_\theta - \mathbf{v}^{\pi_\mu}) \le \min_{\mathbf{w}: \mathbf{w} \ge 1} \|\mathbf{b}_{\boldsymbol{\mu}}\|_{1,\mathbf{w}} \|\mathbf{v}_\theta - \mathbf{v}^{\pi_\mu}\|_{\infty,1/\mathbf{w}}$$

Notice that when $\mathbf{w} = \mathbf{1}$ the above inequality reduces to $\mathbf{b}_{\boldsymbol{\mu}}^\top (\mathbf{v}_\theta - \mathbf{v}^{\pi_\mu}) \le \|\mathbf{b}_{\boldsymbol{\mu}}\|_1 \|\mathbf{v}_\theta - \mathbf{v}^{\pi_\mu}\|_\infty$, which as we showed has an upper bound $\frac{2}{1-\gamma} \|\mathbf{v}_\theta - \mathbf{v}^{\pi_\mu}\|_\infty$.

Combining the above upper bounds, we have an upper bound on $\epsilon_{\Theta,N}$:

$$\begin{aligned} \epsilon_{\Theta,N} = r_{ep}(x; y_x^*) - r_{ep}(x; x_\theta) &\le \frac{1}{1-\gamma} \|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}^*\|_1 + \min_{\mathbf{w}: \mathbf{w} \ge 1} \|\mathbf{b}_{\boldsymbol{\mu}}\|_{1,\mathbf{w}} \|\mathbf{v}_\theta - \mathbf{v}^{\pi_\mu}\|_{\infty,1/\mathbf{w}} \\ &\le \frac{1}{1-\gamma} \left(\|\boldsymbol{\mu}_\theta - \boldsymbol{\mu}^*\|_1 + 2\|\mathbf{v}_\theta - \mathbf{v}^{\pi_\mu}\|_\infty\right). \end{aligned}$$

Since it holds for any $\theta \in \Theta$, we can minimize the right-hand side over all possible choices. □

D Proof of Sample Complexity of Mirror Descent

Theorem 2. *With probability $1 - \delta$, Algorithm 1 learns an ϵ-optimal policy with $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}| \log(\frac{1}{\delta})}{(1-\gamma)^2 \epsilon^2}\right)$ samples.*

The proof is a combination of the basic property of mirror descent (Lemma 9) and the martingale concentration. Define $K = |\mathcal{S}||\mathcal{A}|$ and $\kappa = \frac{1}{1-\gamma}$ as shorthands. We first slightly modify the per-round loss used to compute the gradient. Recall $l_n(x) \coloneqq \mathbf{p}^\top \mathbf{v} + \boldsymbol{\mu}_n^\top \mathbf{a}_{\mathbf{v}} - \mathbf{p}^\top \mathbf{v}_n - \boldsymbol{\mu}^\top \mathbf{a}_{\mathbf{v}_n}$ and let us consider instead a loss function

$$h_n(x) \coloneqq \mathbf{b}_{\boldsymbol{\mu}_n}^\top \mathbf{v} + \boldsymbol{\mu}^\top (\kappa \mathbf{1} - \mathbf{a}_{\mathbf{v}_n})$$

which shifts l_n by a constant in each round. (Note for all the decisions $(\mathbf{v}_n, \boldsymbol{\mu}_n)$ produced by Algorithm 1 $\boldsymbol{\mu}_n$ always satisfies $\|\boldsymbol{\mu}_n\|_1 = 1$). One can verify that $l_n(x) - l_n(x') = h_n(x) - h_n(x')$, for all $x, x' \in \mathcal{X}$, when $\boldsymbol{\mu}, \boldsymbol{\mu}'$ in x and x' satisfy $\|\boldsymbol{\mu}\|_1 = \|\boldsymbol{\mu}'\|_1$ (which holds for Algorithm 1). As the definition of regret is relative, we may work with h_n in online learning and use it to define the feedback.

The reason for using h_n instead of l_n is to make $\nabla_{\boldsymbol{\mu}} h_n((\mathbf{v}, \boldsymbol{\mu}))$ (and its unbiased approximation) a positive vector (because $\kappa \geq \|\mathbf{a}_{\mathbf{v}}\|_\infty$ for any $\mathbf{v} \in \mathcal{V}$), so that the regret bound can have a better dependency on the dimension for learning $\boldsymbol{\mu}$ that lives in the simplex $\mathcal{M}$. This is a common trick used in the online learning, e.g. in EXP3.

To run mirror descent, we set the first-order feedback g_n received by the learner as an unbiased estimate of $\nabla h_n(x_n)$. For round n, we construct g_n based on *two* calls of the generative model:

$$g_n = \begin{bmatrix} \mathbf{g}_{n,v} \\ \mathbf{g}_{n,\mu} \end{bmatrix} = \begin{bmatrix} \tilde{\mathbf{p}}_n + \frac{1}{1-\gamma}(\gamma \tilde{\mathbf{P}}_n - \mathbf{E}_n)^\top \tilde{\boldsymbol{\mu}}_n \\ K(\kappa \hat{\mathbf{1}}_n - \hat{\mathbf{r}}_n - \frac{1}{1-\gamma}(\gamma \hat{\mathbf{P}}_n - \hat{\mathbf{E}}_n)\mathbf{v}_n) \end{bmatrix}$$

For $\mathbf{g}_{n,v}$, we sample $\mathbf{p}$, then sample $\boldsymbol{\mu}_n$ to get a state-action pair, and finally query the transition dynamics $\mathbf{P}$ at the state-action pair sampled from $\boldsymbol{\mu}_n$. ($\tilde{\mathbf{p}}_n$, $\tilde{\mathbf{P}}_n$, and $\tilde{\boldsymbol{\mu}}_n$ denote the single-sample empirical approximation of these probabilities.) For $\mathbf{g}_{n,\mu}$, we first sample *uniformly* a state-action pair (which explains the factor K), and then query the reward $\mathbf{r}$ and the transition dynamics $\mathbf{P}$. ($\hat{\mathbf{1}}_n$, $\hat{\mathbf{r}}_n$, $\hat{\mathbf{P}}_n$, and $\hat{\mathbf{E}}_n$ denote the single-sample empirical estimates.) To emphasize, we use $\tilde{}$ and $\hat{}$ to distinguish the empirical quantities obtained by these two independent queries. By construction, we have $\mathbf{g}_{n,\mu} \geq 0$. It is clear that this direction g_n is unbiased, i.e. $\mathbb{E}[g_n] = \nabla h_n(x_n)$. Moreover, it is extremely sparse and can be computed using $O(1)$ sample, computational, and memory complexities.

Let $y_N^* = (\mathbf{v}^{\hat{\pi}_N}, \boldsymbol{\mu}^*)$. We bound the regret by the following rearrangement.

$$\begin{aligned}
\text{Regret}_N(y_N^*) &= \sum_{n=1}^N l_n(x_n) - \sum_{n=1}^N l_n(y_N^*) \\
&= \sum_{n=1}^N h_n(x_n) - \sum_{n=1}^N h_n(y_N^*) \\
&= \sum_{n=1}^N \nabla h_n(x_n)^\top (x_n - y_N^*) \\
&= \left(\sum_{n=1}^N (\nabla h_n(x_n) - g_n)^\top x_n\right) + \left(\sum_{n=1}^N g_n^\top (x_n - y_N^*)\right) + \left(\sum_{n=1}^N (g_n - \nabla h_n(x_n))^\top y_N^*\right) \\
&\leq \left(\sum_{n=1}^N (\nabla h_n(x_n) - g_n)^\top x_n\right) + \left(\max_{x \in \mathcal{X}} \sum_{n=1}^N g_n^\top (x_n - x)\right) + \left(\sum_{n=1}^N (g_n - \nabla h_n(x_n))^\top y_N^*\right), \qquad (29)
\end{aligned}$$

where the third equality comes from h_n being linear. We recognize the first term is a martingale $M_N = \sum_{n=1}^N (\nabla h_n(x_n) - g_n)^\top x_n$, because x_n does not depend on g_n. Therefore, we can appeal to standard martingale concentration property. For the second term, it can be upper bounded by standard regret analysis of mirror descent, by treating $g_n^\top x$ as the per-round loss. For the third term, because $y_N^* = (\mathbf{v}^{\hat{\pi}_N}, \boldsymbol{\mu}^*)$ depends on $\{g_n\}_{n=1}^N$, it is not a martingale. Nonetheless, we will be able to handle it through a union bound. Below, we give details for bounding these three terms.

D.1 The First Term: Martingale Concentration

For the first term, $\sum_{n=1}^N (\nabla h_n(x_n) - g_n)^\top x_n$, we use a martingale concentration property. Specifically, we adopt a Bernstein-type inequality (McDiarmid, 1998, Theorem 3.15):

Lemma 6. (McDiarmid, 1998, Theorem 3.15) *Let $M_0, \ldots, M_N$ be a martingale and let $F_0 \subseteq F_1 \subseteq \cdots \subseteq F_n$ be the filtration such that $M_n = \mathbb{E}_{|F_n}[M_{n+1}]$. Suppose there are $b, \sigma < \infty$ such that for all n, given F_{n-1}, $M_n - M_{n-1} \leq b$, and $\mathbb{V}_{|F_{n-1}}[M_n - M_{n-1}] \leq \sigma^2$ almost surely. Then for any $\epsilon \geq 0$,*

$$P(M_N - M_0 \geq \epsilon) \leq \exp\left(\frac{-\epsilon^2}{2N\sigma^2(1 + \frac{b\epsilon}{3N\sigma^2})}\right).$$

Lemma 6 implies, with probability at least $1-\delta$,

$$M_N - M_0 \le \sqrt{2N\sigma^2(1+o(1))\log\left(\frac{1}{\delta}\right)},$$

where $o(1)$ means convergence to 0 as $N \to \infty$.

To apply Lemma 6, we need to provide bounds on the properties of the martingale difference:

$$\begin{aligned} M_n - M_{n-1} &= (\nabla h_n(x_n) - g_n)^\top x_n \\ &= (\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n} - \mathbf{g}_{n,\mu})^\top \boldsymbol{\mu}_n + (\mathbf{b}_{\boldsymbol{\mu}_n} - \mathbf{g}_{n,v})^\top \mathbf{v}_n. \end{aligned}$$

For the first term $(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n} - \mathbf{g}_{n,\mu})^\top \boldsymbol{\mu}_n$, we use the lemma below:

Lemma 7. *Let $\boldsymbol{\mu} \in \mathcal{M}$ be arbitrary, chosen independently from the randomness of $\mathbf{g}_{n,\mu}$ when F_{n-1} is given. Then it holds $|(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n} - \mathbf{g}_{n,\mu})^\top \boldsymbol{\mu}| \le \frac{2(1+K)}{1-\gamma}$ and $\mathbb{V}_{|F_{n-1}}[(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n} - \mathbf{g}_{n,\mu})^\top \boldsymbol{\mu}] \le \frac{4K}{(1-\gamma)^2}$.*

Proof. By triangular inequality,

$$|(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n} - \mathbf{g}_{n,\mu})^\top \boldsymbol{\mu}| \le |(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n})^\top \boldsymbol{\mu}| + |\mathbf{g}_{n,\mu}^\top \boldsymbol{\mu}|.$$

For the deterministic part, using Lemma 5 and Hölder's inequality,

$$|(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n})^\top \boldsymbol{\mu}| \le \kappa + \|\mathbf{a}_{\mathbf{v}_n}\|_\infty \|\boldsymbol{\mu}\|_1 \le \frac{2}{1-\gamma}.$$

For the stochastic part, let i_n be index of the sampled state-action pair and j_n be the index of the transited state sampled at the pair given by i_n. With abuse of notation, we will use i_n to index both $\mathcal{S}\times\mathcal{A}$ and $\mathcal{S}$. With this notation, we may derive

$$\begin{aligned} |\mathbf{g}_{n,\mu}^\top \boldsymbol{\mu}| &= |K\boldsymbol{\mu}^\top(\kappa\hat{\mathbf{1}}_n - \hat{\mathbf{r}}_n - \frac{1}{1-\gamma}(\gamma\hat{\mathbf{P}}_n - \hat{\mathbf{E}}_n)\mathbf{v}_n)| \\ &= K\mu_{i_n}|\kappa - r_{i_n} - \frac{\gamma v_{n,j_n} - v_{n,i_n}}{1-\gamma}| \\ &\le \frac{2K\mu_{i_n}}{1-\gamma} \le \frac{2K}{1-\gamma} \end{aligned}$$

where we use the facts that $r_{i_n}, v_{n,j_n}, v_{n,i_n} \in [0,1]$ and $\mu_{i_n} \le 1$.

For $\mathbb{V}_{|F_{n-1}}[(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n} - \mathbf{g}_{n,\mu})^\top \boldsymbol{\mu}_n]$, we can write it as

$$\begin{aligned} \mathbb{V}_{|F_{n-1}}[(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n} - \mathbf{g}_{n,\mu})^\top \boldsymbol{\mu}] &= \mathbb{V}_{|F_{n-1}}[\mathbf{g}_{n,\mu}^\top \boldsymbol{\mu}] \\ &\le \mathbb{E}_{|F_{n-1}}[|\mathbf{g}_{n,\mu}^\top \boldsymbol{\mu}_n|^2] \\ &= \sum_{i_n} \frac{1}{K}\mathbb{E}_{j_n|i_n}\left[K^2\mu_{i_n}^2\left(\kappa - r_{i_n} - \frac{\gamma v_{n,j_n} - v_{n,i_n}}{1-\gamma}\right)^2\right] \\ &\le \frac{4K}{(1-\gamma)^2}\sum_{i_n}\mu_{i_n}^2 \\ &\le \frac{4K}{(1-\gamma)^2}\left(\sum_{i_n}\mu_{i_n}\right)^2 \le \frac{4K}{(1-\gamma)^2} \end{aligned}$$

where in the second inequality we use the fact that $|\kappa - r_{i_n} - \frac{\gamma v_{n,j_n} - v_{n,i_n}}{1-\gamma}| \le \frac{2}{1-\gamma}$. □

For the second term $(\mathbf{b}_{\boldsymbol{\mu}_n} - \mathbf{g}_{n,v})^\top \mathbf{v}_n$, we use the following lemma.

Lemma 8. *Let $\mathbf{v} \in \mathcal{V}$ be arbitrary, chosen independently from the randomness of $\mathbf{g}_{n,v}$ when F_{n-1} is given.. Then it holds that $|(\mathbf{b}_{\boldsymbol{\mu}_n} - \mathbf{g}_{n,v})^\top \mathbf{v}| \le \frac{4}{1-\gamma}$ and $\mathbb{V}_{|F_{n-1}}[(\mathbf{b}_{\boldsymbol{\mu}_n} - \mathbf{g}_{n,v})^\top \mathbf{v}] \le \frac{4}{(1-\gamma)^2}$.*

Proof. We appeal to Lemma 5, which shows $\|\mathbf{b}_{\boldsymbol{\mu}_n}\|_1, \|\mathbf{g}_{n,v}\|_1 \le \frac{2}{1-\gamma}$, and derive

$$|(\mathbf{b}_{\boldsymbol{\mu}_n} - \mathbf{g}_{n,v})^\top \mathbf{v}| \le (\|\mathbf{b}_{\boldsymbol{\mu}_n}\|_1 + \|\mathbf{g}_{n,v}\|_1)\|\mathbf{v}\|_\infty \le \frac{4}{1-\gamma}.$$

Similarly, for the variance, we can write

$$\mathbb{V}_{|F_{n-1}}[(\mathbf{b}_{\boldsymbol{\mu}_n} - \mathbf{g}_{n,v})^\top \mathbf{v}] = \mathbb{V}_{|F_{n-1}}[\mathbf{g}_{n,v}^\top \mathbf{v}] \le \mathbb{E}_{|F_{n-1}}[(\mathbf{g}_{n,v}^\top \mathbf{v})^2] \le \frac{4}{(1-\gamma)^2}.$$ □

Thus, with helps from the two lemmas above, we are able to show

$$M_n - M_{n-1} \le |(\kappa \mathbf{1} - \mathbf{a}_{\mathbf{v}_n} - \mathbf{g}_{n,\mu})^\top \boldsymbol{\mu}_n| + |(\mathbf{b}_{\boldsymbol{\mu}_n} - \mathbf{g}_{n,v})^\top \mathbf{v}_n| \le \frac{4 + 2(1+K)}{1-\gamma}$$

as well as (because $\mathbf{g}_{n,\mu}$ and $\mathbf{g}_{n,b}$ are computed using independent samples)

$$\mathbb{V}_{|F_{n-1}}[M_n - M_{n-1}] \le \mathbb{E}_{|F_{n-1}}[|(\kappa \mathbf{1} - \mathbf{a}_{\mathbf{v}_n} - \mathbf{g}_{n,\mu})^\top \boldsymbol{\mu}_n|^2] + \mathbb{E}_{|F_{n-1}}[|(\mathbf{b}_{\boldsymbol{\mu}_n} - \mathbf{g}_{n,v})^\top \mathbf{v}_n|^2] \le \frac{4(1+K)}{(1-\gamma)^2}$$

Now, since $M_0 = 0$, by martingale concentration in Lemma 6, we have

$$P\left(\sum_{n=1}^N (\nabla h_n(x_n) - g_n)^\top x_n > \epsilon\right) \le \exp\left(\frac{-\epsilon^2}{2N\sigma^2(1 + \frac{b\epsilon}{3N\sigma^2})}\right)$$

with $b = \frac{6+2K}{1-\gamma}$ and $\sigma^2 = \frac{4(1+K)}{(1-\gamma)^2}$. This implies that, with probability at least $1-\delta$, it holds

$$\sum_{n=1}^N (\nabla h_n(x_n) - g_n)^\top x_n \le \sqrt{N \frac{8(1+K)}{(1-\gamma)^2}(1 + o(1)) \log\left(\frac{1}{\delta}\right)} = \tilde{O}\left(\frac{\sqrt{NK \log(\frac{1}{\delta})}}{1-\gamma}\right)$$

D.2 Static Regret of Mirror Descent

Next we move onto deriving the regret bound of mirror descent with respect to the online loss sequence:

$$\max_{x \in \mathcal{X}} \sum_{n=1}^N g_n^\top (x_n - x)$$

This part is quite standard; nonetheless, we provide complete derivations below.

We first recall a basic property of mirror descent

Lemma 9. *Let $\mathcal{X}$ be a convex set. Suppose R is 1-strongly convex with respect to some norm $\|\cdot\|$. Let g be an arbitrary vector and define, for $x \in \mathcal{X}$,*

$$y = \arg\min_{x' \in \mathcal{X}} \langle g, x' \rangle + B_R(x'||x)$$

Then for all $z \in \mathcal{X}$,

$$\langle g, y - z \rangle \le B_R(z||x) - B_R(z||y) - B_R(y||x) \tag{30}$$

Proof. Recall the definition $B_R(x'||x) = R(x') - R(x) - \langle \nabla R(x), x' - x \rangle$. The optimality of the proximal map can be written as

$$\langle g + \nabla R(y) - \nabla R(x), y - z \rangle \le 0, \qquad \forall z \in \mathcal{X}.$$

By rearranging the terms, we can rewrite the above inequality in terms of Bregman divergences as follows and derive the first inequality (30):

$$\begin{aligned}
\langle g, y - z \rangle &\le \langle \nabla R(x) - \nabla R(y), y - z \rangle \\
&= B_R(z||x) - B_R(z||y) + \langle \nabla R(x) - \nabla R(y), y \rangle - \langle \nabla R(x), x \rangle + \langle \nabla R(y), y \rangle + R(x) - R(y) \\
&= B_R(z||x) - B_R(z||y) + \langle \nabla R(x), y - x \rangle + R(x) - R(y) \\
&= B_R(z||x) - B_R(z||y) - B_R(y||x),
\end{aligned}$$

which concludes the lemma. □

Let $x' \in \mathcal{X}$ be arbitrary. Applying this lemma to the nth iteration of mirror descent in (20), we get

$$\langle g_n, x_{n+1} - x' \rangle \leq \frac{1}{\eta} \left(B_R(x'||x_n) - B_R(x'||x_{n+1}) - B_R(x_{n+1}||x_n) \right)$$

By a telescoping sum, we then have

$$\sum_{n=1}^{N} \langle g_n, x_n - x' \rangle \leq \frac{1}{\eta} B_R(x'||x_1) + \sum_{n=1}^{N} \langle g_n, x_{n+1} - x_n \rangle - \frac{1}{\eta} B_R(x_{n+1}||x_n).$$

We bound the right-hand side as follows. Recall that based on the geometry of $\mathcal{X} = \mathcal{V} \times \mathcal{M}$, we considered a natural Bregman divergence of the form:

$$B_R(x'||x) = \frac{1}{2|\mathcal{S}|} \|\mathbf{v}' - \mathbf{v}\|_2^2 + KL(\boldsymbol{\mu}'||\boldsymbol{\mu})$$

Let $x_1 = (\mathbf{v}_1, \boldsymbol{\mu}_1)$ where $\boldsymbol{\mu}_1$ is uniform. By this choice, we have:

$$\frac{1}{\eta} B_R(x'||x_1) \leq \frac{1}{\eta} \max_{x \in \mathcal{X}} B_R(x||x_1) \leq \frac{1}{\eta} \left(\frac{1}{2} + \log(K) \right).$$

We now decompose each item in the above sum as:

$$\begin{aligned} \langle g_n, x_{n+1} - x_n \rangle - \frac{1}{\eta} B_R(x_{n+1}||x_n) = &\left(\mathbf{g}_{n,v}^\top (\mathbf{v}_{n+1} - \mathbf{v}_n) - \frac{1}{2\eta|\mathcal{S}|} \|\mathbf{v}_n - \mathbf{v}_{n+1}\|_2^2 \right) \\ &+ \left(\mathbf{g}_{n,\mu}^\top (\boldsymbol{\mu}_{n+1} - \boldsymbol{\mu}_n) - \frac{1}{\eta} KL(\boldsymbol{\mu}_{n+1}||\boldsymbol{\mu}_n) \right) \end{aligned}$$

and we upper bound them using the two lemmas below (recall $\mathbf{g}_{n,\mu} \geq 0$ due to the added $\kappa \mathbf{1}$ term).

Lemma 10. *For any vector x, y, g and scalar $\eta > 0$, it holds $\langle g, x - y \rangle - \frac{1}{2\eta} \|x - y\|_2^2 \leq \frac{\eta \|g\|_2^2}{2}$.*

Proof. By Cauchy-Swartz inequality, $\langle g, x - y \rangle - \frac{1}{2\eta} \|x - y\|_2^2 \leq \|g\|_2 \|x - y\|_2 - \frac{1}{2\eta} \|x - y\|_2^2 \leq \frac{\eta \|g\|_2^2}{2}$. □

Lemma 11. *Suppose $B_R(x||y) = KL(x||y)$ and x, y are probability distributions, and $g \geq 0$ element-wise. Then, for $\eta > 0$,*

$$-\frac{1}{\eta} B_R(y||x) + \langle g, x - y \rangle \leq \frac{\eta}{2} \sum_i x_i (g_i)^2 = \frac{\eta}{2} \|g\|_x^2.$$

Proof. Let Δ denotes the unit simplex.

$$\begin{aligned} -B_R(y||x) + \langle \eta g, x - y \rangle &\leq \langle \eta g, x \rangle + \max_{y' \in \Delta} \langle -\eta g, y \rangle - B_R(y'||x) \\ &= \langle \eta g, x \rangle + \log \left(\sum_i x_i \exp(-\eta g_i) \right) && (\because \text{convex conjugate of KL divergence}) \\ &\leq \langle \eta g, x \rangle + \log \left(\sum_i x_i \left(1 - \eta g_i + \frac{1}{2} (\eta g_i)^2 \right) \right) && (\because e^x \leq 1 + x + \frac{1}{2} x^2 \text{ for } x \leq 0) \\ &= \langle \eta g, x \rangle + \log \left(1 + \sum_i x_i \left(-\eta g_i + \frac{1}{2} (\eta g_i)^2 \right) \right) \\ &\leq \langle \eta g, x \rangle + \sum_i x_i \left(-\eta g_i + \frac{1}{2} (\eta g_i)^2 \right) && (\because \log(x) \leq x - 1) \\ &= \frac{1}{2} \sum_i x_i (\eta g_i)^2 = \frac{\eta^2}{2} \|g\|_x^2. \end{aligned}$$

Finally, dividing both sides by η, we get the desired result. □

Thus, we have the upper bound $\langle g_n, x_{n+1} - x_n\rangle - \frac{1}{\eta}B_R(x_{n+1}||x_n) = \frac{\eta|\mathcal{S}|\|\mathbf{g}_{n,v}\|_2^2}{2} + \frac{\eta\|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2}{2}$. Together with the upper bound on $\frac{1}{\eta}B_R(x'||x_1)$, it implies that

$$\begin{aligned}\sum_{n=1}^N \langle g_n, x_n - x'\rangle &\le \frac{1}{\eta}B_R(x'||x_1) + \sum_{n=1}^N \langle g_n, x_{n+1} - x_n\rangle - \frac{1}{\eta}B_R(x_{n+1}||x_n)\\ &\le \frac{1}{\eta}\left(\frac{1}{2} + \log(K)\right) + \frac{\eta}{2}\sum_{n=1}^N |\mathcal{S}|\|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2. \end{aligned} \tag{31}$$

We can expect, with high probability, $\sum_{n=1}^N |\mathcal{S}|\|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2$ concentrates toward its expectation, i.e.

$$\sum_{n=1}^N |\mathcal{S}|\|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2 \le \sum_{n=1}^N \mathbb{E}[|\mathcal{S}|\|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2] + o(N).$$

Below we quantify this relationship using martingale concentration. First we bound the expectation.

Lemma 12. $\mathbb{E}[\|\mathbf{g}_{n,v}\|_2^2] \le \frac{4}{(1-\gamma)^2}$ *and* $\mathbb{E}[\|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2] \le \frac{4K}{(1-\gamma)^2}$.

Proof. For the first statement, using the fact that $\|\cdot\|_2 \le \|\cdot\|_1$ and Lemma 5, we can write

$$\mathbb{E}[\|\mathbf{g}_{n,v}\|_2^2] \le \mathbb{E}[\|\mathbf{g}_{n,v}\|_1^2] = \mathbb{E}[\|\tilde{\mathbf{p}}_n + \frac{1}{1-\gamma}(\gamma\tilde{\mathbf{P}}_n - \mathbf{E}_n)^\top\tilde{\boldsymbol{\mu}}_n\|_1^2] \le \frac{4}{(1-\gamma)^2}.$$

For the second statement, let i_n be the index of the sampled state-action pair and j_n be the index of the transited-to state sampled at the pair given by i_n. With abuse of notation, we will use i_n to index both $\mathcal{S}\times\mathcal{A}$ and $\mathcal{S}$.

$$\begin{aligned}\mathbb{E}[\|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2] &= \mathbb{E}\left[\sum_{i_n}\frac{1}{K}\mathbb{E}_{j_n|i_n}\left[K^2\mu_{i_n}\left(\kappa - r_{i_n} - \frac{\gamma v_{n,j_n} - v_{n,i_n}}{1-\gamma}\right)^2\right]\right]\\ &\le \frac{4K}{(1-\gamma)^2}\mathbb{E}\left[\sum_{i_n}\mu_{i_n}\right] \le \frac{4K}{(1-\gamma)^2}.\end{aligned}$$

□

To bound the tail, we resort to the Höffding-Azuma inequality of martingale (McDiarmid, 1998, Theorem 3.14).

Lemma 13 (Azuma-Hoeffding)**.** *Let $M_0, \dots, M_N$ be a martingale and let $F_0 \subseteq F_1 \subseteq \cdots \subseteq F_n$ be the filtration such that $M_n = \mathbb{E}_{|F_n}[M_{n+1}]$. Suppose there exists $b < \infty$ such that for all n, given F_{n-1}, $|M_n - M_{n-1}| \le b$. Then for any $\epsilon \ge 0$,*

$$P(M_N - M_0 \ge \epsilon) \le \exp\left(\frac{-2\epsilon^2}{Nb^2}\right)$$

To apply Lemma 13, we consider the martingale

$$M_N = \sum_{n=1}^N |\mathcal{S}|\|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2 - \left(\sum_{n=1}^N \mathbb{E}[|\mathcal{S}|\|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2]\right)$$

To bound the change of the size of martingale difference $|M_n - M_{n-1}|$, we follow similar steps to Lemma 12.

Lemma 14. $\|\mathbf{g}_{n,v}\|_2^2 \le \frac{4}{(1-\gamma)^2}$ *and* $\|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2 \le \frac{4K^2}{(1-\gamma)^2}$.

Note $\|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}}^2$ is K-factor larger than $\mathbb{E}[\|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}}^2])$ and $K \ge 1$. Therefore, we have

$$|M_n - M_{n-1}| \le |\mathcal{S}|\|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2 + |\mathcal{S}|\mathbb{E}[\|\mathbf{g}_{n,v}\|_2^2] + \mathbb{E}[\|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2] \le \frac{8(|\mathcal{S}| + K^2)}{(1-\gamma)^2}$$

Combining these results, we have, with probability as least $1-\delta$,

$$\sum_{n=1}^{N} |\mathcal{S}|\|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2 \le \sum_{n=1}^{N} \mathbb{E}[|\mathcal{S}|\|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2] + \frac{4\sqrt{2}(|\mathcal{S}|+K^2)}{(1-\gamma)^2}\sqrt{N\log\left(\frac{1}{\delta}\right)}$$
$$\le \frac{4(K+|\mathcal{S}|)}{(1-\gamma)^2}N + \frac{4\sqrt{2}(|\mathcal{S}|+K^2)}{(1-\gamma)^2}\sqrt{N\log\left(\frac{1}{\delta}\right)}$$

Now we suppose we set $\eta = \frac{1-\gamma}{\sqrt{KN}}$. From (37), we then have

$$\sum_{n=1}^{N}\langle g_n, x_n - x'\rangle \le \frac{1}{\eta}\left(\frac{1}{2}+\log(K)\right) + \frac{\eta}{2}\sum_{n=1}^{N}|\mathcal{S}|\|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\mu}_n}^2$$
$$\le \frac{\sqrt{KN}}{1-\gamma}\left(\frac{1}{2}+\log(K)\right) + \frac{1-\gamma}{\sqrt{KN}}\left(\frac{2(K+|\mathcal{S}|)}{(1-\gamma)^2}N + \frac{2\sqrt{2}(|\mathcal{S}|+K^2)}{(1-\gamma)^2}\sqrt{N\log\left(\frac{1}{\delta}\right)}\right)$$
$$\le \tilde{O}\left(\frac{\sqrt{KN}}{1-\gamma} + \frac{\sqrt{K^3\log\frac{1}{\delta}}}{1-\gamma}\right).$$

D.3 Union Bound

Lastly, we provide an upper bound on the last component:

$$\sum_{n=1}^{N}(g_n - \nabla h_n(x_n))^\top y_N^*.$$

Because y_N^* depends on g_n, this term does not obey martingale concentration like the first component $\sum_{n=1}^{N}(\nabla h_n(x_n) - g_n)^\top x_n$ which we analyzed in Appendix D.1 To resolve this issue, we utilize the concept of covering number and derive a union bound.

Recall for a compact set $\mathcal{Z}$ in a norm space, the covering number $\mathcal{N}(\mathcal{Z},\epsilon)$ with $\epsilon>0$ is the minimal number of ϵ-balls that covers $\mathcal{Z}$. That is, there is a set $\{z_i\in\mathcal{Z}\}_{i=1}^{\mathcal{N}(\mathcal{Z},\epsilon)}$ such that $\max_{z\in\mathcal{Z}}\min_{z'\in B(\mathcal{Z},\epsilon)}\|z-z'\|\le\epsilon$. Usually the covering number $\mathcal{N}(\mathcal{Z},\epsilon)$ is polynomial in ϵ and perhaps exponential in the ambient dimension of $\mathcal{Z}$.

The idea of covering number can be used to provide a union bound of concentration over compact sets, which we summarize as a lemma below.

Lemma 15. *Let f,g be two random L-Lipschitz functions. Suppose for some $a>0$ and some fixed $z\in\mathcal{Z}$ selected independently of f,g, it holds*

$$P\left(|f(z)-g(z)|>\epsilon\right)\le\exp\left(-a\epsilon^2\right)$$

Then it holds that

$$P\left(\sup_{z\in\mathcal{Z}}|f(z)-g(z)|>\epsilon\right)\le\mathcal{N}\left(\mathcal{Z},\frac{\epsilon}{4L}\right)\exp\left(\frac{-a\epsilon^2}{4}\right)$$

Proof. Let $\mathcal{C}$ denote a set of covers of size $\mathcal{N}(\mathcal{Z},\frac{\epsilon}{4L})$ Then, for any $z\in\mathcal{Z}$ which could depend on f,g,

$$|f(z)-g(z)| \le \min_{z'\in\mathcal{C}}|f(z)-f(z')|+|f(z')-g(z')|+|g(z')-g(z)|$$
$$\le \min_{z'\in\mathcal{C}} 2L\|z-z'\|+|f(z')-g(z')]|$$
$$\le \frac{\epsilon}{2}+\max_{z'\in\mathcal{C}}|f(z')-g(z')|$$

Thus, $\sup_{z\in\mathcal{Z}}|f(z)-g(z)|>\epsilon \implies \max_{z'\in\mathcal{C}}|f(z')-g(z')|>\frac{\epsilon}{2}$. Therefore, we have the union bound.

$$P\left(\sup_{z\in\mathcal{Z}}|f(z)-\mathbb{E}[f(z)]|>\epsilon\right)\le\mathcal{N}\left(\mathcal{Z},\frac{\epsilon}{4L}\right)\exp\left(\frac{-a\epsilon^2}{4}\right).$$ □

We now use Lemma 15 to bound the component $\sum_{n=1}^N (g_n - \nabla h_n(x_n))^\top y_N^*$. We recall by definition, for $x = (\mathbf{v}, \boldsymbol{\mu})$,

$$(\nabla h_n(x_n) - g_n)^\top x = (\kappa \mathbf{1} - \mathbf{a}_{\mathbf{v}_n} - \mathbf{g}_{n,\mu})^\top \boldsymbol{\mu} + (\mathbf{b}_{\boldsymbol{\mu}_n} - \mathbf{g}_{n,v})^\top \mathbf{v}$$

Because $y_N^* = (\mathbf{v}^{\hat{\pi}_N}, \boldsymbol{\mu}^*)$, we can write the sum of interest as

$$\sum_{n=1}^N (g_n - \nabla h_n(x_n))^\top y_N^* = \sum_{n=1}^N (\mathbf{g}_{n,\mu} - \kappa \mathbf{1} + \mathbf{a}_{\mathbf{v}_n})^\top \boldsymbol{\mu}^* + \sum_{n=1}^N (\mathbf{g}_{n,v} - \mathbf{b}_{\boldsymbol{\mu}_n})^\top \mathbf{v}^{\hat{\pi}_N}$$

For the first term, because $\boldsymbol{\mu}^*$ is set beforehand by the MDP definition and does not depend on the randomness during learning, it is a martingale and we can apply the steps in Appendix D.1 to show,

$$\sum_{n=1}^N (\mathbf{g}_{n,\mu} - \kappa \mathbf{1} + \mathbf{a}_{\mathbf{v}_n})^\top \boldsymbol{\mu}^* = \tilde{O}\left(\frac{\sqrt{NK \log(\frac{1}{\delta})}}{1-\gamma} \right)$$

For the second term, because $\mathbf{v}^{\hat{\pi}_N}$ depends on the randomness in the learning process, we need to use a union bound. Following the steps in Appendix D.1, we see that for some fixed $\mathbf{v} \in \mathcal{V}$, it holds

$$P\left(\left| \sum_{n=1}^N (\mathbf{g}_{n,v} - \mathbf{b}_{\boldsymbol{\mu}_n})^\top \mathbf{v} \right| > \epsilon \right) \leq \exp\left(-\frac{(1-\gamma)^2}{N} \epsilon^2 \right)$$

where some constants were ignored for the sake of conciseness. Note also that it does not have the $\sqrt{K}$ factor because of Lemma 8. To apply Lemma 15, we need to know the order of covering number of $\mathcal{V}$. Since $\mathcal{V}$ is an $|\mathcal{S}|$-dimensional unit cube in the positive orthant, it is straightforward to show $\mathcal{N}(\mathcal{V}, \epsilon) \leq \max(1, (1/\epsilon)^{|\mathcal{S}|})$ (by simply discretizing evenly in each dimension). Moreover, the functions $\sum_{n=1}^N \mathbf{g}_{n,v}^\top \mathbf{v}$ and $\sum_{n=1}^N \mathbf{b}_{\boldsymbol{\mu}_n}^\top \mathbf{v}$ are $\frac{N}{1-\gamma}$-Lipschitz in $\|\cdot\|_\infty$.

Applying Lemma 15 then gives us:

$$P\left(\sup_{v \in \mathcal{V}} \left| \sum_{n=1}^N (\mathbf{g}_{n,v} - \mathbf{b}_{\boldsymbol{\mu}_n})^\top \mathbf{v} \right| > \epsilon \right) \leq \mathcal{N}\left(\mathcal{V}, \frac{\epsilon(1-\gamma)}{4N} \right) \exp\left(-\frac{(1-\gamma)^2}{4N} \epsilon^2 \right).$$

For a given δ, we thus want to find the smallest ϵ such that:

$$\delta \geq \mathcal{N}\left(\mathcal{V}, \frac{\epsilon(1-\gamma)}{4N} \right) \exp\left(-\frac{(1-\gamma)^2}{4N} \epsilon^2 \right).$$

That is:

$$\log(\frac{1}{\delta}) \leq \frac{(1-\gamma)^2}{4N} \epsilon^2 + |\mathcal{S}| \min(0, \log(\frac{\epsilon(1-\gamma)}{4N})).$$

Picking $\epsilon = O\left(\log(N) \frac{\sqrt{N \log(\frac{1}{\delta})}}{1-\gamma} \right) = \tilde{O}\left(\frac{\sqrt{N \log(\frac{1}{\delta})}}{1-\gamma} \right)$ guarantees that the inequality is verified asymptotically.

Combining these two steps, we have shown overall, with probability at least $1-\delta$,

$$\sum_{n=1}^N (g_n - \nabla h_n(x_n))^\top y_N^* = \tilde{O}\left(\frac{\sqrt{NK \log(\frac{1}{\delta})}}{1-\gamma} \right).$$

D.4 Summary

In the previous subsections, we have provided high probability upper bounds for each term in the decomposition

$$\text{Regret}_N(y_N^*) \leq \left(\sum_{n=1}^N (\nabla h_n(x_n) - g_n)^\top x_n \right) + \left(\max_{x \in \mathcal{X}} \sum_{n=1}^N g_n^\top (x_n - x) \right) + \left(\sum_{n=1}^N (g_n - \nabla h_n(x_n))^\top y_N^* \right)$$

implying with probability at least $1-\delta$,

$$\text{Regret}_N(y_N^*) \le \tilde{O}\left(\frac{\sqrt{NK\log(\frac{1}{\delta})}}{1-\gamma}\right) + \tilde{O}\left(\frac{\sqrt{KN}}{1-\gamma} + \frac{\sqrt{K^3\log\frac{1}{\delta}}}{1-\gamma}\right) = \tilde{O}\left(\frac{\sqrt{N|\mathcal{S}||\mathcal{A}|\log(\frac{1}{\delta})}}{1-\gamma}\right)$$

By Theorem 1, this would imply with probability at least $1-\delta$,

$$V^{\hat{\pi}_N}(p) \ge V^*(p) - \frac{\text{Regret}_N(y_N^*)}{N} \ge V^*(p) - \tilde{O}\left(\frac{\sqrt{|\mathcal{S}||\mathcal{A}|\log(\frac{1}{\delta})}}{(1-\gamma)\sqrt{N}}\right)$$

In other words, the sample complexity of mirror descent to obtain an ϵ approximately optimal policy (i.e. $V^*(p) - V^{\hat{\pi}_N}(p) \le \epsilon$) is at most $\tilde{O}\left(\frac{|\mathcal{S}||\mathcal{A}|\log(\frac{1}{\delta})}{(1-\gamma)^2\epsilon^2}\right)$.

E Sample Complexity of Mirror Descent with Basis Functions

Here we provide further discussions on the sample complexity of running Algorithm 1 with linearly parameterized function approximators and the proof of Theorem 3.

Theorem 3. *Under a proper choice of Θ and B_R, with probability $1-\delta$, Algorithm 1 learns an $(\epsilon+\epsilon_{\Theta,N})$-optimal policy with $\tilde{O}\left(\frac{dim(\Theta)\log(\frac{1}{\delta})}{(1-\gamma)^2\epsilon^2}\right)$ samples.*

E.1 Setup

We suppose that the decision variable is parameterized in the form $x_\theta = (\mathbf{\Phi}\boldsymbol{\theta}_v, \mathbf{\Psi}\boldsymbol{\theta}_\mu)$, where $\mathbf{\Phi}, \mathbf{\Psi}$ are given nonlinear basis functions and $(\boldsymbol{\theta}_v, \boldsymbol{\theta}_\mu) \in \Theta$ are the parameters to learn. For modeling the value function, we suppose each column in $\mathbf{\Phi}$ is a vector (i.e. function) such that its $\|\cdot\|_\infty$ is less than one. For modeling the state-action distribution, we suppose each column in $\mathbf{\Psi}$ is a state-action distribution from which we can draw samples. This choice implies that every column of $\mathbf{\Phi}$ belongs to $\mathcal{V}$, and every column of $\mathbf{\Psi}$ belongs to $\mathcal{M}$.

Considering the geometry of $\mathbf{\Phi}$ and $\mathbf{\Psi}$, we consider a compact and convex parameter set

$$\Theta = \{\theta = (\boldsymbol{\theta}_v, \boldsymbol{\theta}_\mu) : \|\boldsymbol{\theta}_v\|_2 \le \frac{C_v}{\sqrt{dim(\boldsymbol{\theta}_v)}}, \boldsymbol{\theta}_\mu \ge 0, \|\boldsymbol{\theta}_\mu\|_1 = 1\}$$

where $C_v < \infty$. The constant C_v acts as a regularization in learning: if it is too small, the bias (captured as $\epsilon_{\Theta,N}$ in Corollary 1 restated below) becomes larger; if it is too large, the learning becomes slower.

This choice of Θ makes sure, for $\theta = (\boldsymbol{\theta}_v, \boldsymbol{\theta}_\mu) \in \Theta$, $\mathbf{\Psi}\boldsymbol{\theta}_\mu \in \mathcal{M}$ and $\|\mathbf{\Phi}\boldsymbol{\theta}_v\|_\infty \le \|\boldsymbol{\theta}_v\|_1 \le C_v$. Therefore, by the above construction, we can verify that the requirement in Corollary 1 is satisfied, i.e. for $\theta = (\boldsymbol{\theta}_v, \boldsymbol{\theta}_\mu) \in \Theta$, we have $(\mathbf{\Phi}\boldsymbol{\theta}_v, \mathbf{\Psi}\boldsymbol{\theta}_\mu) \in \mathcal{X}_\Theta$.

Corollary 1. *Let $X_N = \{x_n \in \mathcal{X}_\theta\}_{n=1}^N$ be* any *sequence. Let $\hat{\pi}_N$ be the policy given either by the average or the best decision in X_N. It holds that*

$$V^{\hat{\pi}_N}(p) \ge V^*(p) - \frac{\text{Regret}_N(\Theta)}{N} - \epsilon_{\Theta,N}$$

where $\epsilon_{\Theta,N} = \min_{x_\theta \in \mathcal{X}_\theta} r_{ep}(\hat{x}_N; y_N^*) - r_{ep}(\hat{x}_N; x_\theta)$ measures the expressiveness of X_θ, and $\text{Regret}_N(\Theta) \coloneqq \sum_{n=1}^N l_n(x_n) - \min_{x\in\mathcal{X}_\Theta} \sum_{n=1}^N l_n(x)$.

E.2 Online Loss and Sampled Gradient

Let $\theta = (\boldsymbol{\theta}_v, \boldsymbol{\theta}_\mu) \in \Theta$. In view of the parameterization above, we can identify the online loss in (22) in the parameter space as

$$h_n(\theta) \coloneqq \mathbf{b}_{\boldsymbol{\mu}_n}^\top \mathbf{\Phi}\boldsymbol{\theta}_v + \boldsymbol{\theta}_\mu^\top \mathbf{\Psi}^\top (\tfrac{1}{1-\gamma}\mathbf{1} - \mathbf{a}_{\mathbf{v}_n}) \tag{32}$$

where we have the natural identification $x_n = (\mathbf{v}_n, \boldsymbol{\mu}_n) = (\boldsymbol{\Phi}\boldsymbol{\theta}_{v,n}, \boldsymbol{\Psi}\boldsymbol{\theta}_{\mu,n})$ and $\theta_n = (\boldsymbol{\theta}_{v,n}, \boldsymbol{\theta}_{\mu,n}) \in \Theta$ is the decision made by the online learner in the nth round. Note that because this extension of Algorithm 1 makes sure $\|\boldsymbol{\theta}_{\mu,n}\|_1 = 1$ for every iteration, we can still use h_n. For writing convenience, we will continue to overload h_n as a function of parameter θ in the following analyses.

Mirror descent requires gradient estimates of $\nabla h_n(\theta_n)$. Here we construct an unbiased stochastic estimate of $\nabla h_n(\theta_n)$ as

$$g_n = \begin{bmatrix} \mathbf{g}_{n,v} \\ \mathbf{g}_{n,\mu} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Phi}^\top (\tilde{\mathbf{p}}_n + \frac{1}{1-\gamma}(\gamma \tilde{\mathbf{P}}_n - \mathbf{E}_n)^\top \tilde{\boldsymbol{\mu}}_n) \\ dim(\boldsymbol{\theta}_\mu) \hat{\boldsymbol{\Psi}}_n^\top (\frac{1}{1-\gamma}\hat{\mathbf{i}}_n - \hat{\mathbf{r}}_n - \frac{1}{1-\gamma}(\gamma \hat{\mathbf{P}}_n - \hat{\mathbf{E}}_n)\mathbf{v}_n) \end{bmatrix} \tag{33}$$

using two calls of the generative model (again we overload the symbol g_n for the analyses in this section):

- The upper part $\mathbf{g}_{n,v}$ is constructed similarly as before in (23): First we sample the initial state from the initial distribution, the state-action pair using the learned state-action distribution, and then the transited-to state at the sampled state-action pair. We evaluate $\boldsymbol{\Phi}$'s values at those samples to construct $\mathbf{g}_{n,v}$. Thus, $\mathbf{g}_{n,v}$ would generally be a dense vector of size $dim(\boldsymbol{\theta}_v)$ (unless the columns of $\boldsymbol{\Phi}$ are sparse to begin with).

- The lower part $\mathbf{g}_{n,\mu}$ is constructed slightly differently. Recall for the tabular version in (23), we uniformly sample over the state and action spaces. Here instead we first sample uniformly a column (i.e. a basis function) in $\boldsymbol{\Psi}$ and then sample a state-action pair according to the sampled column, which is a distribution by design. Therefore, the multiplier due to uniform sampling in the second row of (33) is now $dim(\boldsymbol{\theta}_\mu)$ rather than $|\mathcal{S}||\mathcal{A}|$ in (23). The matrix $\hat{\boldsymbol{\Psi}}_n$ is extremely sparse, where only the single sampled entry (the column and the state-action pair) is one and the others are zero. In fact, one can think of the tabular version as simply using basis functions $\boldsymbol{\Psi} = \mathbf{I}$, i.e. each column is a delta distribution. Under this identification, the expression in (33) matches the one in (23).

It is straightforward to verify that $\mathbb{E}[g_n] = \nabla h_n(\theta_n)$ for g_n in (33).

E.3 Proof of Theorem 3

We follow the same steps of the analysis of the tabular version. We will highlight the differences/improvement due to using function approximations.

First, we use Corollary 1 in place of Theorem 1. To properly handle the randomness, we revisit its derivation to slightly tighten the statement, which was simplified for the sake of cleaner exposition. Define

$$y^*_{N,\theta} = (\mathbf{v}^*_{N,\theta}, \boldsymbol{\mu}^*_\theta) \coloneqq \arg\max_{x_\theta \in \mathcal{X}_\theta} r_{ep}(\hat{x}_N; x_\theta).$$

For writing convenience, let us also denote $\theta^*_N = (\boldsymbol{\theta}^*_{v,N}, \boldsymbol{\theta}^*_\mu) \in \Theta$ as the corresponding parameter of $y^*_{N,\theta}$. We remark that $\boldsymbol{\mu}^*_\theta$ (i.e. $\boldsymbol{\theta}^*_\mu$), which tries to approximate $\boldsymbol{\mu}^*$, is fixed before the learning process, whereas $\mathbf{v}^*_{N,\theta}$ (i.e. $\boldsymbol{\theta}^*_{v,N}$) could depend on the stochasticity in the learning. Using this new notation and the steps in the proof of Corollary 1, we can write

$$\begin{aligned} V^*(p) - V^{\hat{\pi}_N}(p) &= r_{ep}(\hat{x}_N; y^*_N) \\ &= \epsilon_{\Theta,N} + r_{ep}(\hat{x}_N; y^*_{N,\theta}) \leq \epsilon_{\Theta,N} + \frac{\text{Regret}_N(y^*_{N,\theta})}{N} \end{aligned}$$

where the first equality is Proposition 4, the last inequality follows the proof of Theorem 1, and we recall the definition $\epsilon_{\Theta,N} = r_{ep}(\hat{x}_N; y^*_N) - r_{ep}(\hat{x}_N; y^*_{N,\theta})$.

The rest of the proof is very similar to that of Theorem 1, because linear parameterization does not change the

convexity of the loss sequence. Let $y_N^* = (\mathbf{v}^{\hat{\pi}_N}, \boldsymbol{\mu}^*)$. We bound the regret by the following rearrangement.

$$\begin{aligned}
\text{Regret}_N(y^*_{N,\theta}) &= \sum_{n=1}^N l_n(x_n) - \sum_{n=1}^N l_n(y^*_{N,\theta}) \\
&= \sum_{n=1}^N h_n(\theta_n) - \sum_{n=1}^N h_n(\theta^*_N) \\
&= \sum_{n=1}^N \nabla h_n(\theta_n)^\top (\theta_n - \theta^*_N) \\
&= \left(\sum_{n=1}^N (\nabla h_n(\theta_n) - g_n)^\top \theta_n\right) + \left(\sum_{n=1}^N g_n^\top (\theta_n - \theta^*_N)\right) + \left(\sum_{n=1}^N (g_n - \nabla h_n(\theta_n))^\top \theta^*_N\right) \\
&\le \left(\sum_{n=1}^N (\nabla h_n(\theta_n) - g_n)^\top \theta_n\right) + \left(\max_{\theta\in\Theta} \sum_{n=1}^N g_n^\top (\theta_n - \theta)\right) + \left(\sum_{n=1}^N (g_n - \nabla h_n(\theta_n))^\top \theta^*_N\right) \qquad (34)
\end{aligned}$$

where the second equality is due to the identifcation in (32).

We will solve this online learning problem with mirror descent

$$\theta_{n+1} = \arg\min_{\theta\in\Theta} \langle g_n, \theta\rangle + \frac{1}{\eta} B_R(\theta||\theta_n) \qquad (35)$$

with step size $\eta > 0$ and a Bregman divergence that is a straightforward extension of (21)

$$B_R(\theta'||\theta) = \tfrac{1}{2}\tfrac{dim(\theta_v)}{C_v^2}\|\boldsymbol{\theta}'_v - \boldsymbol{\theta}_v\|_2^2 + KL(\boldsymbol{\theta}'_\mu||\boldsymbol{\theta}_\mu) \qquad (36)$$

where the constant $\frac{dim(\theta_v)}{C_v^2}$ is chosen to make the size of Bregman divergence dimension-free (at least up to log factors). Below we analyze the size of the three terms in (34) like what we did for Theorem 2.

E.4 The First Term: Martingale Concentration

The first term is a martingale. We will use this part to highlight the different properties due to using basis functions. The proof follows the steps in Appendix D.1, but now the martingale difference of interest is instead

$$\begin{aligned}
M_n - M_{n-1} &= (\nabla h_n(\theta_n) - g_n)^\top \theta_n \\
&= (\boldsymbol{\Psi}^\top(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n}) - \mathbf{g}_{n,\mu})^\top \boldsymbol{\theta}_{\mu,n} + (\boldsymbol{\Phi}^\top \mathbf{b}_{\boldsymbol{\mu}_n} - \mathbf{g}_{n,v})^\top \boldsymbol{\theta}_{v,n}
\end{aligned}$$

They now have nicer properties due to the way $\mathbf{g}_{n,\mu}$ is sampled.

For the first term $(\boldsymbol{\Psi}^\top(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n}) - \mathbf{g}_{n,\mu})^\top \boldsymbol{\theta}_{\mu,n}$, we use the lemma below, where we recall the filtration F_n is naturally defined as $\{\theta_1, \ldots, \theta_n\}$.

Lemma 16. *Let $\theta = (\boldsymbol{\theta}_v, \boldsymbol{\theta}_\mu) \in \Theta$ be arbitrary that is chosen independent of the randomness of $\mathbf{g}_{n,\mu}$ when F_{n-1} is given. Then it holds $|(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n} - \mathbf{g}_{n,\mu})^\top \boldsymbol{\theta}| \le \frac{2(1+dim(\boldsymbol{\theta}_{\boldsymbol{\mu}}))}{1-\gamma}$ and $\mathbb{V}_{|F_{n-1}}[(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n} - \mathbf{g}_{n,\mu})^\top \boldsymbol{\theta}_n] \le \frac{4dim(\boldsymbol{\theta}_{\boldsymbol{\mu}})}{(1-\gamma)^2}$.*

Proof. By triangular inequality,

$$|(\boldsymbol{\Psi}^\top(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n}) - \mathbf{g}_{n,\mu})^\top \boldsymbol{\theta}_\mu| \le |(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n})^\top \boldsymbol{\Psi}\boldsymbol{\theta}_\mu| + |\mathbf{g}_{n,\mu}^\top \boldsymbol{\theta}_\mu|$$

For the deterministic part, using Lemma 5 and Hölder's inequality,

$$|(\kappa\mathbf{1} - \mathbf{a}_{\mathbf{v}_n})^\top \boldsymbol{\Psi}\boldsymbol{\theta}_\mu| \le \kappa + \|\mathbf{a}_{\mathbf{v}_n}\|_\infty \|\boldsymbol{\Psi}\boldsymbol{\theta}_\mu\|_1 \le \frac{2}{1-\gamma}$$

For the stochastic part, let k_n denote the sampled column index, i_n be index of the sampled state-action pair using the column of k_n, and j_n be the index of the transited state sampled at the pair given by i_n. With abuse of

notation, we will use i_n to index both $\mathcal{S}\times\mathcal{A}$ and $\mathcal{S}$. Let $\boldsymbol{\mu}=\boldsymbol{\Psi}\boldsymbol{\theta}_\mu$. With this notation, we may derive

$$\begin{aligned}|\mathbf{g}_{n,\mu}^\top\boldsymbol{\theta}_\mu| &= |dim(\boldsymbol{\theta}_\mu)\boldsymbol{\theta}_{\boldsymbol{\mu}}^\top\hat{\boldsymbol{\Psi}}_n^\top(\kappa\hat{\mathbf{1}}_n-\hat{\mathbf{r}}_n-\frac{1}{1-\gamma}(\gamma\hat{\mathbf{P}}_n-\hat{\mathbf{E}}_n)\mathbf{v}_n)|\\ &= dim(\boldsymbol{\theta}_\mu)\theta_{\mu,k_n}|\kappa-r_{i_n}-\frac{\gamma v_{n,j_n}-v_{n,i_n}}{1-\gamma}|\\ &\le \frac{2dim(\boldsymbol{\theta}_\mu)\theta_{\mu,k_n}}{1-\gamma}\le\frac{2dim(\boldsymbol{\theta}_\mu)}{1-\gamma}\end{aligned}$$

where we use the facts $r_{i_n}, v_{n,j_n}, v_{n,i_n}\in[0,1]$ and $\theta_{\mu,k_n}\le 1$.

Let $\boldsymbol{\psi}_\mu^{(k)}$ denote the kth column of $\boldsymbol{\Psi}$. For $\mathbb{V}_{|F_{n-1}}[(\kappa\mathbf{1}-\mathbf{a}_{\mathbf{v}_n}-\mathbf{g}_{n,\mu})^\top\boldsymbol{\theta}_n]$, we can write it as

$$\begin{aligned}\mathbb{V}_{|F_{n-1}}[(\boldsymbol{\Psi}^\top(\kappa\mathbf{1}-\mathbf{a}_{\mathbf{v}_n})-\mathbf{g}_{n,\mu})^\top\boldsymbol{\theta}_\mu] &= \mathbb{V}_{|F_{n-1}}[\mathbf{g}_{n,\mu}^\top\boldsymbol{\theta}_n]\\ &\le \mathbb{E}_{|F_{n-1}}[|\mathbf{g}_{n,\mu}^\top\boldsymbol{\theta}_n|^2]\\ &= \sum_{k_n}\frac{1}{dim(\boldsymbol{\theta}_\mu)}\sum_{i_n}\psi_{\mu,i_n}^{(k_n)}\mathbb{E}_{j_n|i_n}\left[dim(\boldsymbol{\theta}_\mu)^2\theta_{\mu,k_n}^2\left(\kappa-r_{i_n}-\frac{\gamma v_{n,j_n}-v_{n,i_n}}{1-\gamma}\right)^2\right]\\ &\le \frac{4dim(\boldsymbol{\theta}_\mu)}{(1-\gamma)^2}\sum_{k_n}\theta_{\mu,k_n}^2\sum_{i_n}\psi_{\mu,i_n}^{(k_n)}\\ &\le \frac{4dim(\boldsymbol{\theta}_\mu)}{(1-\gamma)^2}\left(\sum_{k_n}\theta_{\mu,k_n}\right)^2\le\frac{4dim(\boldsymbol{\theta}_\mu)}{(1-\gamma)^2}\end{aligned}$$

where in the second inequality we use the fact that $|\kappa-r_{i_n}-\frac{\gamma v_{n,j_n}-v_{n,i_n}}{1-\gamma}|\le\frac{2}{1-\gamma}$. □

For the second term $(\boldsymbol{\Phi}^\top\mathbf{b}_{\boldsymbol{\mu}_n}-\mathbf{g}_{n,v})^\top\boldsymbol{\theta}_{v,n}$, we use this lemma.

Lemma 17. *Let $\boldsymbol{\theta}\in\mathcal{V}$ be arbitrary that is chosen independent of the randomness of $\mathbf{g}_{n,v}$ when F_{n-1} is given. Then it holds $|(\boldsymbol{\Phi}^\top\mathbf{b}_{\boldsymbol{\mu}_n}-\mathbf{g}_{n,v})^\top\boldsymbol{\theta}|\le\frac{4C_v}{1-\gamma}$ and $\mathbb{V}_{|F_{n-1}}[(\boldsymbol{\Phi}^\top\mathbf{b}_{\boldsymbol{\mu}_n}-\mathbf{g}_{n,v})^\top\boldsymbol{\theta}]\le\frac{4C_v^2}{(1-\gamma)^2}$.*

Proof. We appeal to Lemma 5, which shows $\|\mathbf{b}_{\boldsymbol{\mu}_n}\|_1\le\frac{2}{1-\gamma}$ and

$$\|\tilde{\mathbf{p}}_n+\frac{1}{1-\gamma}(\gamma\tilde{\mathbf{P}}_n-\mathbf{E}_n)^\top\tilde{\boldsymbol{\mu}}_n\|_1\le\frac{2}{1-\gamma}$$

Therefore, overall we can derive

$$|(\boldsymbol{\Phi}^\top\mathbf{b}_{\boldsymbol{\mu}_n}-\mathbf{g}_{n,v})^\top\boldsymbol{\theta}|\le\left(\|\mathbf{b}_{\boldsymbol{\mu}_n}\|_1+\|\tilde{\mathbf{p}}_n+\frac{1}{1-\gamma}(\gamma\tilde{\mathbf{P}}_n-\mathbf{E}_n)^\top\tilde{\boldsymbol{\mu}}_n\|_1\right)\|\boldsymbol{\Phi}\boldsymbol{\theta}_v\|_\infty\le\frac{4C_v}{1-\gamma}$$

where we use again each column in $\boldsymbol{\Phi}$ has $\|\cdot\|_\infty$ less than one, and $\|\cdot\|_\infty\le\|\cdot\|_2$. Similarly, for the variance, we can write

$$\mathbb{V}_{|F_{n-1}}[(\boldsymbol{\Phi}^\top\mathbf{b}_{\boldsymbol{\mu}_n}-\mathbf{g}_{n,v})^\top\boldsymbol{\theta}]=\mathbb{V}_{|F_{n-1}}[\mathbf{g}_{n,v}^\top\boldsymbol{\theta}]\le\mathbb{E}_{|F_{n-1}}[(\mathbf{g}_{n,v}^\top\boldsymbol{\theta})^2]\le\frac{4C_v^2}{(1-\gamma)^2}$$ □

From the above two lemmas, we see the main difference from the what we had in Appendix D.1 for the tabular case is that, the martingale difference now scales in $O\left(\frac{C_v+dim(\boldsymbol{\theta}_\mu)}{1-\gamma}\right)$ instead of $O\left(\frac{|\mathcal{S}||\mathcal{A}|}{1-\gamma}\right)$, and its variance scales in $O\left(\frac{C_v^2+dim(\boldsymbol{\theta}_\mu)}{(1-\gamma)^2}\right)$ instead of $O\left(\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}\right)$. We note the constant C_v is universal, independent of the problem size. Following the similar steps in Appendix D.1, these new results imply that

$$P\left(\sum_{n=1}^N(\nabla h_n(\theta_n)-g_n)^\top\theta_n>\epsilon\right)\le\exp\left(\frac{-\epsilon^2}{2N\sigma^2(1+\frac{b\epsilon}{3N\sigma^2})}\right)$$

with $b = O\left(\frac{C_v + dim(\boldsymbol{\theta_\mu})}{1-\gamma}\right)$ and $O\left(\frac{C_v^2 + dim(\boldsymbol{\theta_\mu})}{(1-\gamma)^2}\right)$. This implies that, with probability at least $1-\delta$, it hold

$$\sum_{n=1}^{N} (\nabla h_n(\theta_n) - g_n)^\top \theta_n = \tilde{O}\left(\frac{\sqrt{N(C_v^2 + dim(\boldsymbol{\theta_\mu}))\log(\frac{1}{\delta})}}{1-\gamma}\right)$$

E.5 Static Regret of Mirror Descent

Again the steps here are very similar to those in Appendix D.2. We concern bounding the static regret.

$$\max_{\theta \in \Theta} \sum_{n=1}^{N} g_n^\top (\theta_n - \theta)$$

From Appendix D.2, we recall this can be achieved by the mirror descent's optimality condition. The below inequality is true, for any $\theta' \in \Theta$:

$$\sum_{n=1}^{N} \langle g_n, \theta_n - \theta' \rangle \leq \frac{1}{\eta} B_R(\theta' || \theta_1) + \sum_{n=1}^{N} \langle g_n, \theta_{n+1} - \theta_n \rangle - \frac{1}{\eta} B_R(\theta_{n+1} || \theta_n)$$

Based on our choice of Bregman divergence given in (36), i.e.

$$B_R(\theta' || \theta) = \tfrac{1}{2} \tfrac{dim(\theta_v)}{C_v^2} \|\boldsymbol{\theta}'_v - \boldsymbol{\theta}_v\|_2^2 + KL(\boldsymbol{\theta}'_\mu || \boldsymbol{\theta}_\mu), \tag{36}$$

we have $\frac{1}{\eta} B_R(\theta' || \theta_1) \leq \frac{\tilde{O}(1)}{\eta}$. For each $\langle g_n, \theta_{n+1} - \theta_n \rangle - \frac{1}{\eta} B_R(\theta_{n+1} || \theta_n)$, we will use again the two basic lemmas we proved in Appendix D.2.

Lemma 10. *For any vector x, y, g and scalar $\eta > 0$, it holds $\langle g, x - y \rangle - \frac{1}{2\eta}\|x - y\|_2^2 \leq \frac{\eta \|g\|_2^2}{2}$.*

Lemma 11. *Suppose $B_R(x||y) = KL(x||y)$ and x, y are probability distributions, and $g \geq 0$ element-wise. Then, for $\eta > 0$,*

$$-\frac{1}{\eta} B_R(y||x) + \langle g, x - y \rangle \leq \frac{\eta}{2} \sum_i x_i (g_i)^2 = \frac{\eta}{2} \|g\|_x^2.$$

Thus, we have the upper bound

$$\langle g_n, \theta_{n+1} - \theta_n \rangle - \frac{1}{\eta} B_R(\theta_{n+1} || \theta_n) = \frac{C_v^2}{dim(\theta_v)} \frac{\eta \|\mathbf{g}_{n,v}\|_2^2}{2} + \frac{\eta \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\theta}_{\mu,n}}^2}{2}$$

Together with the upper bound on $\frac{1}{\eta} B_R(x' || x_1)$, it implies that

$$\begin{aligned} \sum_{n=1}^{N} \langle g_n, x_n - x' \rangle &\leq \frac{1}{\eta} B_R(x' || x_1) + \sum_{n=1}^{N} \langle g_n, x_{n+1} - x_n \rangle - \frac{1}{\eta} B_R(x_{n+1} || x_n) \\ &\leq \frac{\tilde{O}(1)}{\eta} + \frac{\eta}{2} \sum_{n=1}^{N} \frac{C_v^2}{dim(\theta_v)} \|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\theta}_{\mu,n}}^2 \end{aligned} \tag{37}$$

We can expect, with high probability, $\sum_{n=1}^{N} \frac{C_v^2}{dim(\theta_v)} \|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\theta}_{\mu,n}}^2$ concentrates toward its expectation, i.e.

$$\sum_{n=1}^{N} \frac{C_v^2}{dim(\theta_v)} \|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\theta}_{\mu,n}}^2 \leq \sum_{n=1}^{N} \mathbb{E}\left[\frac{C_v^2}{dim(\theta_v)} \|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|_{\boldsymbol{\theta}_{\mu,n}}^2\right] + o(N)$$

To bound the right-hand side, we will use the upper bounds below, which largely follow the proof of Lemma 16 and Lemma 17.

Lemma 18. $\mathbb{E}[\|\mathbf{g}_{n,v}\|_2^2] \le \frac{4dim(\boldsymbol{\theta}_v)}{(1-\gamma)^2}$ *and* $\mathbb{E}[\|\mathbf{g}_{n,\mu}\|^2_{\boldsymbol{\theta}_{\mu,n}}] \le \frac{4dim(\boldsymbol{\theta}_\mu)}{(1-\gamma)^2}$.

Lemma 19. $\|\mathbf{g}_{n,v}\|_2^2 \le \frac{4dim(\boldsymbol{\theta}_v)}{(1-\gamma)^2}$ *and* $\|\mathbf{g}_{n,\mu}\|^2_{\boldsymbol{\theta}_{\mu,n}} \le \frac{4dim(\boldsymbol{\theta}_\mu)^2}{(1-\gamma)^2}$.

By Azuma-Hoeffding's inequality in Lemma 13,

$$\begin{aligned}\sum_{n=1}^N \frac{C_v^2}{dim(\theta_v)}\|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|^2_{\boldsymbol{\theta}_{\boldsymbol{\mu},n}} &\le \sum_{n=1}^N \mathbb{E}\left[\frac{C_v^2}{dim(\theta_v)}\|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|^2_{\boldsymbol{\theta}_{\boldsymbol{\mu},n}}\right] + O\left(\frac{C_v^2 + dim(\boldsymbol{\theta}_\mu)^2}{(1-\gamma)^2}\sqrt{N\log\left(\frac{1}{\delta}\right)}\right)\\ &\le O\left(\frac{C_v^2 + dim(\boldsymbol{\theta}_\mu)}{(1-\gamma)^2}N\right) + O\left(\frac{C_v^2 + dim(\boldsymbol{\theta}_\mu)^2}{(1-\gamma)^2}\sqrt{N\log\left(\frac{1}{\delta}\right)}\right)\end{aligned}$$

Now we suppose we set $\eta = O\left(\frac{1-\gamma}{\sqrt{N(C_v^2+dim(\boldsymbol{\theta}_\mu))}}\right)$. We have

$$\sum_{n=1}^N \langle g_n, \theta_n - \theta'\rangle \le \frac{\tilde{O}(1)}{\eta} + \frac{\eta}{2}\sum_{n=1}^N \frac{C_v^2}{dim(\theta_v)}\|\mathbf{g}_{n,v}\|_2^2 + \|\mathbf{g}_{n,\mu}\|^2_{\boldsymbol{\theta}_{\mu,n}} \le \tilde{O}\left(\frac{\sqrt{(C_v^2 + dim(\boldsymbol{\theta}_\mu))N}}{1-\gamma}\right)$$

E.5.1 Union Bound

Lastly we use an union bound to handle the term

$$\sum_{n=1}^N (g_n - \nabla h_n(\theta_n))^\top \theta_N^*$$

We follow the steps in Appendix D.3: we will use again the fact that $\theta_N^* = (\boldsymbol{\theta}_{v,N}^*, \boldsymbol{\theta}_\mu^*) \in \Theta$, so we can handle the part with $\boldsymbol{\theta}_\mu^*$ using the standard martingale concentration, and the part with $\boldsymbol{\theta}_{v,N}^*$ using the union bound.

Using the previous analyses, we see can first show that the martingale due to the part $\boldsymbol{\theta}_\mu^*$ concentrates in $\tilde{O}\left(\frac{\sqrt{Ndim(\boldsymbol{\theta}_\mu)\log(\frac{1}{\delta})}}{1-\gamma}\right)$. Likewise, using the union bound, we can show the martingale due to the part $\boldsymbol{\theta}_{v,N}^*$ concentrates in $\tilde{O}\left(\frac{\sqrt{NC_v^2\log(\frac{\mathcal{N}}{\delta})}}{1-\gamma}\right)$ where $\mathcal{N}$ some proper the covering number of the set $\left\{\boldsymbol{\theta}_v : \|\boldsymbol{\theta}_v\|_2 \le \frac{C_v}{\sqrt{dim(\boldsymbol{\theta}_v)}}\right\}$. Because $\log\mathcal{N} = O(\dim(\boldsymbol{\theta}_v))$ for an Euclidean ball. We can combine the two bounds and show together

$$\sum_{n=1}^N (g_n - \nabla h_n(\theta_n))^\top \theta_N^* = \tilde{O}\left(\frac{\sqrt{N(C_v^2\dim(\boldsymbol{\theta}_v) + dim(\boldsymbol{\theta_\mu}))\log(\frac{1}{\delta})}}{1-\gamma}\right)$$

E.5.2 Summary

Combining the results of the three parts above, we have, with probability $1-\delta$,

$$\begin{aligned}&\mathrm{Regret}_N(y^*_{N,\theta})\\ &\le \left(\sum_{n=1}^N (\nabla h_n(\theta_n) - g_n)^\top\theta_n\right) + \left(\max_{\theta\in\Theta}\sum_{n=1}^N g_n^\top(\theta_n - \theta)\right) + \left(\sum_{n=1}^N (g_n - \nabla h_n(\theta_n))^\top\theta_N^*\right)\\ &= \tilde{O}\left(\frac{\sqrt{N(dim(\boldsymbol{\theta_\mu}) + C_v^2)\log(\frac{1}{\delta})}}{1-\gamma}\right) + \tilde{O}\left(\frac{\sqrt{(C_v^2 + dim(\boldsymbol{\theta}_\mu))N}}{1-\gamma}\right) + \tilde{O}\left(\frac{\sqrt{N(C_v^2\dim(\boldsymbol{\theta}_v) + dim(\boldsymbol{\theta_\mu}))\log(\frac{1}{\delta})}}{1-\gamma}\right)\\ &= \tilde{O}\left(\frac{\sqrt{Ndim(\Theta)\log(\frac{1}{\delta})}}{1-\gamma}\right)\end{aligned}$$

where the last step is due to C_v is a universal constant. Or equivalently, the above bounds means a sample complexity in $\tilde{O}\left(\frac{dim(\Theta)\log(\frac{1}{\delta})}{(1-\gamma)^2\epsilon^2}\right)$. Finally, we recall the policy performance has a bias $\epsilon_{\Theta,N}$ in Corollary 1 due to using function approximators. Considering this effect, we have the final statement.