Efficiently Solving MDPs with Stochastic Mirror Descent

Yujia Jin [1] **Aaron Sidford** [1]

Abstract

We present a unified framework based on primal-dual stochastic mirror descent for approximately solving infinite-horizon Markov decision processes (MDPs) given a generative model. When applied to an average-reward MDP with $\mathrm{A}_{\mathrm{tot}}$ total actions and mixing time bound t_{mix} our method computes an ϵ-optimal policy with an expected $\widetilde{O}(t_{\mathrm{mix}}^2 \mathrm{A}_{\mathrm{tot}} \epsilon^{-2})$ samples from the state-transition matrix, removing the ergodicity dependence of prior art. When applied to a γ-discounted MDP with $\mathrm{A}_{\mathrm{tot}}$ total actions our method computes an ϵ-optimal policy with an expected $\widetilde{O}((1-\gamma)^{-4} \mathrm{A}_{\mathrm{tot}} \epsilon^{-2})$ samples, improving over previous primal-dual methods and matching the state-of-the-art up to a $(1-\gamma)^{-1}$ factor. Both methods are model-free, update state values and policies simultaneously, and run in time linear in the number of samples taken. We achieve these results through a more general stochastic mirror descent framework for solving bilinear saddle-point problems with simplex and box domains and we demonstrate the flexibility of this framework by providing further applications to constrained MDPs.

1 Introduction

Markov decision processes (MDPs) are a fundamental mathematical abstraction for sequential decision making under uncertainty and they serve as a basic modeling tool in reinforcement learning (RL) and stochastic control (Bertsekas & Tsitsiklis, 1995; Puterman, 2014; Sutton & Barto, 2018). Two prominent classes of MDPs are discounted MDPs (DMDPs) and average-reward MDPs (AMDPs). Each have been studied extensively; DMDPs have a number of nice theoretical properties including reward convergence and operator monotonicity (Bertsekas et al., 1995) and AMDPs are applicable to optimal control, learning automata, and various real-world reinforcement learning settings (Mahadevan, 1996; Auer & Ortner, 2007; Ouyang et al., 2017).

In this paper we consider the prevalent computational learning problem of finding an approximately optimal policy of an MDP given only restricted access to the model. In particular, we consider the problem of computing an ϵ-optimal policy, i.e. a policy with an additive ϵ error in expected cumulative reward over infinite horizon, under the standard assumption of a generative model (Kearns & Singh, 1999; Kakade et al., 2003), which allows one to sample from state-transitions given the current state-action pair. This problem is well-studied and there are multiple known upper and lower bounds on its sample complexity (Azar et al., 2012; Wang, 2017a; Sidford et al., 2018a; Wainwright, 2019).

In this work, we provide a unified framework based on primal-dual stochastic mirror descent (SMD) for learning an ϵ-optimal policies for both AMDPs and DMDPs with a generative model. We show that this framework achieves sublinear running times for solving dense bilinear saddle-point problems with simplex and box domains, and (as a special case) ℓ_∞ regression (Sherman, 2017; Sidford & Tian, 2018). As far as we are aware, this is the first such sub-linear running time for this problem. We achieve our results by applying this framework to saddle-point representations of AMDPs and DMDPs and proving that approximate equilibria yield approximately optimal policies.

Our MDP algorithms have sample complexity linear in the total number of possible actions, denoted by $\mathrm{A}_{\mathrm{tot}}$. For an AMDP with bounded mixing time t_{mix} for all policies, we prove a sample complexity of $\widetilde{O}(t_{\mathrm{mix}}^2 \mathrm{A}_{\mathrm{tot}} \epsilon^{-2})$ [1], which removes the ergodicity condition of prior art (Wang, 2017b) (which can in the worst-case be unbounded). For DMDP with discount factor γ, we prove a sample complexity of $\widetilde{O}((1-\gamma)^{-4} \mathrm{A}_{\mathrm{tot}} \epsilon^{-2})$, improving over the best-known achievable sample complexity by primal-dual methods (Wang, 2017a) and matching the state-of-the-art (Sidford et al., 2018a; Wainwright, 2019) and lower bound (Azar et al., 2012) up to a $(1-\gamma)^{-1}$ factor.

[1] Management Science and Engineering (MS&E), Stanford University, Stanford, United States. Correspondence to: Yujia Jin <yujiajin@stanford.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1] Throughout the paper we use $\widetilde{O}$ to hide poly-logarithmic factors in $\mathrm{A}_{\mathrm{tot}}$, t_{mix}, $1/(1-\gamma)$, $1/\epsilon$, and the number of states of the MDP.

We hope our method serves as a building block towards a more unified understanding the complexity of MDPs and RL. By providing a general SMD-based framework which is provably efficient for solving multiple prominent classes of MDPs we hope this paper may lead to a better understanding and broader application of the traditional convex optimization toolkit to modern RL. As a preliminary demonstration of flexibility of our framework, we show that it extends to yield new results for approximately optimizing constrained MDPs and hope it may find further utility.

1.1 Problem Setup

Throughout the paper we denote an MDP instance by a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{r}, \gamma)$ with components defined as follows:

- $\mathcal{S}$ - a finite set of states where each $i \in \mathcal{S}$ is called a *state of the MDP*, in tradition this is also denoted as s.
- $\mathcal{A} = \cup_{i\in[S]}\mathcal{A}_i$ - a finite set of actions that is a collection of sets of actions $\mathcal{A}_i$ for states $i \in \mathcal{S}$. We overload notation slightly and let $(i, a_i) \in \mathcal{A}$ denote *an action a_i at state i*. $\mathrm{A}_{\mathrm{tot}} := |\mathcal{A}| := \sum_{i\in\mathcal{S}} |\mathcal{A}_i|$ denotes the total number of state-action pairs.
- $\mathcal{P}$ - the collection of state-to-state transition probabilities where $\mathcal{P} := \{p_{ij}(a_i)|i, j \in \mathcal{S}, a_i \in \mathcal{A}_i\}$ and $p_{ij}(a_i)$ denotes the probability of transition to state j when taking action a_i at state i.
- $\mathbf{r}$ - the vector of state-action transitional rewards where $\mathbf{r} \in [0, 1]^{\mathcal{A}}$, r_{i,a_i} is the instant reward received when taking action a_i at state $i \in \mathcal{S}$.[2]
- γ - the discount factor of MDP, by which one downweights the reward in the next future step. When $\gamma \in (0, 1)$, we call the instance a *discounted MDP* (DMDP) and when $\gamma = 1$, we call the instance an *average-reward MDP* (AMDP).

We use $\mathbf{P} \in \mathbb{R}^{\mathcal{A}\times\mathcal{S}}$ as the state-transition matrix where its (i, a_i)-th row corresponds to the transition probability from state $i \in \mathcal{S}$ where $a_i \in \mathcal{A}_i$ to state j. Correspondingly we use $\hat{\mathbf{I}}$ as the matrix with a_i-th row corresponding to $\mathbf{e}_i$, for all $i \in \mathcal{S}, a_i \in \mathcal{A}_i$.

Now, the model operates as follows: when at state i, one can pick an action a_i from the given action set $\mathcal{A}_i$. This generates a reward r_{i,a_i}. Also based on the transition model with probability $p_{ij}(a_i)$, it transits to state j and the process repeats. Our goal is to compute a random policy which determines which actions to take at each state. A random policy is a collection of probability distributions $\pi := \{\pi_i\}_{i\in\mathcal{S}}$, where $\pi_i \in \Delta^{\mathcal{A}_i}$ is a vector in the $|\mathcal{A}_i|$-dimensional simplex with $\pi_i(a_i)$ denoting the probability of taking $a_i \in \mathcal{A}_i$ at action j. One can extend π_i to the set of $\Delta^{\mathcal{A}}$ by filling in 0s on entries corresponding to other states $j \neq i$, and denote $\Pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ as the concatenated policy matrix with i-th row being the extended Δ_i. We denote $\mathbf{P}^\pi$ as the trasitional probability matrix of the MDP when using policy π, thus we have $\mathbf{P}^\pi(i, j) := \sum_{a_i\in\mathcal{A}_i} \pi_i(a_i)p_{ij}(a_i) = \Pi \cdot \mathbf{P}$ for all $i, j \in \mathcal{S}$, where $\cdot$ in the right-hand side (RHS) denotes matrix-matrix multiplication. Further, we let $\mathbf{r}^\pi$ denote corresponding average reward under policy π defined as $\mathbf{r}^\pi := \Pi \cdot \mathbf{r}$, where $\cdot$ in RHS denotes matrix-vector multiplication. We overload notation and still use $\mathbf{I}$ to denote the standard identity matrix if computing with regards to probability transition matrix $\Pi^\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$.

Given an MDP instance $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{r}, \gamma)$ and an initial distribution over states $\mathbf{q} \in \Delta^{\mathcal{S}}$, we are interested in finding the optimal π^* among all policy π that maximizes the following cumulative reward $\bar{v}^\pi$ of the MDP:

$$\pi^* := \arg\max_\pi \bar{v}^\pi \quad \text{where}$$
$$\bar{v}^\pi := \begin{cases} \mathbb{E}^\pi\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_{i_t,a_t} | i_1 \sim \mathbf{q}\right], & \text{i.e., DMDPs} \\ \lim_{T\to\infty} \frac{1}{T}\mathbb{E}^\pi\left[\sum_{t=1}^{T} r_{i_t,a_t} | i_1 \sim \mathbf{q}\right], & \text{i.e., AMDPs} . \end{cases}$$

Here $\{i_1, a_1, i_2, a_2, \cdots, i_t, a_t\}$ are state-action transitions generated by the MDP under policy π. For the DMDP case, it also holds by definition that $\bar{v}^\pi := \mathbf{q}^\top(\mathbf{I} - \gamma\mathbf{P}^\pi)^{-1}\mathbf{r}^\pi$.

For the AMDP case (i.e. when $\gamma = 1$), we also define $\boldsymbol{\nu}^\pi$ as the stationary distribution under policy π satisfying $\boldsymbol{\nu}^\pi = (\mathbf{P}^\pi)^\top\boldsymbol{\nu}^\pi$. To ensure the value of $\bar{v}^\pi$ is well-defined, we restrict our attention to a subgroup which we call *mixing AMDP* satisfying the following mixing assumption:

Assumption A. *An AMDP instance is* mixing *if t_{mix}, defined as follows, is bounded by $1/2$, i.e.*

$$t_{\mathrm{mix}} := \max_\pi \left[\arg\min_{t\geq 1} \max_{\mathbf{q}\in\Delta^{\mathcal{S}}} \|(\mathbf{P}^{\pi\top})^t\mathbf{q} - \boldsymbol{\nu}^\pi\|_1\right] \leq \tfrac{1}{2}.$$

The mixing condition assumes for arbitrary policy π and arbitrary initial state, the resulting Markov chain leads toward a distribution close enough to its stationary distribution $\boldsymbol{\nu}^\pi$ starting from any initial state i in $O(t_{\mathrm{mix}})$ time steps. This assumption implies the the uniqueness of the stationary distribution, makes $\bar{v}^\pi$ above well-defined with the equivalent $\bar{v}^\pi = (\boldsymbol{\nu}^\pi)^\top\mathbf{r}^\pi$, governing the complexity of our mixing AMDP algorithm, and is key for the results we prove (Theorem 1). Further, the assumption is equivalent as in Wang (2017b), up to constant factors.

By nature of the definition of *mixing AMDP*, we note that the value of a strategy π is independent of initial distribution $\mathbf{q}$ and only dependent of the eventual stationary distribution

[2] The assumption that $\mathbf{r}$ only depends on state action pair i, a_i is a common practice (Sidford et al., 2018a).

as long as the AMDP is mixing, which also implies $\bar{v}^\pi$ is always well-defined. For this reason, sometimes we also omit $i_1 \sim \mathbf{q}$ in the corresponding definition of $\bar{v}^\pi$.

We call a policy π an ϵ-*(approximate) optimal policy* for the MDP problem, if it satisfies $\bar{v}^\pi \geq \bar{v}^* - \epsilon$.[3] We call a policy an expected ϵ-(approximate) optimal policy if it satisfies the condition in expectation, i.e. $\mathbb{E}\bar{v}^\pi \geq \bar{v}^* - \epsilon$. The goal of paper is to develop efficient algorithms that find (expected) ϵ-approximate policy for the given MDP instance assuming access to a generative model.

1.2 Main Results

The main result of the paper is a unified framework based on randomized primal-dual stochastic mirror descent (SMD) that with high probability finds an (expected) ϵ-optimal policy with some sample complexity guarantee. Formally we provide two algorithms (see Algorithm 1 for both cases) with the following guarantees respectively.

Theorem 1. *Given a mixing AMDP tuple* $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{r})$*, let* $\epsilon \in (0, 1)$*, one can construct an expected* ϵ*-optimal policy* π^ϵ *from the decomposition (see Section 5) of output* $\boldsymbol{\mu}^\epsilon$ *of Algorithm 1 with sample complexity* $O\left(t_{\mathrm{mix}}^2 \mathrm{A}_{\mathrm{tot}} \epsilon^{-2} \log(\mathrm{A}_{\mathrm{tot}})\right)$.

Theorem 2. *Given a DMDP tuple* $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{r}, \gamma)$ *with discount factor* $\gamma \in (0, 1)$*, let* $\epsilon \in (0, 1)$*, one can construct an expected* ϵ*-optimal policy* π^ϵ *from the decomposition (see Section 5) of output* $\boldsymbol{\mu}^\epsilon$ *of Algorithm 1 with sample complexity* $O\left((1-\gamma)^{-4} \mathrm{A}_{\mathrm{tot}} \epsilon^{-2} \log(\mathrm{A}_{\mathrm{tot}})\right)$.

We remark that for both problems, the algorithm also gives with high probability an ϵ-optimal policy at the cost of an extra $\log(1/\delta)$ factor to the sample complexity through a reduction from high-probability to expected optimal policy (see Wang (2017b) for more details). Note that we only get randomized policies, and we leave the question of getting directly deterministic policies ass an interesting open direction.

Table 1 gives a comparison of sample complexity between our methods and prior methods[4] for computing an ϵ-approximate policy in DMDPs and AMDPs given a generative model.

As a generalization, we show how to solve constrained average-reward MDPs (cf. (Altman, 1999), a generalization of average-reward MDP) using the primal-dual stochastic mirror descent framework in Section 6. We build an algorithm that solves the constrained problem (13) to ϵ-accuracy within sample complexity $O((t_{\mathrm{mix}}^2 + K) D^2 \mathrm{A}_{\mathrm{tot}} \epsilon^{-2} \log(\mathrm{A}_{\mathrm{tot}}))$, where K and D^2 are number and size of the constraints. To the best of our knowledge this is the first sample complexity results for constrained MDPs given by a generative model.

[3]Hereinafter, we use superscript * and $^{\pi^*}$ interchangeably.

[4]Most methods assume a uniform action set $\mathcal{A}$ for each of the $|\mathcal{S}|$ states, but can also be generalizedd to the non-uniform case parameterized by $\mathrm{A}_{\mathrm{tot}}$.

Algorithm 1 SMD for mixing AMDP / DMDPs

Input: MDP tuple $\mathcal{M} = (\mathcal{S}, \cup_{i \in \mathcal{S}} \mathcal{A}_i, \mathcal{P}, \mathbf{r}, \gamma)$, initial $(\mathbf{v}_0, \boldsymbol{\mu}_0) \in \mathbb{B}_{2M}^{\mathcal{S}} \times \Delta^{\mathcal{A}}$.
Output: An expected ϵ-approximate solution $(\mathbf{v}^\epsilon, \boldsymbol{\mu}^\epsilon)$ for problem (5).
Parameter: Step-size η^{v}, η^{μ}, number of iterations T, accuracy level ϵ.
for $t = 1, \ldots, T$ **do**
  // $\mathbf{v}$ *gradient estimation*
  Sample $(i, a_i) \sim [\boldsymbol{\mu}]_{i,a_i}$, $j \sim p_{ij}(a_i)$, $i' \sim q_{i'}$
  Set $\tilde{g}_{t-1}^{\mathsf{v}} = \begin{cases} \mathbf{e}_j - \mathbf{e}_i & \text{mixing} \\ (1-\gamma)\mathbf{e}_{i'} + \gamma \mathbf{e}_j - \mathbf{e}_i & \text{discounted} \end{cases}$
  // $\boldsymbol{\mu}$ *gradient estimation*
  Sample $(i, a_i) \sim \frac{1}{\mathrm{A}_{\mathrm{tot}}}$, $j \sim p_{ij}(a_i)$
  Set $\tilde{g}_{t-1}^{\mu} = \begin{cases} \mathrm{A}_{\mathrm{tot}}(v_i - v_j - r_{i,a_i})\mathbf{e}_{i,a_i} & \text{mixing} \\ \mathrm{A}_{\mathrm{tot}}(v_i - \gamma v_j - r_{i,a_i})\mathbf{e}_{i,a_i} & \text{discounted} \end{cases}$
  // *Stochastic mirror descent steps* (Π *as projection)*
  $\mathbf{v}_t \leftarrow \Pi_{\mathbb{B}_{2M}^{\mathcal{S}}}(\mathbf{v}_{t-1} - \eta^{\mathsf{v}} \tilde{g}_{t-1}^{\mathsf{v}})$
  $\boldsymbol{\mu}_t \leftarrow \Pi_{\Delta^{\mathcal{A}}}(\boldsymbol{\mu}_{t-1} \circ \exp(-\eta^{\mu} \tilde{g}_{t-1}^{\mu}))$
end for
Return $(\mathbf{v}^\epsilon, \boldsymbol{\mu}^\epsilon) \leftarrow \frac{1}{T} \sum_{t \in [T]} (\mathbf{v}_t, \boldsymbol{\mu}_t)$

As a byproduct, the framework we build in Section 3 also gives a stochastic algorithms (see Algorithm 2) that finds an expected ϵ-approximate solution of ℓ_∞-ℓ_1 bilinear minimax problems of form $\min_{\mathbf{x} \in [-1,1]^n} \max_{\mathbf{y} \in \Delta^m} \mathbf{y}^\top \mathbf{M} \mathbf{x} + \mathbf{b}^\top \mathbf{x} - \mathbf{c}^\top \mathbf{y}$ to ϵ-additive accuracy with runtime $\widetilde{O}(((m + n)\left\|\mathbf{M}\right\|_\infty^2 + n\|\mathbf{b}\|_1^2 + m\left\|\mathbf{c}\right\|_\infty^2)\epsilon^{-2})$ given ℓ_1 sampler of iterate $\mathbf{y}$, and ℓ_1 samplers based on the input entries of $\mathbf{M}$, $\mathbf{b}$ and $\mathbf{c}$ (see Corollary 1 for details). Consequently, it solves (box constrained) ℓ_∞ regression problems of form $\min_{\mathbf{x} \in [-1,1]^n} \left\|\mathbf{M}\mathbf{x} - \mathbf{c}\right\|_\infty$ to ϵ-additive accuracy within runtime $\widetilde{O}(((m+n)\left\|\mathbf{M}\right\|_\infty^2 + m\left\|\mathbf{c}\right\|_\infty^2)\epsilon^{-2})$ given similar sampling access (see Remark 1 for details and Table 3 for comparison with previous results).

1.3 Technique Overview

We adopt the idea of formulating the MDP problem as a bilinear saddle point problem in light of linear duality, following the line of randomized model-free primal-dual π learning studied in Wang (2017a;b). This formulation relates MDP to solving bilinear saddle point problems with box and simplex domains, which falls into well-studied generalizations of convex optimization (Nemirovski, 2004a; Carmon et al., 2019).

Type	Method	Sample Complexity
mixing AMDP	Primal-Dual Method (Wang, 2017b)	$\widetilde{O}\left(\tau^2 t_{\mathrm{mix}}^2 \mathrm{A_{tot}} \epsilon^{-2}\right)$
	Our method (Theorem 1)	$\widetilde{O}\left(t_{\mathrm{mix}}^2 \mathrm{A_{tot}} \epsilon^{-2}\right)$
DMDP	Empirical QVI (Azar et al., 2012)	$\widetilde{O}\left((1-\gamma)^{-5} \mathrm{A_{tot}} \epsilon^{-2}\right)$
	Primal-Dual Method (Wang, 2017a)	$\widetilde{O}\left((1-\gamma)^{-6} \lvert\mathcal{S}\rvert^2 \mathrm{A_{tot}} \epsilon^{-2}\right)$
	Primal-Dual Method (Wang, 2017a)	$\widetilde{O}\left(\tau^4 (1-\gamma)^{-4} \mathrm{A_{tot}} \epsilon^{-2}\right)$
	Variance-reduced QVI (Sidford et al., 2018a)	$\widetilde{O}\left((1-\gamma)^{-3} \mathrm{A_{tot}} \epsilon^{-2}\right)$
	Empirical MDP + Blackbox (Agarwal et al., 2020)	$\widetilde{O}\left((1-\gamma)^{-3} \mathrm{A_{tot}} \epsilon^{-2}\right)$
	Variance-reduced Q-learning (Wainwright, 2019)	$\widetilde{O}\left((1-\gamma)^{-3} \mathrm{A_{tot}} \epsilon^{-2}\right)$
	Our method (Theorem 2)	$\widetilde{O}\left((1-\gamma)^{-4} \mathrm{A_{tot}} \epsilon^{-2}\right)$

Table 1. **Comparison of sample complexity to get ϵ-optimal policy among stochastic methods (complete version see Appendix A).** Here $\mathcal{S}$ denotes state space, $\mathrm{A_{tot}}$ denotes number of state-action pair, t_{mix} is mixing time for mixing AMDP, and γ is discount factor for DMDP. Parameter τ shows up whenever the designed algorithm requires additional ergodic condition for MDP, i.e. there exists some distribution $\mathbf{q}$ and $\tau > 0$ satisfying $\sqrt{1/\tau}\mathbf{q} \le \boldsymbol{\nu}^\pi \le \sqrt{\tau}\mathbf{q}$, $\forall$ policy π and its induced stationary distribution $\boldsymbol{\nu}^\pi$.

We study the efficiency of standard stochastic mirror descent (SMD) for this bilinear saddle point problem where the minimization (primal) variables are constrained to in a rescaled box domain and the maximization (dual) variables are constrained to lie in the simplex. We use the idea of local-norm variance bounds emerging in Shalev-Shwartz et al. (2012); Carmon et al. (2019), to design and analyze efficient stochastic estimators for the gradient of this problem that have low-variance under the corresponding local norms. We provide a new analytical way to bound the quality of an approximately-optimal policy constructed from the approximately optimal solution of bilinear saddle point problem, which utilizes the influence of the dual constraints under minimax optimality. Compared with prior work, we eliminate the assumption of ergodicity through extending the primal space by a constant size. Putting all these together, we prove a simple and natural SMD algorithm which solves both mixing AMDP and DMDP problems efficiently, with clear dependence on hardness parameters like mixing time t_{mix} or discount factor γ.

1.4 Related Work

1.4.1 On Solving MDP

Within the tremendous body of study on MDPs, and more generally reinforcement learning, stands the well-studied classic problem of computational efficiency (i.e. iteration number, runtime, etc.) of finding optimal policy, given the entire MDP instance as an input. Traditional deterministic methods for the problems are value iteration, policy iteration, and linear programming. (Bertsekas et al., 1995; Ye, 2011), which find an approximately optimal policy to high-accuracy but have superlinear runtime in the usually high problem dimension $\Omega(\lvert\mathcal{S}\rvert \cdot \mathrm{A_{tot}})$.

To avoid the necessity of knowing the whole problem instance and having superlinear runtime dependence, more recently, researchers have designed stochastic algorithms assuming only a generative model that samples from state-transitions (Kakade et al., 2003). Azar et al. (2012) proved a lower bound of $\Omega((1-\gamma)^{-3}\mathrm{A_{tot}}\epsilon^{-2})$ while also giving a Q-value-iteration algorithm with a higher guaranteed sample complexity. This was recently improved in Sidford et al. (2018b) using variance-reduction ideas, and was further improved to match(up to logarithmic factors) lower bound in (Sidford et al., 2018a) using a type of variance-reduced randomized value iteration. Soon later in Wainwright (2019), a variance-reduced Q-learning method also achieves nearly tight sample complexity for the discounted case. In Agarwal et al. (2020) the authors use a non-algorithmic approach that shows $\widetilde{O}((1-\gamma)^{-3}\mathrm{A_{tot}}\epsilon^{-2})$ samples suffice. On the other hand, Wang (2017a) designed a randomized primal-dual method, an instance of SMD with slightly different sampling distribution and form of update for the estimators, which has superlinear sample complexity guarantee unless additional ergodicity assumptions are made. Whether such randomized primal-dual methods necessarily incur this higher computational cost is unclear and a key motivation for our work.

While a few methods match (up to logarithmic factors) the lower bound shown for sample complexity for solving DMDP (Sidford et al., 2018a; Wainwright, 2019), it is unclear how and if one can design a similar method for average-reward MDP and get the optimal sample complexity dependence. The only related work under average-reward MDP setting uses primal-dual π-learning (Wang, 2017b), following the stochastic primal-dual method in (Wang, 2017a). Being also a variant of SMD methods, their algorithm has a different domain setup, different update forms, and a more ad-hoc analysis compared with ours. The sample complexity of their method is $\widetilde{O}(\tau^2 t_{\text{mix}}^2 \mathrm{A}_{\text{tot}}\epsilon^{-2})$, which still requires an additional bound on the ergodicity parameter, $\tau > 0$, and depends on this parameter polynomially.

This mismatch of the theoretical efficiency of SMD methods as opposed to value iteration and Q-learning in solving DMDPs and the dependence of ergodicity when solving AMDPs with SMD methods motivated our study of a general SMD framework that could provably efficiently solve both problem instances. Table 1 includes a complete comparison between our results and the prior art for both cases.

1.4.2 On ℓ_∞ Regression and Bilinear Saddle Point Problesm

Our framework gives a stochastic method for solving ℓ_∞ regression, which is a core problem in both combinatorics and continuous optimization due to its connection with maximum flow and linear programming (Lee & Sidford, 2014; 2015). Classic methods build on solving a smooth approximations of the problem (Nesterov, 2005) or finding the right regularizers and algorithms for its correspondingly primal-dual minimax problem (Nemirovski, 2004b; Nesterov, 2007). These methods have recently been improved to $\widetilde{O}(\mathrm{nnz}\,\|\mathbf{M}\|_\infty\,\epsilon^{-1})$ using a joint regularizer with nice area-convexity properties in Sherman (2017) or using accelerated coordinate method with a matching runtime bound in sparse-column case in Sidford & Tian (2018) .

In comparison to all the state-of-the-art, for dense input data matrix our method gives the first algorithm with sublinear runtime dependence $O(m+n)$ instead of $O(\mathrm{nnz})$. For completeness we include a comparison of runtimes for methods mentioned above in Appendix B (see Table 3).

Our sublinear method for ℓ_∞-regression is closely related to a line of work on obtaining efficient stochastic methods for approximately solving *matrix games*, i.e. bilinear saddle point problems (Grigoriadis & Khachiyan, 1995; Clarkson et al., 2012; Palaniappan & Bach, 2016), and, in particular, a recent line of work by the authors and collaborators (Carmon et al., 2019; 2020) that explores the benefit of careful sampling and variance reduction in matrix games. In Carmon et al. (2019) we provide a framework to analyze variance-reduced SMD under local norms to obtain better complexity bounds for different domain setups, i.e. ℓ_1-ℓ_1, ℓ_1-ℓ_2, and ℓ_2-ℓ_2 where ℓ_1 corresponds to the simplex and ℓ_2 corresponds to the Euclidean ball. In Carmon et al. (2020) we study the improved sublinear and variance-reduced coordinate methods for these domain setups utilizing the desgn of optimal gradient estimators. This paper adapts the local norm analysis and coordinate-wise gradient estimator design in Carmon et al. (2019; 2020) to obtain our SMD algorithm and analysis for ℓ_1-ℓ_∞ games.

2 Preliminaries

First, we introduce several known tools for studying MDPs.

2.1 Bellman Equation.

For mixing AMDP, $\bar{v}^*$ is the optimal average reward if and only if there exists a vector $\mathbf{v}^* = (v_i^*)_{i\in\mathcal{S}}$ satisfying its corresponding *Bellman equation* (Bertsekas et al., 1995)

$$\bar{v}^* + v_i^* = \max_{a_i\in\mathcal{A}_i}\left\{\sum_{j\in\mathcal{S}} p_{ij}(a_i)v_j^* + r_{i,a_i}\right\}, \forall i\in\mathcal{S}. \quad (1)$$

When considering a mixing AMDP as in the paper, the existence of solution to the above equation can be guaranteed. However, it is important to note that one cannot guarantee the uniqueness of the optimal $\mathbf{v}^*$. In fact, for each optimal solution $\mathbf{v}^*$, $\mathbf{v}^* + c\mathbf{1}$ is also an optimal solution.

For DMDP, one can show that at optimal policy π^*, each state $i\in\mathcal{S}$ can be assigned an optimal cost-to-go value v_i^* satisfying the following *Bellman equation* (Bertsekas et al., 1995)

$$v_i^* = \max_{a_i\in\mathcal{A}_i}\left\{\sum_{j\in\mathcal{S}} \gamma p_{ij}(a_i)v_j^* + r_{i,a_i}\right\}, \forall i\in\mathcal{S}. \quad (2)$$

When $\gamma\in(0,1)$, it is straightforward to guarantee the existence and uniqueness of the optimal solution $\mathbf{v}^* := (v_i^*)_{i\in\mathcal{S}}$ to the system.

2.2 Linear Programming (LP) Formulation.

We can further write the above Bellman equations equivalently as the following primal or dual linear programming problems. We define the domain as $\mathbb{B}_m^{\mathcal{S}} := m\cdot[-1,1]^{S}$ where $\mathbb{B}$ stands for box, and $\Delta^n := \{\Delta\in\mathbb{R}^n, \Delta_i\geq 0, \sum_{i\in[n]}\Delta_i = 1\}$ for standard n-dimension simplex.

For mixing AMDP case, the linear programming formulation leveraging matrix notation is (with (P), (D) representing (equivalently) the primal form and the dual form respectively)

$$\begin{aligned}
&(\mathrm{P}) && \min_{\bar{v},\mathbf{v}} && \bar{v}\\
& && \text{subject to} && \bar{v}\cdot\mathbf{1} + (\hat{\mathbf{I}}-\mathbf{P})\mathbf{v} - \mathbf{r}\geq 0,\\
&(\mathrm{D}) && \max_{\boldsymbol{\mu}\in\Delta^{\mathcal{A}}} && \boldsymbol{\mu}^\top\mathbf{r}\\
& && \text{subject to} && (\hat{\mathbf{I}}-\mathbf{P})^\top\boldsymbol{\mu} = \mathbf{0}.
\end{aligned} \quad (3)$$

The optimal values of both systems are the optimal expected cumulative reward $\bar{v}^*$ under optimal policy π^*, thus hereinafter we use $\bar{v}^*$ and $\bar{v}^{\pi^*}$ interchangeably. Given the optimal dual solution $\boldsymbol{\mu}^*$, one can without loss of generality impose the constraint of $\langle \mathbf{I}^\top \boldsymbol{\mu}^*, \mathbf{v}^* \rangle = 0$ [5] to ensure uniqueness of the primal problem (P).

For DMDP case, the equivalent linear programming is

$$\begin{aligned} &(\mathrm{P}) \quad \min_{\mathbf{v}\in\mathbb{B}^{\mathcal{S}}_{2M}} \quad (1-\gamma)\mathbf{q}^\top\mathbf{v} \\ &\quad\quad \text{subject to} \quad (\hat{\mathbf{I}}-\gamma\mathbf{P})\mathbf{v}-\mathbf{r}\ge 0, \\ &(\mathrm{D}) \quad \max_{\boldsymbol{\mu}\in\Delta^{\mathcal{A}}} \quad \boldsymbol{\mu}^\top\mathbf{r} \\ &\quad\quad \text{subject to} \quad (\hat{\mathbf{I}}-\gamma\mathbf{P})^\top\boldsymbol{\mu}=(1-\gamma)\mathbf{q}. \end{aligned} \tag{4}$$

Given a fixed initial distribution $\mathbf{q}$, the optimal values of both systems are a $(1-\gamma)$ factor of the optimal expected cumulative reward , i.e. $(1-\gamma)\bar{v}^*$ under optimal policy π^*.

2.3 Minimax Formulation.

By standard linear duality, we can recast the problem formulation in Section 2.2 using the method of Lagrangian multipliers, as bilinear saddle-point (minimax) problems. For AMDPs the minimax formulation is

$$\begin{aligned} &\min_{\bar{v},\mathbf{v}\in\mathbb{B}^{\mathcal{S}}_{2M}} \max_{\boldsymbol{\mu}\in\Delta^{\mathcal{A}}} \quad f(\bar{v},\mathbf{v},\boldsymbol{\mu}), \\ &\text{where } f(\bar{v},\mathbf{v},\boldsymbol{\mu}) \quad := \bar{v}+\boldsymbol{\mu}^\top(-\bar{v}\cdot\mathbf{1}+(\mathbf{P}-\hat{\mathbf{I}})\mathbf{v}+\mathbf{r}) \\ &\qquad\qquad\qquad\quad = \boldsymbol{\mu}^\top((\mathbf{P}-\hat{\mathbf{I}})\mathbf{v}+\mathbf{r}) \end{aligned} \tag{5}$$

For DMDPs the minimax formulation is

$$\begin{aligned} &\min_{\mathbf{v}\in\mathbb{B}^{\mathcal{S}}_{2M}} \max_{\boldsymbol{\mu}\in\Delta^{\mathcal{A}}} f_{\mathbf{q}}(\mathbf{v},\boldsymbol{\mu}), \\ &\text{where } f_{\mathbf{q}}(\mathbf{v},\boldsymbol{\mu}) := (1-\gamma)\mathbf{q}^\top\mathbf{v}+\boldsymbol{\mu}^\top((\gamma\mathbf{P}-\hat{\mathbf{I}})\mathbf{v}+\mathbf{r}). \end{aligned} \tag{6}$$

Note in both cases we have added the constriant of $\mathbf{v}\in\mathbb{B}^{\mathcal{S}}_{2M}$. The M is different for each case, and will be specified in Section 4 to ensure that $\mathbf{v}^*\in\mathbb{B}^{\mathcal{S}}_{M}$. As a result, constraining the bilinear saddle point problem on a restriced domain for primal variables will not affect the optimality of the original optimal solution due to it global optimality, but will considerably save work for the algorithm by considering a smaller domain.

For each problem we define the duality gap of the minimax problem $\min_{\mathbf{v}\in\mathbb{B}^{\mathcal{S}}_{2M}}\max_{\boldsymbol{\mu}\in\Delta^{\mathcal{A}}} f(\mathbf{v},\boldsymbol{\mu})$ at a given pair of feasible solution $(\mathbf{v},\boldsymbol{\mu})$ as $\mathrm{Gap}(\mathbf{v},\boldsymbol{\mu}) := \max_{\boldsymbol{\mu}'\in\Delta^{\mathcal{A}}} f(\mathbf{v},\boldsymbol{\mu}') - \min_{\mathbf{v}'\in\mathbb{B}^{\mathcal{S}}_{2M}} f(\mathbf{v}',\boldsymbol{\mu})$.

An ϵ-approximate solution of the minimax problem is a pair of feasible solution $(\mathbf{v}^\epsilon,\boldsymbol{\mu}^\epsilon)\in\mathbb{B}^{\mathcal{S}}_{2M}\times\Delta^{\mathcal{A}}$ with its duality gap bounded by ϵ, i.e. $\mathrm{Gap}(\mathbf{v}^\epsilon,\boldsymbol{\mu}^\epsilon)\le\epsilon$. An expected ϵ-approximate solution is one satisfying $\mathbb{E}\mathrm{Gap}(\mathbf{v}^\epsilon,\boldsymbol{\mu}^\epsilon)\le\epsilon$.

[5] $\hat{\mathbf{I}}^\top\boldsymbol{\mu}^*$ represents the stationary distribution over states given optimal policy π^* constructed from optimal dual variable $\boldsymbol{\mu}^*$.

3 Stochastic Mirror Descent Framework

In this section, we consider the following ℓ_∞-ℓ_1 bilinear games as an abstraction of the MDP minimax problems of interest. Such games are induced by one player minimizing over the box domain (ℓ_∞) and the other maximizing over the simplex domain (ℓ_1) a bilinear objective:

$$\min_{\mathbf{x}\in\mathbb{B}^n_b}\max_{y\in\Delta^m} f(\mathbf{x},\mathbf{y}) := \mathbf{y}^\top\mathbf{M}\mathbf{x}+\mathbf{b}^\top\mathbf{x}-\mathbf{c}^\top\mathbf{y}, \tag{7}$$

We study the efficiency of coordinate stochastic mirror descent algorithms onto this ℓ_∞-ℓ_1 minimax problem. The analysis follows from extending a fine-grained analysis of mirror descent with Bregman divergence using local norm arguments in Shalev-Shwartz et al. (2012); Carmon et al. (2019) to the ℓ_∞-ℓ_1 domain. We defer all proofs in this section to Appendix C.

At a given iterate $(\mathbf{x},\mathbf{y})\in B^n_b\times\Delta^m$, our algorithm computes an estimate of the gradients for both sides defined as $g^{\mathsf{x}}(\mathbf{x},\mathbf{y}) := \mathbf{M}^\top\mathbf{y}+\mathbf{b}\in\mathbb{R}^n$ (x-gradient, or g^{x}); $g^{\mathsf{y}}(\mathbf{x},\mathbf{y}) := -\mathbf{M}\mathbf{x}+\mathbf{c}\in\mathbb{R}^m$ (y-gradient, or g^{y}).

The norm we use to measure these gradients are induced by Bregman divergence, a natural extension of Euclidean norm. For our analysis we choose to use $V_{\mathbf{x}}(\mathbf{x}') := \frac{1}{2}\|\mathbf{x}-\mathbf{x}'\|_2^2$ and $V_{\mathbf{y}}(\mathbf{y}') := \sum_{i\in[m]} y_i\log(\frac{y'_i}{y_i})$ (KL-divergence), which are also common practice (Wang, 2017a;b; Nesterov, 2005) catering to the geometry of each domain, and induce the dual norms on the gradients in form $\|g^{\mathsf{x}}\| := \|g^{\mathsf{x}}\|_2 = \sqrt{\sum_{j\in[n]} g_j^{\mathsf{x}2}}$ (standard ℓ_2-norm)for x side, and $\|g^{\mathsf{y}}\|^2_{\mathbf{y}'} := \sum_{i\in[m]} y'_i(g_i^{\mathsf{y}})^2$ (a weighted ℓ_2-norm) for y side.

To describe the properties of estimators needed for our algorithm, we introduce the following definition of *bounded estimator* as follows.

Definition 1 (Bounded Estimator)**.** *Given the following properties on mean, scale and variance of an estimator:*
(i) unbiasedness: $\mathbb{E}\tilde{g}=g$*;*
(ii) bounded maximum entry: $\|\tilde{g}\|_\infty\le c$ *with probability* 1*;*
(iii) bounded second-moment: $\mathbb{E}\|\tilde{g}\|^2\le v$
we call $\tilde{g}$ *a* $(c,v,\|\cdot\|)$-bounded estimator *if satisfying* (i) *and* (iii)*, call it and a* $(c,v,\|\cdot\|^m_\Delta)$-bounded estimator *if besides* (i) *and* (ii)*, it also satisfies* (iii) *with local norm* $\|\cdot\|_{\mathbf{y}}$ *for all* $\mathbf{y}\in\Delta^m$*.*

Now we give Algorithm 2, our general algorithmic framework for solving (7) given efficient bounded estimators for the gradient. Its theoretical guarantees are given in Theorem 3 which bounds the number of iterations needed to obtain expected ϵ-approximate solution.

Theorem 3. *Given an* ℓ_∞-ℓ_1 *game, i.e.* (7)*, and desired accuracy* ϵ*, a* $(v^{\mathsf{x}},\|\cdot\|_2)$*-bounded estimator* $\tilde{g}^{\mathsf{x}}$ *, and a* $(\frac{4v^{\mathsf{y}}}{\epsilon},v^{\mathsf{y}},\|\cdot\|_{\Delta^m})$*-bounded estimator* $\tilde{g}^{\mathsf{y}}$*, Algorithm 2 with*

Algorithm 2 SMD for ℓ_∞-ℓ_1 game

Input: Desired accuracy ϵ, primal domain size b, $(v^{\mathsf{x}}, \|\cdot\|_2)$-bounded estimator g^{x}, $(\frac{20v^{\mathsf{y}}}{\epsilon}, v^{\mathsf{y}}, \|\cdot\|_{\Delta^m})$-bounded estimator g^{y}
Output: An expected ϵ-approximate solution $(\mathbf{x}^\epsilon, \mathbf{y}^\epsilon)$ for problem (7).
Parameter: Step-size η^{x}, η^{y}, total iteration number T.
for $t = 0, \ldots, T-1$ **do**
  Get $\tilde{g}^{\mathsf{x}}$ estimator for g^{x}, $\tilde{g}^{\mathsf{y}}$ estimator for g^{y}
  Update $\mathbf{x}_{t+1} \leftarrow \arg\min_{\mathbf{x}\in\mathbb{B}_b^n} \langle \eta^{\mathsf{x}} \tilde{g}^{\mathsf{x}}(\mathbf{x}_t, \mathbf{y}_t), \mathbf{x}\rangle + V_{\mathbf{x}_t}(\mathbf{x})$
  Update $\mathbf{y}_{t+1} \leftarrow \arg\min_{\mathbf{y}\in\Delta^m} \langle \eta^{\mathsf{y}} \tilde{g}^{\mathsf{y}}(\mathbf{x}_t, \mathbf{y}_t), \mathbf{y}\rangle + V_{\mathbf{y}_t}(\mathbf{y})$
end for
Return $(\mathbf{x}^\epsilon, \mathbf{y}^\epsilon) \leftarrow \frac{1}{T}\sum_{t\in[T]}(\mathbf{x}_t, \mathbf{y}_t)$

choice of parameters $\eta_x \le \frac{\epsilon}{4v^{\mathsf{x}}}$, $\eta_y \le \frac{\epsilon}{4v^{\mathsf{y}}}$ outputs an expected ϵ-approximate optimal solution within any iteration number $T \ge \max\{\frac{16nb^2}{\epsilon\eta_x}, \frac{8\log m}{\epsilon\eta_y}\}$.

Sketch of Proof For simplicity we only include a proof sketch here and defer the complete proof details to Appendix C.2.

Regret bounds with local norms. The core statement is a standard regret bound using local norms (see Lemma 10 and Lemma 11) which summing together gives the following guarantee (let $\tilde{g}_t^{\mathsf{x}}, \tilde{g}_t^{\mathsf{y}}$ denote $\tilde{g}^{\mathsf{x}}(\mathbf{x}_t, \mathbf{y}_t), \tilde{g}^{\mathsf{y}}(\mathbf{x}_t, \mathbf{y}_t)$)

$$\sum_{t\in[T]} \langle \tilde{g}_t^{\mathsf{x}}, \mathbf{x}_t - \mathbf{x}\rangle + \sum_{t\in[T]} \langle \tilde{g}_t^{\mathsf{y}}, \mathbf{y}_t - \mathbf{y}\rangle \le \frac{V_{\mathbf{x}_0}(\mathbf{x})}{\eta^{\mathsf{x}}} + \frac{\sum_{t=0}^{T} \eta^{\mathsf{x}} \|\tilde{g}_t^{\mathsf{x}}\|_2^2}{2} + \frac{V_{\mathbf{y}_0}(\mathbf{y})}{\eta^{\mathsf{y}}} + \frac{\sum_{t=0}^{T} \eta^{\mathsf{y}} \|\tilde{g}_t^{\mathsf{y}}\|_{\mathbf{y}_t}^2}{2}. \tag{8}$$

Note one needs the bounded maximum entry condition for $\tilde{g}^{\mathsf{y}}$ as the condition to use Lemma 11.

Domain size. The domain size can be bounded as $\max_{\mathbf{x}\in\mathbb{B}_b^n} V_{\mathbf{x}_0}(\mathbf{x}) \le 2nb^2$, $\max_{\mathbf{y}\in\Delta^m} V_{\mathbf{y}_0}(\mathbf{y}) \le \log m$ by definition of their corresponding Bregman divergences.

Second-moment bounds. This is given through the bounded second-moment properties of estimators directly.

Ghost-iterate analysis.[6] In order to substitute $\tilde{g}^{\mathsf{x}}, \tilde{g}^{\mathsf{y}}$ with $g^{\mathsf{x}}, g^{\mathsf{y}}$ for LHS of Eq. (8), one can apply the regret bounds again to ghost iterates generated by taking gradient step with $\hat{g} = g - \tilde{g}$ coupled with each iteration. The additional terms coming from this extra regret bounds are in expectation 0 through conditional expectation computation.

Optimal tradeoff. One pick η_x, η_y, T accordingly to get the desired guarantee as stated in Theorem 3. □

[6]For standard SMD on convex problems this step is unnecessary. One can directly use conditional expectation by fixing $\mathbf{x} = \mathbf{x}^*$. However, for saddle-point problems, the same technique only gives $\max_{\mathbf{x},\mathbf{y}} \mathbb{E}[\text{regret}] \le \epsilon$. The ghost iterates analysis is standard (Nemirovski, 2004a; Carmon et al., 2019) and necessary to get a bound in terms of $\mathbb{E}\max_{\mathbf{x},\mathbf{y}}[\text{regret}] \le \epsilon$ instead.

Now we design gradient estimators assuming certain sampling oracles to ensure good bounded properties.

When $\mathbf{x} \in \mathbb{B}_1^n$, this leads to the theoretical guarantee as stated formally in Corollary 1. We defer design of estimators and all proofs to Appendix C.3 and simply state the theoretical runtime guarantee of Algorithm 2 here.

Corollary 1. *Given an ℓ_∞-ℓ_1 game (7) with domains $\mathbf{x} \in \mathbb{B}_1^n$, $\mathbf{y} \in \Delta^m$, $\epsilon \in (0,1)$ and $\|\mathbf{M}\|_\infty + \|\mathbf{c}\|_\infty = \Omega(1)$. If one has all sampling oracles needed*[7]*, Algorithm 2 with certain gradient estimators (see (19) and (20)) finds an expected ϵ-approximate solution in runtime (equivalent as sample complexity here) $O([(n + m\log m)\|\mathbf{M}\|_\infty^2 + n\|\mathbf{b}\|_1^2 + m\log m\|\mathbf{c}\|_\infty^2] \cdot \epsilon^{-2})$.*

Finally, we remark that one can also use Algorithm 2 to solve ℓ_∞-regression, i.e. the problem of finding $\mathbf{x}^* := \arg\min_{\mathbf{x}\in\mathbb{B}_1^n} \|\mathbf{M}\mathbf{x} - \mathbf{c}\|_\infty$ by simply writing it in equivalent minimax form of $\min_{\mathbf{x}\in\mathbb{B}_1^n}\max_{\mathbf{y}\in\Delta^m} \mathbf{y}^\top(\hat{\mathbf{M}}\mathbf{x} - \hat{\mathbf{c}})$ where $\hat{\mathbf{M}} := [\mathbf{M}; -\mathbf{M}]$ and $\hat{\mathbf{c}} := [\mathbf{c}; -\mathbf{c}]$.

Remark 1. *Algorithm 2 produces an expected ϵ-approximate solution $\mathbf{x}^\epsilon$ satisfying $\mathbb{E}\|\mathbf{M}\mathbf{x}^\epsilon - \mathbf{c}\|_\infty \le \|\mathbf{M}\mathbf{x}^* - \mathbf{c}\|_\infty + \epsilon$, within runtime $\widetilde{O}\left(\left[(m+n)\|\mathbf{M}\|_\infty^2 + m\|\mathbf{c}\|_\infty^2\right] \cdot \epsilon^{-2}\right)$.*

4 Gradient Estimators for Mixing AMDP and DMDP

Specifically, both MDP problems induced by solving mixing or DMDP are in minimax form of (7) if we let $\mathbf{y} \leftarrow \boldsymbol{\mu}$ with $m = \mathrm{A}_{\text{tot}}$, $\mathbf{x} \leftarrow \mathbf{v}$ with $n = |\mathcal{S}|$ and $b \leftarrow 2M$ with M chosen to be $M = 3t_{\text{mix}}$ for mixing AMDP and $M = \frac{1}{1-\gamma}$ for DMDP so that $\mathbf{v}^* \in \mathbb{B}_M^{\mathcal{S}}$; we defer readers to proofs of Lemma 6 and 9 for a complete argument on it.

In this section, we give a cleaner way to construct gradient estimators with desired properties for mixing and discounted cases utilizing problem structure and the generative model at hand. Such a gradient estimator samples state-action pair for $\mathbf{v}$-side using a dynamic distribution induced by $\boldsymbol{\mu}$, while sampling state-action pair for $\boldsymbol{\mu}$-side using a uniform distribution. We defer all proofs in this section to Appendix D.

4.1 Mixing AMDPs

For the mixing case, we set $M = 3t_{\text{mix}}$ to guarantee $\mathbf{v}^* \in \mathbb{B}_{2M}^{\mathcal{S}}$.[8] This follows immediately from a lemma that relates

[7]Note all the sampling oracles needed are essentially ℓ_1 samplers proportional to the matrix / vector entries, and an ℓ_1 sampler induced by $y \in \Delta^m$.

[8]Note the one can show $\|\mathbf{v}^*\|_\infty \le M$ and the extra coefficient 2 in box size $2M$ is to ensure a stronger condition on dual constraints for approximate solutions, which can be seen more clearly in proof of Lemma 6

matrix norm of interest to the mixing property of MDP, which we defer readers to Appendix D.1 for details.

Given domain setups, now we describe formally the gradient estimators used in Algorithm 1 and their properties.

For the $\mathbf{v}$-side, we consider the following gradient estimator

$$\begin{aligned} &\text{Sample } (i, a_i) \sim [\boldsymbol{\mu}]_{i,a_i}, j \sim p_{ij}(a_i). \\ &\text{Set } \tilde{g}^{\mathsf{v}}(\mathbf{v}, \boldsymbol{\mu}) = \mathbf{e}_j - \mathbf{e}_i. \end{aligned} \tag{9}$$

This is a bounded gradient estimator for the box domain.

Lemma 1. *$\tilde{g}^{\mathsf{v}}$ as in (9) is a $(2, \|\cdot\|_2)$-bounded estimator.*

For the $\boldsymbol{\mu}$-side, we consider the following gradient estimator

$$\begin{aligned} &\text{Sample } (i, a_i) \sim 1/\mathrm{A}_{\mathrm{tot}}, j \sim p_{ij}(a_i). \\ &\text{Set } \tilde{g}^{\mu}(\mathbf{v}, \boldsymbol{\mu}) = \mathrm{A}_{\mathrm{tot}}(v_i - \gamma v_j - r_{i,a_i})\mathbf{e}_{i,a_i}. \end{aligned} \tag{10}$$

This is a bounded gradient estimator for the simplex domain.

Lemma 2. *$\tilde{g}^{\mu}$ defined in (10) is a $((2M+1)\mathrm{A}_{\mathrm{tot}}, 9(M^2+1)\mathrm{A}_{\mathrm{tot}}, \|\cdot\|_{\Delta^{\mathcal{A}}})$-bounded estimator.*

Theorem 3 together with guarantees of designed gradient estimators in Lemma 1, 2 and choice of $M = 3t_{\mathrm{mix}}$ gives Corollary 4.

Corollary 2. *Given mixing AMDP tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{r})$ with desired accuracy $\epsilon \in (0,1)$, Algorithm 1 with parameter choice $\eta^{\mathsf{v}} = O(\epsilon)$, $\eta^{\mu} = O(\epsilon t_{\mathrm{mix}}^{-2} {\mathrm{A}_{\mathrm{tot}}}^{-1})$ outputs an expected ϵ-approximate solution to mixing minimax problem (5) with sample complexity $O(t_{\mathrm{mix}}^2 \mathrm{A}_{\mathrm{tot}} \epsilon^{-2} \log(\mathrm{A}_{\mathrm{tot}}))$.*

The proof follows immediately by noticing each iteration costs $O(1)$ sample generation, thus directly transferring the total iteration number to sample complexity.

4.2 DMDPs

Here, we pick $M = (1-\gamma)^{-1}$, with the guarantee that $\mathbf{v}^* \in \mathbb{B}^{\mathcal{S}}_{2M}$ following from Lemma 15 we state and prove in Appendix D.1.

For discounted case one construct gradient estimators in a similar way. For the $\mathbf{v}$-side, we consider the following gradient estimator

$$\begin{aligned} &\text{Sample } (i, a_i) \sim [\boldsymbol{\mu}]_{i,a_i}, j \sim p_{ij}(a_i), i' \sim q_{i'} \\ &\text{Set } \tilde{g}^{\mathsf{v}}(\mathbf{v}, \boldsymbol{\mu}) = (1-\gamma)\mathbf{e}_{i'} + \gamma \mathbf{e}_j - \mathbf{e}_i. \end{aligned} \tag{11}$$

Lemma 3. *$\tilde{g}^{\mathsf{v}}$ as in (11) is a $(2, \|\cdot\|_2)$-bounded estimator.*

For the $\boldsymbol{\mu}$-side, we consider the following gradient estimator

$$\begin{aligned} &\text{Sample } (i, a_i) \sim \frac{1}{\mathrm{A}_{\mathrm{tot}}}, j \sim p_{ij}(a_i). \\ &\text{Set } \tilde{g}^{\mu}(\mathbf{v}, \boldsymbol{\mu}) = \mathrm{A}_{\mathrm{tot}}(v_i - \gamma v_j - r_{i,a_i})\mathbf{e}_{i,a_i}. \end{aligned} \tag{12}$$

Lemma 4. *$\tilde{g}^{\mu}$ defined in (12) is a $((2M+1)\mathrm{A}_{\mathrm{tot}}, 9(M^2+1)\mathrm{A}_{\mathrm{tot}}, \|\cdot\|_{\Delta^{\mathcal{A}}})$-bounded estimator.*

Theorem 3 together with guarantees of gradient estimators in use in Lemma 3, 4 and choice of $M = (1-\gamma)^{-1}$ gives Corollary 3.

Corollary 3. *Given DMDP tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathbf{r}, \gamma)$ with desired accuracy $\epsilon \in (0,1)$, Algorithm 1 outputs an expected ϵ-approximate solution to discounted minimax problem (6) with sample complexity $O((1-\gamma)^{-2}\mathrm{A}_{\mathrm{tot}}\epsilon^{-2}\log(\mathrm{A}_{\mathrm{tot}}))$.*

5 From Optimal Solution to Optimal Policy

In this section, we relate the quality of an approximate solution to minimax problem to the quality of approximate policy one can construct from it, for both cases. To do that, we show that the enlarged primal space (size $2M$ instead of M) allows one to bound the quality of dual variables $\boldsymbol{\mu}^\epsilon$ better. We defer all proofs of this section to Appendix E.

5.1 Mixing AMDPs

Now we proceed to show how to convert an ϵ-approximate solution of (5) to an $\Theta(\epsilon)$-approximate policy for (3).

First we introduce a lemma that relates the dual variable $\boldsymbol{\mu}^\epsilon$ with optimal cost-to-go values $\mathbf{v}^*$ and expected reward $\bar{v}^*$.

Lemma 5. *If $(\mathbf{v}^\epsilon, \boldsymbol{\mu}^\epsilon)$ is an expected ϵ-approximate optimal solution to mixing AMDP minimax problem (5) , then for any optimal $\mathbf{v}^*$ and $\bar{v}^*$, $\mathbb{E}\left[{\boldsymbol{\mu}^{\epsilon}}^{\top}\left[(\hat{\mathbf{I}} - \mathbf{P})\mathbf{v}^* - \mathbf{r}\right] + \bar{v}^*\right] \le \epsilon$.*

Next we transfer an optimal solution to an optimal policy, formally through Lemma 6.

Lemma 6. *Given an ϵ-approximate solution $(\mathbf{v}^\epsilon, \boldsymbol{\mu}^\epsilon)$ for mixing minimax problem as defined in (5), let π^ϵ be the unique decomposition (in terms of $\boldsymbol{\lambda}^\epsilon$) such that $\mu^\epsilon_{i,a_i} = \lambda^\epsilon_i \cdot \pi^\epsilon_{i,a_i}, \forall i \in \mathcal{S}, a_i \in \mathcal{A}_i$, where $\boldsymbol{\lambda} \in \Delta^{\mathcal{S}}, \pi^\epsilon_i \in \Delta^{\mathcal{A}_i}, \forall i \in \mathcal{S}$. Taking $\pi := \pi^\epsilon$ as our policy, it holds that $\bar{v}^* \le \mathbb{E}\bar{v}^\pi + 3\epsilon$.*

To prove the lemma, we need the following helper lemma bounding the matrix norm.

Lemma 7. *Given a mixing AMDP, policy π, and its probability transition matrix $\mathbf{P}^\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$ and stationary distribution $\boldsymbol{\nu}^\pi$, $\left\|(\mathbf{I} - \mathbf{P}^\pi + \mathbf{1}(\boldsymbol{\nu}^\pi)^\top)^{-1}\right\|_\infty \le 3t_{\mathrm{mix}}$.*

Using this fact one can prove Lemma 6 by showing the linear constraints in dual formulation (D) of (3) are approximately satisfied given an ϵ-approximate optimal solution $(\mathbf{v}^\epsilon, \boldsymbol{\mu}^\epsilon)$ to minimax problem (5).

Lemma 6 shows one can construct an expected ϵ-optimal policy from an expected $\epsilon/3$-approximate solution of the minimax problem (5). Thus, using Corollary 4 one directly obtains our desired total sample complexity for Algorithm 1 to solve mixing AMDPs to desired accuracy, as stated in Theorem 1.

5.2 DMDPs

Now we show how to convert an ϵ-approximate solution of (6) to an $\Theta((1-\gamma)^{-1}\epsilon)$-approximate policy of (4).

First we introduce a lemma similar to Lemma 5 that relates the dual variable $\boldsymbol{\mu}^\epsilon$ with values $\mathbf{v}^*$ under ϵ-approximation.

Lemma 8. *If $(\mathbf{v}^\epsilon, \boldsymbol{\mu}^\epsilon)$ is an ϵ-approximate optimal solution to the DMDP minimax problem (6), then for optimal $\mathbf{v}^*$, $\mathbb{E}\boldsymbol{\mu}^{\epsilon\top}\left[(\hat{\mathbf{I}}-\gamma\mathbf{P})\mathbf{v}^*-\mathbf{r}\right]\le\epsilon$.*

Next we transfer an optimal solution to an optimal policy, formally through Lemma 9.

Lemma 9. *Given an expected ϵ-approximate solution $(\mathbf{v}^\epsilon, \boldsymbol{\mu}^\epsilon)$ for discounted minimax problem as defined in (6), let π^ϵ be the unique decomposition (in terms of $\boldsymbol{\lambda}^\epsilon$) such that $\mu^\epsilon_{i,a_i}=\lambda^\epsilon_i\cdot\pi^\epsilon_{i,a_i},\forall i\in\mathcal{S},a_i\in\mathcal{A}_i$, where $\boldsymbol{\lambda}\in\Delta^{\mathcal{S}},\pi^\epsilon_i\in\Delta^{\mathcal{A}_i},\forall i\in\mathcal{S}$. Taking $\pi:=\pi^\epsilon$ as our policy, it holds that $\bar{v}^*\le\mathbb{E}\bar{v}^\pi+3\epsilon/(1-\gamma)$.*

Lemma 9 shows it suffices to find an expected $(1-\gamma)\epsilon$-approximate solution to problem (6) to get an expected ϵ-optimal policy. Together with Corollary 3 this directly yields the sample complexity as claimed in Theorem 2.

6 Constrained MDP

In this section, we consider solving a generalization of the mixing AMDP problem with additional linear constraints, which has been an important and well-known problem class along the study of MDP (Altman, 1999); we defer readers to Appendix G for derivation ommitted in this section.

Formally, we focus on approximately solving the following dual formulation of constrained mixing AMDPs [9] :

$$\begin{aligned}(\mathrm{D})\quad &\max_{\boldsymbol{\mu}\in\Delta^{\mathcal{A}}} && 0\\ &\text{subject to} && (\hat{\mathbf{I}}-\mathbf{P})^\top\boldsymbol{\mu}=\mathbf{0},\quad \mathbf{D}^\top\boldsymbol{\mu}\ge\mathbf{1},\end{aligned}\tag{13}$$

where $\mathbf{D}=\begin{bmatrix}\mathbf{d}_1 & \cdots & \mathbf{d}_K\end{bmatrix}$ under the additional assumptions that $\mathbf{d}_k\ge\mathbf{0},\forall k\in[K]$ and the problem is strictly feasible (with an inner point in its feasible set). Our goal is to compute ϵ-approximate policies and solutions for (13) defined as follows.

Definition 2. *Given a policy π with its stationary distribution $\boldsymbol{\nu}^\pi$, it is an ϵ-approximate policy of system (13) if for $\boldsymbol{\mu}$ defined as $\mu_{i,a_i}=\nu^\pi_i\pi_{i,a_i},\forall i\in\mathcal{S},a_i\in\mathcal{A}_i$ it is an ϵ-approximate solution of (13), i.e. it satisfies*

$$\boldsymbol{\mu}^\top(\hat{\mathbf{I}}-\mathbf{P})=\mathbf{0},\quad \mathbf{D}^\top\boldsymbol{\mu}\ge(1-\epsilon)\mathbf{1}.\tag{14}$$

For $D:=\|\mathbf{D}\|_{\max}:=\max_{i,a_i,k}|[d_k]_{i,a_i}|$ and $M:=3Dt_{\mathrm{mix}}$ we consider the following equivalent problem:

$$\min_{\mathbf{v}\in\mathbb{B}^{\mathcal{S}}_{2M},\mathbf{s}:\sum_k s_k\le 2,\mathbf{s}\ge\mathbf{0}}\ \max_{\boldsymbol{\mu}\in\Delta^{\mathcal{A}}}\ f(\mathbf{v},\mathbf{s},\boldsymbol{\mu})\tag{15}$$

where $f(\mathbf{v},\mathbf{s},\boldsymbol{\mu}):=\boldsymbol{\mu}^\top\left[(\hat{\mathbf{I}}-\mathbf{P})\mathbf{v}+\mathbf{D}\mathbf{s}\right]-\mathbf{1}^\top\mathbf{s}$.

Note in the formulation we pose the additional constraints on $\mathbf{v},\mathbf{s}$ for the sake of analysis. These constraints don't change the problem optimality by noticing $\mathbf{v}^*\in\mathbb{B}^{\mathcal{S}}_{2M}$, $\mathbf{s}^*\in\Delta^K$; see Appendix G for details.

By designing gradient estimators and choosing divergence terms properly, one can obtain an approximately optimal solution efficiently, and thus an approximately optimal policy.

Corollary 4. *Given mixing AMDP tuple $\mathcal{M}=(\mathcal{S},\mathcal{A},\mathcal{P},\mathbf{r})$ with constraints $D:=\max_{i,a_i,k}|[d_k]_{i,a_i}|$, for accuracy $\epsilon\in(0,1)$, Algorithm 1 with parameter choice $\eta^{\mathsf{v}}=O(\epsilon)$, $\eta^{\mathsf{s}}=O(\epsilon K^{-1}D^{-2})$, $\eta^\mu=O(\epsilon t_{\mathrm{mix}}^{-2}D^{-2}\mathrm{A_{tot}}^{-1})$ outputs an expected ϵ-approximate solution to constrained mixing minimax problem (15) with sample complexity $O((t^2_{\mathrm{mix}}\mathrm{A_{tot}}+K)D^2\epsilon^{-2}\log(\mathrm{A_{tot}}))$.*

Following the similar rounding technique as in Section 5, one can show the approximate solution $\boldsymbol{\mu}^\epsilon$ for minimax problem (15) leads to an approximately optimal policy π^ϵ an of problem (13).

Corollary 5. *Following the setting of Corollary 4, the policy π^ϵ induced by the unique decomposition of $\boldsymbol{\mu}^\epsilon$ from the output satisfying $\mu^\epsilon_{i,a_i}=\lambda^\epsilon_i\cdot\pi^\epsilon_{i,a_i}$, is an $O(\epsilon)$-approximate policy for system (13).*

7 Conclusion

This work offers a general framework based on stochastic mirror descent to find an ϵ-optimal policy for AMDPs and DMDPs. It improves over previous convex optimization approaches for solving these MDPs, achieving a better sample complexity and removing an ergodicity condition for mixing AMDP, while matching the known nearly-optimal algorithms up to $(1-\gamma)^{-1}$ factor for DMDPs.

This work reveals an interesting connection MDP problems and ℓ_∞-regression. We believe there are a number of interesting directions and open problems for future work, including getting optimal sample complexity for discounted case, obtaining high-precision algorithms, extending the framework to broader classes of MDPs, etc. See Appendix F for a more detailed discussion of these open directions. We hope a better understanding of these problems could lead to a more complete picture of solving MDP and RL using convex-optimization methods.

[9] One can reduce the general case of $\mathbf{D}^\top\boldsymbol{\mu}\ge\mathbf{c}$ for some $\mathbf{c}>0$ to this case by taking $\mathbf{d}_k\leftarrow\mathbf{d}_k/c_k$, under which an ϵ-approximate solution as defined in (14) of the modified problem corresponds to a multiplicatively approximate solution satisfying $\mathbf{D}^\top\boldsymbol{\mu}\ge(1-\epsilon)\mathbf{c}$.

Acknowledgements

This research was partially supported by NSF CAREER Award CCF-1844855, a PayPal research gift award, and a Stanford Graduate Fellowship. We thank Yair Carmon and Kevin Tian for helpful discussions on coordinate methods for matrix games; we thank Mengdi Wang, Xian Wu, Lin F. Yang, and Yinyu Ye for helpful discussions regarding MDPs; we also thank the anonymous reviewers who helped improve the completeness and readability of this paper by providing many helpful comments.

References

Agarwal, A., Kakade, S., and Yang, L. F. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83, 2020.

Altman, E. *Constrained Markov decision processes*, volume 7. CRC Press, 1999.

Auer, P. and Ortner, R. Logarithmic online regret bounds for undiscounted reinforcement learning. In *Advances in Neural Information Processing Systems*, pp. 49–56, 2007.

Azar, M. G., Munos, R., and Kappen, B. On the sample complexity of reinforcement learning with a generative model. *arXiv preprint arXiv:1206.6461*, 2012.

Bertsekas, D. P. and Tsitsiklis, J. N. Neuro-dynamic programming: an overview. In *Proceedings of 1995 34th IEEE Conference on Decision and Control*, volume 1, pp. 560–564. IEEE, 1995.

Bertsekas, D. P., Bertsekas, D. P., Bertsekas, D. P., and Bertsekas, D. P. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 1995.

Carmon, Y., Jin, Y., Sidford, A., and Tian, K. Variance reduction for matrix games. In *Advances in Neural Information Processing Systems*, pp. 11377–11388, 2019.

Carmon, Y., Jin, Y., Sidford, A., and Tian, K. Coordinate methods for matrix games. *to appear in Symposium on Foundations of Computer Science*, 2020.

Clarkson, K. L., Hazan, E., and Woodruff, D. P. Sublinear optimization for machine learning. *Journal of the ACM (JACM)*, 59(5):1–49, 2012.

Cohen, M. B., Kelner, J., Peebles, J., Peng, R., Sidford, A., and Vladu, A. Faster algorithms for computing the stationary distribution, simulating random walks, and more. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 583–592. IEEE, 2016.

Grigoriadis, M. D. and Khachiyan, L. G. A sublinear-time randomized approximation algorithm for matrix games. *Operations Research Letters*, 18(2):53–58, 1995.

Kakade, S. M. et al. *On the sample complexity of reinforcement learning*. PhD thesis, University of London London, England, 2003.

Kearns, M. J. and Singh, S. P. Finite-sample convergence rates for q-learning and indirect algorithms. In *Advances in neural information processing systems*, pp. 996–1002, 1999.

Lee, Y. T. and Sidford, A. Path finding methods for linear programming: Solving linear programs in o (vrank) iterations and faster algorithms for maximum flow. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pp. 424–433. IEEE, 2014.

Lee, Y. T. and Sidford, A. Efficient inverse maintenance and faster algorithms for linear programming. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pp. 230–249. IEEE, 2015.

Mahadevan, S. Average reward reinforcement learning: Foundations, algorithms, and empirical results. *Machine learning*, 22(1-3):159–195, 1996.

Nemirovski, A. Prox-method with rate of convergence o(1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004a.

Nemirovski, A. Prox-method with rate of convergence o (1/t) for variational inequalities with lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1): 229–251, 2004b.

Nesterov, Y. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1):127–152, 2005.

Nesterov, Y. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2-3):319–344, 2007.

Ouyang, Y., Gagrani, M., Nayyar, A., and Jain, R. Learning unknown markov decision processes: A thompson sampling approach. In *Advances in Neural Information Processing Systems*, pp. 1333–1342, 2017.

Palaniappan, B. and Bach, F. Stochastic variance reduction methods for saddle-point problems. In *Advances in Neural Information Processing Systems*, pp. 1416–1424, 2016.

Puterman, M. L. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Shalev-Shwartz, S. et al. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Sherman, J. Area-convexity, l-infinity regularization, and undirected multicommodity flow. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 452–460, 2017.

Sidford, A. and Tian, K. Coordinate methods for accelerating l-infinity regression and faster approximate maximum flow. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 922–933. IEEE, 2018.

Sidford, A., Wang, M., Wu, X., Yang, L., and Ye, Y. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pp. 5186–5196, 2018a.

Sidford, A., Wang, M., Wu, X., and Ye, Y. Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 770–787. SIAM, 2018b.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Wainwright, M. J. Variance-reduced q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019.

Wang, M. Randomized linear programming solves the discounted markov decision problem in nearly-linear running time. *arXiv preprint arXiv:1704.01869*, 2017a.

Wang, M. Primal-dual pi learning: Sample complexity and sublinear run time for ergodic markov decision problems. *arXiv preprint arXiv:1710.06100*, 2017b.

Ye, Y. The simplex and policy-iteration methods are strongly polynomial for the markov decision problem with a fixed discount rate. *Mathematics of Operations Research*, 36 (4):593–603, 2011.

Supplementary material

Appendix

A Complete Table for results on Sample Complexity of MDPs

Type	Method	Sample Complexity
mixing AMDP	Primal-Dual Method (Wang, 2017b)	$\widetilde{O}\left(\tau^2 t_{\mathrm{mix}}^2 \mathrm{A}_{\mathrm{tot}}\epsilon^{-2}\right)$
	Our method (Theorem 1)	$\widetilde{O}\left(t_{\mathrm{mix}}^2 \mathrm{A}_{\mathrm{tot}}\epsilon^{-2}\right)$
DMDP	Empirical QVI (Azar et al., 2012)	$\widetilde{O}\left((1-\gamma)^{-5}\mathrm{A}_{\mathrm{tot}}\epsilon^{-2}\right)$
	Empirical QVI (Azar et al., 2012)	$\widetilde{O}\left((1-\gamma)^{-3}\mathrm{A}_{\mathrm{tot}}\epsilon^{-2}\right)$, $\epsilon = \widetilde{O}(\frac{1}{\sqrt{(1-\gamma)\lvert\mathcal{S}\rvert}})$
	Primal-Dual Method (Wang, 2017a)	$\widetilde{O}\left((1-\gamma)^{-6}\lvert\mathcal{S}\rvert^2\mathrm{A}_{\mathrm{tot}}\epsilon^{-2}\right)$
	Primal-Dual Method (Wang, 2017a)	$\widetilde{O}\left(\tau^4(1-\gamma)^{-4}\mathrm{A}_{\mathrm{tot}}\epsilon^{-2}\right)$
	Variance-reduced Value Iteration (Sidford et al., 2018b)	$\widetilde{O}\left((1-\gamma)^{-4}\mathrm{A}_{\mathrm{tot}}\epsilon^{-2}\right)$
	Variance-reduced QVI (Sidford et al., 2018a)	$\widetilde{O}\left((1-\gamma)^{-3}\mathrm{A}_{\mathrm{tot}}\epsilon^{-2}\right)$
	Empirical MDP + Blackbox (Agarwal et al., 2020)	$\widetilde{O}\left((1-\gamma)^{-3}\mathrm{A}_{\mathrm{tot}}\epsilon^{-2}\right)$
	Variance-reduced Q-learning (Wainwright, 2019)	$\widetilde{O}\left((1-\gamma)^{-3}\mathrm{A}_{\mathrm{tot}}\epsilon^{-2}\right)$
	Our method (Theorem 2)	$\widetilde{O}\left((1-\gamma)^{-4}\mathrm{A}_{\mathrm{tot}}\epsilon^{-2}\right)$

Table 2. **Complete comparison of sample complexity to get ϵ-optimal policy among stochastic methods.** Here $\mathcal{S}$ denotes state space, $\mathrm{A}_{\mathrm{tot}}$ denotes number of state-action pair, t_{mix} is mixing time for mixing AMDP, and γ is discount factor for DMDP. Parameter τ shows up whenever the designed algorithm requires additional ergodic condition for MDP, i.e. there exists some distribution $\mathbf{q}$ and $\tau > 0$ satisfying $\sqrt{1/\tau}\mathbf{q} \le \boldsymbol{\nu}^\pi \le \sqrt{\tau}\mathbf{q}$, $\forall$ policy π and its induced stationary distribution $\boldsymbol{\nu}^\pi$.

B Table for ℓ_∞-regression Runtimes

Here we include Table 3 that make comparisons between our sublinear ℓ_∞ regression solver and prior art. We remark for dense matrix $\mathbf{M}$, our method is the only sublinear method along this line of work for approximately solving ℓ_∞ regression problem.

Method	Runtime
Smooth Approximation (Nesterov, 2005)	$\widetilde{O}(\mathrm{nnz}\,\lVert\mathbf{M}\rVert_\infty^2\,\epsilon^{-2})$ or $\widetilde{O}(\mathrm{nnz}\sqrt{n}\,\lVert\mathbf{M}\rVert_\infty\,\epsilon^{-1})$
Mirror-prox Method (Nemirovski, 2004b)	$\widetilde{O}(\mathrm{nnz}\,\lVert\mathbf{M}\rVert_\infty^2\,\epsilon^{-2})$
Dual Extrapolation (Nesterov, 2007)	$\widetilde{O}(\mathrm{nnz}\,\lVert\mathbf{M}\rVert_\infty^2\,\epsilon^{-2})$
Dual Extrapolation with Joint Regularizer (Sherman, 2017)	$\widetilde{O}(\mathrm{nnz}\,\lVert\mathbf{M}\rVert_\infty\,\epsilon^{-1})$
Accelerated Coordinate Method (Sidford & Tian, 2018)	$\widetilde{O}\left(nd^{2.5}\,\lVert\mathbf{M}\rVert_\infty\,\epsilon^{-1}\right)$
Our method (Remark 1)	$\widetilde{O}\left((m+n)\,\lVert\mathbf{M}\rVert_\infty^2\,\epsilon^{-2}\right)$

Table 3. **Comparison of method runtime to get an ϵ-approximate solution to ℓ_∞-regression.** For simplicity here we only state for the simplified problem in form $\min_{\mathbf{x}\in\mathbb{B}^n}\lVert\mathbf{M}\mathbf{x}\rVert_\infty$ where $\mathbf{M}\in\mathbb{R}^{m\times n}$ with nnz nonzero entries and d-sparse columns.

C Proofs for Section 3

C.1 Standard Mirror-Descent Analysis under Local Norms

Here we recast a few standard results on the analysis of mirror-descent using local norm (Shalev-Shwartz et al., 2012), which we use for proving Theorem 3. These are standard regret bounds for ℓ_2 and simplex respectively. First, we provide the well-known regret guarantee for $\mathbf{x} \in \mathbb{B}^n$, when choosing $V_{\mathbf{x}}(\mathbf{x}') \coloneqq \frac{1}{2}\|\mathbf{x}-\mathbf{x}'\|_2^2$.

Lemma 10 (cf. Lemma 12 in Carmon et al. (2019), restated)**.** *Let $T \in \mathbb{N}$ and let $\mathbf{x}_0 \in \mathcal{X}$, $\gamma_0, \gamma_1, \ldots, \gamma_T \in \mathcal{X}^*$, V is 1-strongly convex in $\|\cdot\|_2$. The sequence $\mathbf{x}_1, \ldots, \mathbf{x}_T$ defined by*

$$\mathbf{x}_t = \arg\min_{\mathbf{x}\in\mathcal{X}} \left\{ \left\langle \gamma_{t-1}, \mathbf{x} \right\rangle + V_{\mathbf{x}_{t-1}}(\mathbf{x}) \right\}$$

satisfies for all $\mathbf{x} \in \mathcal{X}$ (denoting $\mathbf{x}_{T+1} \coloneqq \mathbf{x}$),

$$\begin{aligned}\sum_{t\in[T]} \langle \gamma_t, \mathbf{x}_t - \mathbf{x}\rangle &\le V_{\mathbf{x}_0}(\mathbf{x}) + \sum_{t=0}^{T} \{\langle \gamma_t, \mathbf{x}_t - \mathbf{x}_{t+1}\rangle - V_{\mathbf{x}_t}(\mathbf{x}_{t+1})\} \\ &\le V_{\mathbf{x}_0}(\mathbf{x}) + \frac{1}{2}\sum_{t=0}^{T} \|\gamma_t\|_2^2 .\end{aligned}$$

Next, one can show a similar property holds true for $\mathbf{y} \in \Delta^m$, by choosing KL-divergence as Bregman divergence $V_{\mathbf{y}}(\mathbf{y}') \coloneqq \sum_{i\in[m]} y_i \log(y_i'/y_i)$, utilizing local norm $\|\cdot\|_{\mathbf{y}'}$.

Lemma 11 (cf. Lemma 13 in Carmon et al. (2019), immediate consequence)**.** *Let $T \in \mathbb{N}$, $\mathbf{y}_0 \in \mathcal{Y}$, $\gamma_0, \gamma_1, \ldots, \gamma_T \in \mathcal{Y}^*$ satisfying $\|\gamma_t\|_\infty \le 1.79, \forall t = 0, 1, \cdots, T$, and $V_{\mathbf{y}}(\mathbf{y}') \coloneqq \sum_{i\in[m]} y_i \log(y_i'/y_i)$. The sequence $\mathbf{y}_1, \ldots, \mathbf{y}_T$ defined by*

$$\mathbf{y}_t = \arg\min_{\mathbf{y}\in\mathcal{Z}} \left\{ \left\langle \gamma_{t-1}, \mathbf{y} \right\rangle + V_{\mathbf{y}_{t-1}}(\mathbf{y}) \right\}$$

satisfies for all $\mathbf{y} \in \mathcal{Y}$ (denoting $\mathbf{y}_{T+1} \coloneqq \mathbf{y}$),

$$\begin{aligned}\sum_{t\in[T]} \langle \gamma_t, \mathbf{y}_t - \mathbf{y}\rangle &\le V_{\mathbf{y}_0}(\mathbf{y}) + \sum_{t=0}^{T} \{\langle \gamma_t, \mathbf{y}_t - \mathbf{y}_{t+1}\rangle - V_{\mathbf{y}_t}(\mathbf{y}_{t+1})\} \\ &\le V_{\mathbf{y}_0}(\mathbf{y}) + \frac{1}{2}\sum_{t=0}^{T} \|\gamma_t\|_{\mathbf{y}_t}^2 .\end{aligned}$$

C.2 Deferred proof for Theorem 3.

Proof of Theorem 3. By the choice of η^{x} and conditions, one can immediately see that

$$\|\eta^{\mathsf{x}}\tilde{g}^{\mathsf{x}}\|_\infty \le 1.$$

Thus we can use regret bound of stochastic mirror descent with local norms in Lemma 11 and Lemma 10 which gives

$$\begin{aligned}\sum_{t\in[T]} \langle \eta^{\mathsf{x}}\tilde{g}^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t), \mathbf{x}_t - \mathbf{x}\rangle &\le V_{\mathbf{x}_0}(\mathbf{x}) + \frac{{\eta^{\mathsf{x}}}^2}{2}\sum_{t=0}^{T} \|\tilde{g}^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t)\|_2^2, \\ \sum_{t\in[T]} \langle \eta^{\mathsf{y}}\tilde{g}^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t), \mathbf{y}_t - \mathbf{y}\rangle &\le V_{\mathbf{y}_0}(\mathbf{y}) + \frac{{\eta^{\mathsf{y}}}^2}{2}\sum_{t=0}^{T} \|\tilde{g}^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t)\|_{\mathbf{y}_t}^2 .\end{aligned} \tag{16}$$

Now, let $\hat{g}^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t) \coloneqq g^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t) - \tilde{g}^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t)$ and $\hat{g}^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t) \coloneqq g^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t) - \tilde{g}^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t)$, defining the sequence $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots, \hat{\mathbf{x}}_T$ and $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_T$ according to

$$\begin{aligned}\hat{\mathbf{x}}_0 = \mathbf{x}_0, \quad \hat{\mathbf{x}}_t = \arg\min_{\mathbf{x}\in\mathbb{B}_b^n} \langle \eta^{\mathsf{x}}\hat{g}^{\mathsf{x}}(\mathbf{x}_{t-1},\mathbf{y}_{t-1}), \mathbf{x}\rangle + V_{\hat{\mathbf{x}}_{t-1}}(\mathbf{x}); \\ \hat{\mathbf{y}}_0 = \mathbf{y}_0, \quad \hat{\mathbf{y}}_t = \arg\min_{\mathbf{y}\in\Delta^m} \langle \eta^{\mathsf{y}}\hat{g}^{\mathsf{y}}(\mathbf{x}_{t-1},\mathbf{y}_{t-1}), \mathbf{y}\rangle + V_{\hat{\mathbf{y}}_{t-1}}(\mathbf{y}).\end{aligned}$$

Using the similar argument we get

$$\sum_{t\in[T]}\langle \eta^{\mathsf{x}}\hat{g}^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t),\hat{\mathbf{x}}_t-\mathbf{x}\rangle\le V_{\mathbf{x}_0}(\mathbf{x})+\frac{\eta^{\mathsf{x}2}}{2}\sum_{t=0}^{T}\|\hat{g}^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t)\|_2^2,$$
$$\sum_{t\in[T]}\langle \eta^{\mathsf{y}}\hat{g}^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t),\hat{\mathbf{y}}_t-\mathbf{y}\rangle\le V_{\mathbf{y}_0}(\mathbf{y})+\frac{\eta^{\mathsf{y}2}}{2}\sum_{t=0}^{T}\|\tilde{g}^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t)\|_{\mathbf{y}_t}^2. \quad (17)$$

If we write $g=\tilde{g}+\hat{g}$ and rearrange terms we get

$$\begin{aligned}
&\sum_{t\in[T]}\left[\langle g^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t),\mathbf{x}_t-\mathbf{x}\rangle+\langle g^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t),\mathbf{y}_t-\mathbf{y}\rangle\right]\\
=&\sum_{t\in[T]}\left[\langle \tilde{g}^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t),\mathbf{x}_t-\mathbf{x}\rangle+\langle \tilde{g}^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t),\mathbf{y}_t-\mathbf{y}\rangle\right]+\sum_{t\in[T]}\left[\langle \hat{g}^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t),\mathbf{x}_t-\mathbf{x}\rangle+\langle \hat{g}^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t),\mathbf{y}_t-\mathbf{y}\rangle\right]\\
\overset{(i)}{\le}&\frac{2}{\eta^{\mathsf{x}}}V_{\mathbf{x}_0}(\mathbf{x})+\sum_{t\in[T]}\left[\frac{\eta^{\mathsf{x}}}{2}\|\tilde{g}^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t)\|_2^2+\frac{\eta^{\mathsf{x}}}{2}\|\hat{g}^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t)\|_2^2+\langle \hat{g}^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t),\mathbf{x}_t-\hat{\mathbf{x}}_t\rangle\right]\\
&+\frac{2}{\eta^{\mathsf{y}}}V_{\mathbf{y}_0}(\mathbf{y})+\sum_{t\in[T]}\left[\frac{\eta^{\mathsf{y}}}{2}\|\tilde{g}^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t)\|_{\mathbf{y}_t}^2+\frac{\eta^{\mathsf{y}}}{2}\|\hat{g}^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t)\|_{\hat{\mathbf{y}}_t}^2+\langle \hat{g}^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t),\mathbf{y}_t-\hat{\mathbf{y}}_t\rangle\right].
\end{aligned} \quad (18)$$

where we use the regret bounds in Eq. (16), (17) for (i).

Now take supremum over $(\mathbf{x},\mathbf{y})$ and then take expectation on both sides, we get

$$\begin{aligned}
\frac{1}{T}\mathbb{E}\sup_{\mathbf{x},\mathbf{y}}\left[\sum_{t\in[T]}\langle g^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t),\mathbf{x}_t-\mathbf{x}\rangle+\sum_{t\in[T]}\langle g^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t),\mathbf{y}_t-\mathbf{y}\rangle\right]&\overset{(i)}{\le}\sup_{\mathbf{x}}\frac{2}{\eta^{\mathsf{x}}T}V_{\mathbf{x}_0}(\mathbf{x})+\eta^{\mathsf{x}}v^{\mathsf{x}}+\sup_{\mathbf{y}}\frac{2}{\eta^{\mathsf{y}}T}V_{\mathbf{y}_0}(\mathbf{y})+\eta^{\mathsf{y}}v^{\mathsf{y}}\\
&\overset{(ii)}{\le}\frac{4nb^2}{\eta^{\mathsf{x}}T}+\eta^{\mathsf{x}}v^{\mathsf{x}}+\frac{2\log m}{\eta^{\mathsf{y}}T}+\eta^{\mathsf{y}}v^{\mathsf{y}}\\
&\overset{(iii)}{\le}\epsilon,
\end{aligned}$$

where we use (i) $\mathbb{E}[\langle \hat{g}^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t),\mathbf{x}_t-\hat{\mathbf{x}}_t\rangle|0,1,2,\cdots,t-1]=0$, $\mathbb{E}[\langle \hat{g}^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t),\mathbf{y}_t-\hat{\mathbf{y}}_t\rangle|0,1,2,\cdots,t-1]=0$ by conditional expectation, that $\mathbb{E}\|\hat{g}^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t)\|_2^2\le\mathbb{E}\|\tilde{g}^{\mathsf{x}}(\mathbf{x}_t,\mathbf{y}_t)\|_2^2$, $\mathbb{E}[\sum_i[\hat{\mathbf{y}}_t]_i[\hat{g}^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t)]_i^2]\le\mathbb{E}[\sum_i[\hat{\mathbf{y}}_t]_i[\tilde{g}^{\mathsf{y}}(\mathbf{x}_t,\mathbf{y}_t)]_i^2]$ due to the fact that $\mathbb{E}[(X-\mathbb{E}X)^2]\le\mathbb{E}[X^2]$ elementwise, and properties of estimators as stated in condition; (ii) $V_{\mathbf{x}_0}(\mathbf{x}):=\frac{1}{2}\|\mathbf{x}-\mathbf{x}_0\|_2^2\le 2nb^2$, $V_{\mathbf{y}_0}(\mathbf{y})\le\log m$ by properties of KL-divergence; (iii) the choice of $\eta^{\mathsf{x}}=\frac{\epsilon}{4v^{\mathsf{x}}}$, $\eta^{\mathsf{y}}=\frac{\epsilon}{4v^{\mathsf{y}}}$, and $T\ge\max(\frac{16nb^2}{\epsilon\eta^{\mathsf{x}}},\frac{8\log m}{\epsilon\eta^{\mathsf{y}}})$.

Together with the bilinear structure of problem and choice of $\mathbf{x}^\epsilon=\frac{1}{T}\sum_{t\in[T]}\mathbf{x}_t$, $\mathbf{y}^\epsilon=\frac{1}{T}\sum_{t\in[T]}\mathbf{y}_t$ we get $\mathbb{E}[\mathrm{Gap}(\mathbf{x}^\epsilon,\mathbf{y}^\epsilon)]\le\epsilon$, proving the output $(\mathbf{x}^\epsilon,\mathbf{y}^\epsilon)$ is indeed an expected ϵ-approximate solution to the minimax problem (7). □

C.3 Solving ℓ_∞-ℓ_1 Problem Approximately

More concretely, we offer one way to construct the gradient estimators, prove its properties and the implied algorithmic complexity.

For $\mathbf{x}$-side, we consider

$$\begin{aligned}
&\text{Sample } i,j \text{ with probability } p_{ij}:=\frac{|M_{ij}|y_i}{\sum_{i,j}|M_{ij}|y_i},\\
&\text{sample } j' \text{ with probability } p_{j'}:=\frac{|b_{j'}|}{\|\mathbf{b}\|_1},\\
&\text{set } \tilde{g}^{\mathsf{x}}(\mathbf{x},\mathbf{y})=\frac{M_{ij}y_i}{p_{ij}}\mathbf{e}_j+\frac{b_{j'}}{p_{j'}}\mathbf{e}_{j'},
\end{aligned} \quad (19)$$

which has properties as stated in Lemma 12.

Lemma 12. *Gradient estimator* $\tilde{g}^{\mathsf{x}}$ *specified in* (19) *is a* $(v^{\mathsf{x}}, \|\cdot\|_2)$*-bounded estimator, with*

$$v^{\mathsf{x}} = 2\left[\|\mathbf{b}\|_1^2 + \|\mathbf{M}\|_\infty^2\right].$$

Proof. The unbiasedness follows directly by definition. For bound on second-moment, one sees

$$\mathbb{E}\left\|\tilde{g}^{\mathsf{x}}(\mathbf{x},\mathbf{y})\right\|_2^2 \overset{(i)}{\leq} 2\left[\sum_{j'}\frac{b_{j'}^2}{p_{j'}} + \sum_{i,j}\frac{M_{ij}^2 y_i^2}{p_{ij}}\right] \overset{(ii)}{=} 2\left[\|\mathbf{b}\|_1^2 + \left(\sum_{i,j} M_{ij} y_i\right)^2\right] \overset{(iii)}{\leq} 2\left[\|\mathbf{b}\|_1^2 + \|\mathbf{M}\|_\infty^2\right],$$

where we use (i) the fact that $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$ and taking expectation, (ii) plugging in the explicit sampling probabilities as stated in (19), and (iii) Cauchy-Schwarz inequality and the fact that $\mathbf{y} \in \Delta^m$. □

For $\mathbf{y}$-side, we consider

$$\begin{aligned}
&\text{Sample } i, j \text{ with probability } q_{ij} := \frac{|M_{ij} x_j|}{\sum_{i,j}|M_{ij} x_j|},\\
&\text{sample } i' \text{ with probability } q_{i'} := \frac{|c_{i'}|}{\|\mathbf{c}\|_1},\\
&\text{set } \tilde{g}^{\mathsf{y}}(\mathbf{x},\mathbf{y}) = -\frac{M_{ij} x_j}{q_{ij}}\mathbf{e}_i + \frac{c_{i'}}{q_{i'}}\mathbf{e}_{i'}.
\end{aligned} \tag{20}$$

Here we remark that we adopt the same indexing notation i, j but it is independently sampled from given distributions as with ones for $\tilde{g}^{\mathsf{x}}$. Such an estimator has properties stated in Lemma 13.

Lemma 13. *Gradient estimator* $\tilde{g}^{\mathsf{y}}$ *specified in* (20) *is a* $(c^{\mathsf{y}}, v^{\mathsf{y}}, \|\cdot\|_{\Delta^m})$*-bounded estimator, with*

$$c^{\mathsf{y}} = m(b\|\mathbf{M}\|_\infty + \|\mathbf{c}\|_\infty),\ v^{\mathsf{y}} = 2m\left[\|\mathbf{c}\|_\infty^2 + b^2\|\mathbf{M}\|_\infty^2\right].$$

Proof. The unbiasedness follows directly by definition. For bounded maximum entry, one has

$$\|\tilde{g}^{\mathsf{y}}\|_\infty \leq \sum_{i,j}|M_{ij} x_j| + \|\mathbf{c}\|_1 \leq m(b\|\mathbf{M}\|_\infty + \|\mathbf{c}\|_\infty),$$

by definition of the probability distributions and $x_j \in \mathbb{B}_b^n, \mathbf{c} \in \mathbb{R}^m$.

For bound on second-moment in local norm with respect to arbitrary $\mathbf{y}' \in \Delta^m$, one has

$$\begin{aligned}
\mathbb{E}\left\|\tilde{g}^{\mathsf{y}}(\mathbf{x},\mathbf{y})\right\|_{\mathbf{y}'}^2 &\overset{(i)}{\leq} 2\left[\sum_{j'} y'_{i'}\frac{c_{i'}^2}{q_{i'}} + \sum_{i,j} y'_i \frac{M_{ij}^2 x_j^2}{q_{ij}}\right]\\
&\overset{(ii)}{=} 2\left[\left(\sum_{i'} y'_{i'} c_{i'}\right)\|\mathbf{c}\|_1 + \left(\sum_{i,j} y'_i |M_{ij} x_j|\right)\left(\sum_{i,j}|M_{ij} x_j|\right)\right]\\
&\overset{(iii)}{\leq} 2\left[m\|\mathbf{c}\|_\infty^2 + mb^2\|\mathbf{M}\|_\infty^2\right],
\end{aligned}$$

where we use (i) the fact that $\|x+y\|^2 \leq 2\|x\|^2 + 2\|y\|^2$ and taking expectation, (ii) plugging in the explicit sampling probabilities as stated in (20), and (iii) Cauchy-Schwarz inequality and the fact that $\mathbf{y}' \in \Delta^m$, $\mathbf{c} \in \mathbb{R}^m$, $\mathbf{x} \in \mathbb{B}_b^n$. □

Proof of Corollary 1. In light of Theorem 3 with Lemma 12 and Lemma 13, whenever $\epsilon \in (0,1)$, $b\left\|\mathbf{M}\right\|_\infty + \left\|\mathbf{c}\right\|_\infty = \Omega(1)$, gradient estimators in (19) and (20) satisfy the desired conditions. As a result, one can pick

$$\eta^{\mathsf{x}} = \Theta\left(\frac{\epsilon}{\|\mathbf{b}\|_1^2 + \|\mathbf{M}\|_\infty^2}\right), \eta^{\mathsf{y}} = \Theta\left(\frac{\epsilon}{m\left(\|\mathbf{c}\|_\infty^2 + b^2\left\|\mathbf{M}\right\|_\infty^2\right)}\right),$$
$$T = O\left(\frac{(n + m\log m)b^2\left\|\mathbf{M}\right\|_\infty^2 + nb^2\|\mathbf{b}\|_1^2 + m\log m\left\|\mathbf{c}\right\|_\infty^2}{\epsilon^2}\right),$$

to get an expected ϵ-approximate solution to the general ℓ_∞-ℓ_1 bilinear saddle-point problem (7), proving the corollary. □

D Proofs for Section 4

D.1 Bounds on Matrix Norm

Here we provide proofs for Lemma 7 and Lemma 15 used in Section 4 and Section 5. These lemmas bound the ℓ_∞ norm of the matrix $(\mathbf{I} - \mathbf{P}^\pi + \mathbf{1}(\boldsymbol{\nu}^\pi)^\top)^\dagger$ given $\boldsymbol{\nu}^\pi$ as a stationary distribution for $\mathbf{P}^\pi$ for mixing-case, and $(\mathbf{I} - \gamma\mathbf{P}^\pi)^{-1}$ for discounted case and under any policy π. The analysis needed for these lemmas are standard; we give their proofs for completeness and to highlight the connection between these quantities and the structure of the MDP.

We first introduce Lemma 7 showing that the mixing assumption A naturally leads to ℓ_∞-norm bound on the interested matrix, which is useful in both in deciding M and in proving Lemma 6 in Section 5.

Lemma 7. *Given a mixing AMDP, policy π, and its probability transition matrix $\mathbf{P}^\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{S}}$ and stationary distribution $\boldsymbol{\nu}^\pi$, $\left\|(\mathbf{I} - \mathbf{P}^\pi + \mathbf{1}(\boldsymbol{\nu}^\pi)^\top)^{-1}\right\|_\infty \le 3t_{\mathrm{mix}}$.*

In order to prove Lemma 7, we will first give a helper lemma adapted from Cohen et al. (2016) capturing the property of $\mathbf{I} - \mathbf{P}^\pi + \boldsymbol{\nu}^\pi\mathbf{1}^\top$.

Lemma 14 (cf. Lemma 22-23 in Cohen et al. (2016), variant)**.** *For a probabilistic transition matrix $\mathbf{P}^\pi$ with mixing time t_{mix} as defined in Assumption A and stationary distribution $\boldsymbol{\nu}^\pi$, one has for all non-negative integer c,*

$$\left\|(\mathbf{P}^\pi)^{ct_{\mathrm{mix}}} - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\right\|_\infty = \left\|\left[(\mathbf{P}^\pi)^{t_{\mathrm{mix}}} - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\right]^c\right\|_\infty \le \frac{1}{2^c},$$

by observing the equivalence that

$$(\mathbf{P}^\pi)^{\alpha+\beta} - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top = \left((\mathbf{P}^\pi)^\alpha - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\right)\left((\mathbf{P}^\pi)^\beta - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\right), \textit{for all nonnegative integers } \alpha, \beta.$$

We use this lemma and additional algebraic properties involving operator norms and mixing time for the proof of Lemma 7, formally as follows.

Proof of Lemma 7. Denote $\hat{\mathbf{P}} := \mathbf{P}^\pi - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top$, we first show the following equality.

$$(\mathbf{I} - \hat{\mathbf{P}})^{-1} \overset{(i)}{=} \sum_{k=0}^{\infty}\sum_{t=kt_{\mathrm{mix}}+1}^{(k+1)t_{\mathrm{mix}}}\hat{\mathbf{P}}^t = \sum_{k=0}^{\infty}\sum_{t=kt_{\mathrm{mix}}+1}^{(k+1)t_{\mathrm{mix}}}\left((\mathbf{P}^\pi)^t - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\right), \tag{21}$$

To show the equality (i), observing that

$$\left\|(\mathbf{P}^\pi)^{ct_{\mathrm{mix}}+t} - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\right\|_\infty \overset{(i)}{\le} \frac{1}{2^c}\left\|(\mathbf{P}^\pi)^t - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\right\|_\infty \overset{(ii)}{\le} \frac{1}{2^{c-1}},$$

as $c, t \to \infty$, where we use (i) Lemma 14, and (ii) the fact that $\left\|(\mathbf{P}^\pi)^t - (\mathbf{1}\boldsymbol{\nu}^\pi)^\top\right\|_\infty \le 2$ for all $t \ge 0$.

This ensures the RHS of Eq. (21) exists, and also one can check that

$$(\mathbf{I} - \hat{\mathbf{P}})\left(\sum_{t=0}^{\infty}\hat{\mathbf{P}}^t\right) = (\mathbf{I} - \hat{\mathbf{P}})\left(\sum_{t=0}^{\infty}\hat{\mathbf{P}}^t\right) = \mathbf{I},$$

which indicates that equality (21) is valid.

Now by triangle inequality, definition of matrix norm and mixing time, one has

$$\begin{aligned}\left\|(\mathbf{I}-\hat{\mathbf{P}})^{-1}\right\|_\infty &:= \max_{\mathbf{v}\in\mathbb{R}^{\mathcal{S}}}\frac{\left\|(\mathbf{I}-\hat{\mathbf{P}})^{-1}\mathbf{v}\right\|_\infty}{\|\mathbf{v}\|_\infty}\\ &\le \sum_{k=0}^{\infty}\sum_{t=kt_{\mathrm{mix}}+1}^{(k+1)t_{\mathrm{mix}}}\left\|(\mathbf{P}^\pi)^t-\mathbf{1}\boldsymbol{\nu}^{\pi\top}\right\|_1\\ &\le t_{\mathrm{mix}}(1+2\sum_{k=1}^{\infty}(1/2)^k)=3t_{\mathrm{mix}},\end{aligned}$$

thus proving the claim. □

For discounted case, we can alternatively show an upper bound on matrix norm using discount factor γ, formally stated in Lemma 15, and used for bounding M for discounted case in Section 4 and proof of Lemma 9 in Section 5.

Lemma 15. *Given a DMDP with discount factor $\gamma\in(0,1)$, for any probability transition matrix $\mathbf{P}^\pi\in\mathbb{R}^{\mathcal{S}\times\mathcal{S}}$ under certain policy π, it holds that $(\mathbf{I}-\gamma\mathbf{P}^\pi)^{-1}$ is invertible with*

$$\|(\mathbf{I}-\gamma\mathbf{P}^\pi)^{-1}\|_\infty\le\frac{1}{1-\gamma}.$$

Proof of Lemma 15. First, we claim that

$$\min_{v\in\mathbb{R}^{\mathcal{S}}:\|v\|_\infty=1}\left\|(\mathbf{I}-\gamma\mathbf{P}^\pi)^{-1}v\right\|\ge 1-\gamma\ . \tag{22}$$

To see this, let $\mathbf{v}\in\mathbb{R}^{\mathcal{S}}$ with $\|\mathbf{v}\|_\infty=1$ be arbitrary and let $i\in\mathcal{S}$ be such that $|v_i|=1$. We have

$$\begin{aligned}|[(\mathbf{I}-\gamma\mathbf{P}^\pi)\mathbf{v}]_i| &= \left|v_i-\gamma\sum_{j\in\mathcal{S}}\mathbf{P}^\pi(i,j)v_j\right|\ge|v_i|-\left|\gamma\sum_{j\in\mathcal{S}}\mathbf{P}^\pi(i,j)v_j\right|\\ &\ge 1-\gamma\sum_{j\in\mathcal{S}}\mathbf{P}^\pi(i,j)|v_j|\ge 1-\gamma.\end{aligned}$$

Applying the claim yields the result as (22) implies invertibility of $\mathbf{I}-\gamma\mathbf{P}^\pi$ and

$$\begin{aligned}\left\|(\mathbf{I}-\gamma\mathbf{P}^\pi)^{-1}\right\|_\infty &:= \max_{\mathbf{v}\in\mathbb{R}^{\mathcal{S}}}\frac{\left\|(\mathbf{I}-\gamma\mathbf{P}^\pi)^{-1}\mathbf{v}\right\|_\infty}{\|\mathbf{v}\|_\infty}\\ &\overset{(i)}{=}\max_{\hat{\mathbf{v}}}\frac{\left\|(\mathbf{I}-\gamma\mathbf{P}^\pi)^{-1}(\mathbf{I}-\gamma\mathbf{P}^\pi)\hat{\mathbf{v}}\right\|_\infty}{\|(\mathbf{I}-\gamma\mathbf{P}^\pi)\hat{\mathbf{v}}\|_\infty}\\ &\overset{(ii)}{=}\max_{\hat{\mathbf{v}}:\|\hat{\mathbf{v}}\|_\infty=1}\frac{\left\|(\mathbf{I}-\gamma\mathbf{P}^\pi)^{-1}(\mathbf{I}-\gamma\mathbf{P}^\pi)\hat{\mathbf{v}}\right\|_\infty}{\|(\mathbf{I}-\gamma\mathbf{P}^\pi)\hat{\mathbf{v}}\|_\infty}\\ &=\frac{1}{\min_{\hat{\mathbf{v}}:\|\hat{\mathbf{v}}\|_\infty=1}\|(\mathbf{I}-\gamma\mathbf{P}^\pi)\hat{\mathbf{v}}\|_\infty},\end{aligned}$$

where in (i) we replaced $\mathbf{v}$ with $(\mathbf{I}-\gamma\mathbf{P}^\pi)\hat{\mathbf{v}}$ for some $\hat{\mathbf{v}}$ since $\mathbf{I}-\gamma\mathbf{P}^\pi$ is invertible and in (ii) we rescaled $\hat{\mathbf{v}}$ to satisfy $\|\hat{\mathbf{v}}\|_\infty=1$ as scaling $\hat{\mathbf{v}}$ does not affect the ratio so long as $\hat{\mathbf{v}}\neq 0$. □

D.2 Additional proofs for Section 4.1

Proof of Lemma 1. For unbiasedness, direct computation reveals that

$$\mathbb{E}\left[\tilde{g}^{\mathsf{v}}(\mathbf{v},\boldsymbol{\mu})\right]=\sum_{i,a_i,j}\mu_{i,a_i}p_{ij}(a_i)(\mathbf{e}_j-\mathbf{e}_i)=\boldsymbol{\mu}^\top(\mathbf{P}-\hat{\mathbf{I}}).$$

For bound on second-moment, note $\|\tilde{g}^{\mathsf{v}}(\mathbf{v},\boldsymbol{\mu})\|_2^2\le 2$ with probability 1 by definition, the result follows immediately. □

Proof of Lemma 2. For unbiasedness, direct computation reveals that

$$\mathbb{E}\left[\tilde{g}^{\mu}(\mathbf{v},\boldsymbol{\mu})\right] = \sum_{i,a_i,j} p_{ij}(a_i)(v_i - v_j - r_{i,a_i})\mathbf{e}_{i,a_i} = (\hat{\mathbf{I}} - \mathbf{P})\mathbf{v} - \mathbf{r}\;.$$

For bound on ℓ_∞ norm, note that with probability 1 we have $\|\tilde{g}^{\mu}(\mathbf{v},\boldsymbol{\mu})\|_\infty \le (2M+1)\mathrm{A_{tot}}$ given $|v_i - v_j - r_{i,a_i}| \le \max\{2M, 2M+1\} \le 2M+1$ by domain bounds on $\mathbf{v}$. For bound on second-moment, given any $\boldsymbol{\mu}' \in \Delta^{\mathcal{A}}$ we have

$$\mathbb{E}[\|\tilde{g}^{\mu}(\mathbf{v},\boldsymbol{\mu})\|_{\boldsymbol{\mu}'}^2] \le \sum_{i,a_i} \frac{1}{\mathrm{A_{tot}}}\mu'_{i,a_i}\left\{(2M)^2,(2M+1)^2\right\}\mathrm{A_{tot}}^2 \le 9(M^2+1)\mathrm{A_{tot}},$$

where the first inequality follows similarly from $|v_i - v_j - r_{i,a_i}| \le \max\{2M, 2M+1\}, \forall i,j,a_i$. □

D.3 Additional proofs for Section 4.2

Proof of Lemma 3. For unbiasedness, one compute directly that

$$\mathbb{E}\left[\tilde{g}^{\mathsf{v}}(\mathbf{v},\boldsymbol{\mu})\right] = (1-\gamma)\mathbf{q} + \sum_{i,a_i,j}\mu_{i,a_i}p_{ij}(a_i)(\gamma\mathbf{e}_j - \mathbf{e}_i) = (1-\gamma)\mathbf{q} + \boldsymbol{\mu}^\top(\gamma\mathbf{P} - \hat{\mathbf{I}}).$$

For bound on second-moment, note $\|\tilde{g}^{\mathsf{v}}(\mathbf{v},\boldsymbol{\mu})\|_2^2 \le 2$ with probability 1 by definition and the fact that $\mathbf{q} \in \Delta^{\mathcal{S}}$, the result follows immediately. □

Proof of Lemma 4. For unbiasedness, one compute directly that

$$\mathbb{E}\left[\tilde{g}^{\mu}(\mathbf{v},\boldsymbol{\mu})\right] = \sum_{i,a_i}\sum_{j} p_{ij}(a_i)(v_i - \gamma v_j - r_{i,a_i})\mathbf{e}_{i,a_i} = (\hat{\mathbf{I}} - \gamma\mathbf{P})\mathbf{v} - \mathbf{r}.$$

For bound on ℓ_∞ norm, note that with probability 1 we have $\|\tilde{g}^{\mu}(\mathbf{v},\boldsymbol{\mu})\|_\infty \le (2M+1)\mathrm{A_{tot}}$ given $|v_i - \gamma\cdot v_j - r_{i,a_i}| \le \max\{2M, \gamma\cdot 2M+1\} \le 2M+1$ by domain bounds on $\mathbf{v}$. For bound on second-moment, for any $\boldsymbol{\mu}' \in \Delta^{\mathcal{A}}$ we have

$$\mathbb{E}[\|\tilde{g}^{\mu}(\mathbf{v},\boldsymbol{\mu})\|_{\boldsymbol{\mu}'}^2] \le \sum_{i,a_i} \frac{1}{\mathrm{A_{tot}}}\mu'_{i,a_i}\left\{(2M)^2,(2M+1)^2\right\}\mathrm{A_{tot}}^2 \le 9(M^2+1)\mathrm{A_{tot}},$$

where the first inequality follows by directly bounding $|v_i - \gamma v_j - r_{i,a_i}| \le \max\{2M, \gamma\cdot 2M+1\}, \forall i,j,a_i$. □

E Proofs for Section 5

E.1 Additional Proofs for Section 5.1

Proof of Lemma 5. Note by definition

$$\epsilon \ge \mathbb{E}\mathrm{Gap}(\mathbf{v}^\epsilon,\boldsymbol{\mu}^\epsilon) := \mathbb{E}\max_{\hat{\mathbf{v}}\in\mathbb{B}^{\mathcal{S}}_{2M},\hat{\boldsymbol{\mu}}\in\Delta^{\mathrm{A_{tot}}}}\left[(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}^\epsilon)^\top((\mathbf{P}-\mathbf{I})\mathbf{v}^\epsilon + \mathbf{r}) + {\boldsymbol{\mu}^\epsilon}^\top(\mathbf{P}-\mathbf{I})(\mathbf{v}^\epsilon - \hat{\mathbf{v}})\right].$$

When picking $\hat{\mathbf{v}} = \mathbf{v}^*$ and $\hat{\boldsymbol{\mu}} = \boldsymbol{\mu}^*$, i.e. optimizers of the minimax problem, this inequality yields

$$\begin{aligned}
\epsilon &\ge \mathbb{E}\left[(\boldsymbol{\mu}^* - \boldsymbol{\mu}^\epsilon)^\top((\mathbf{P}-\hat{\mathbf{I}})\mathbf{v}^\epsilon + \mathbf{r}) + {\boldsymbol{\mu}^\epsilon}^\top(\mathbf{P}-\hat{\mathbf{I}})(\mathbf{v}^\epsilon - \mathbf{v}^*)\right]\\
&= \mathbb{E}\left[{\boldsymbol{\mu}^*}^\top((\mathbf{P}-\hat{\mathbf{I}})\mathbf{v}^\epsilon + \mathbf{r}) - {\boldsymbol{\mu}^\epsilon}^\top\mathbf{r} - {\boldsymbol{\mu}^\epsilon}^\top(\mathbf{P}-\hat{\mathbf{I}})\mathbf{v}^*\right]\\
&\overset{(i)}{=} \mathbb{E}\left[{\boldsymbol{\mu}^\epsilon}^\top\left((\hat{\mathbf{I}}-\mathbf{P})\mathbf{v}^* - \mathbf{r}\right)\right] + {\boldsymbol{\mu}^*}^\top\mathbf{r}\\
&\overset{(ii)}{=} \mathbb{E}\left[{\boldsymbol{\mu}^\epsilon}^\top\left((\hat{\mathbf{I}}-\mathbf{P})\mathbf{v}^*\right] - \mathbf{r}\right) + \bar{v}^*,
\end{aligned}$$

where we use (i) the fact that ${\boldsymbol{\mu}^*}^\top(\mathbf{P}-\hat{\mathbf{I}}) = 0$ by duality feasibility and (ii) $\bar{v}^* := {\boldsymbol{\mu}^*}^\top\mathbf{r}$ by strong duality of (P) and (D) in (3). □

Proof of Lemma 6. Say $(\mathbf{v}^\epsilon, \boldsymbol{\mu}^\epsilon)$ is an ϵ-optimal solution in the form $\mu^\epsilon_{i,a_i} = \lambda^\epsilon_i \pi^\epsilon_{i,a_i}$, for some $\boldsymbol{\lambda}^\epsilon, \pi^\epsilon$, we still denote the induced policy as π and correspondingly probability transition matrix $\mathbf{P}^\pi$ and expected reward vector $\mathbf{r}^\pi$.

Given the optimality condition, we have

$$\mathbb{E}\left[f(\mathbf{v}^*, \boldsymbol{\mu}^\epsilon) - \min_{\mathbf{v}\in\mathbb{B}^{\mathcal{S}}_{2M}} f(\mathbf{v}, \boldsymbol{\mu}^\epsilon)\right] \le \epsilon,$$

which is also equivalent to

$$\mathbb{E}\left[\max_{\mathbf{v}\in\mathbb{B}^{\mathcal{S}}_{2M}} \boldsymbol{\lambda}^{\epsilon\top}(\mathbf{P}^\pi - \mathbf{I})(\mathbf{v}^* - \mathbf{v})\right] \le \epsilon.$$

We first show $\|\mathbf{v}^*\|_\infty \le M = 5t_{\mathrm{mix}}/3$ which follows from

$$(\mathbf{I} - \mathbf{P}^* + \mathbf{1}(\boldsymbol{\nu}^*)^\top)\mathbf{v}^* = \mathbf{r}^* - \bar{v}^*\mathbf{1}$$

which gives

$$\|\mathbf{v}^*\|_\infty = \left\|(\mathbf{I} - \mathbf{P}^* + \mathbf{1}(\boldsymbol{\nu}^*)^\top)^{-1}(\mathbf{r}^* - \bar{v})\right\|_\infty \le \left\|(\mathbf{I} - \mathbf{P}^* + \mathbf{1}(\boldsymbol{\nu}^*)^\top)^{-1}\right\|_\infty \|\mathbf{r}^* - \bar{v}\|_\infty \le M$$

by definition of M, optimality conditions $(\mathbf{I} - \mathbf{P}^* + \mathbf{1}(\boldsymbol{\nu}^*)^\top)\mathbf{v}^* = \mathbf{r}^* - \bar{v}^*\mathbf{1}$, $\langle \boldsymbol{\nu}^*, \mathbf{v}^*\rangle = 0$, and Lemma 7.

Now notice $\mathbf{v} \in \mathbb{B}^{\mathcal{S}}_{2M}$, we get $\mathbb{E}\|\boldsymbol{\lambda}^{\epsilon\top}(\mathbf{P}^\pi - \mathbf{I})\|_1 \le \frac{1}{M}\epsilon$ by noticing

$$\begin{aligned}
2M\cdot\mathbb{E}\left\|\boldsymbol{\lambda}^{\epsilon\top}(\mathbf{P}^\pi - \mathbf{I})\right\|_1 &= \mathbb{E}\left[\max_{v\in\mathbb{B}^{\mathcal{S}}_{2M}} \boldsymbol{\lambda}^{\epsilon\top}(\mathbf{P}^\pi - \mathbf{I})(-\mathbf{v})\right]\\
&= \mathbb{E}\left[\max_{v\in\mathbb{B}^{\mathcal{S}}_{2M}} \boldsymbol{\lambda}^{\epsilon\top}(\mathbf{P}^\pi - \mathbf{I})(\mathbf{v}^* - \mathbf{v}) - \boldsymbol{\lambda}^{\epsilon\top}(\mathbf{P}^\pi - \mathbf{I})\mathbf{v}^*\right]\\
&\le \epsilon + \mathbb{E}\left\|\boldsymbol{\lambda}^{\epsilon\top}(\mathbf{P}^\pi - \mathbf{I})\right\|_1 \|v^*\|_\infty \le \epsilon + M\cdot\mathbb{E}\left\|\boldsymbol{\lambda}^{\epsilon\top}(\mathbf{P}^\pi - \mathbf{I})\right\|_1.
\end{aligned}$$

This is the part of analysis where expanding the domain size of $\mathbf{v}$ from M to $2M$ will be helpful.

Now suppose that $\boldsymbol{\nu}^\pi$ is the stationary distribution under policy $\pi := \pi^\epsilon$. By definition, this implies

$$\boldsymbol{\nu}^{\pi\top}(\mathbf{P}^\pi - \mathbf{I}) = 0.$$

Therefore, combining this fact with $\mathbb{E}\|\boldsymbol{\lambda}^{\epsilon\top}(\mathbf{P}^\pi - \mathbf{I})\|_1 \le \frac{1}{M}\epsilon$ as we have shown earlier yields

$$\mathbb{E}\left\|(\boldsymbol{\lambda}^\epsilon - \boldsymbol{\nu}^\pi)^\top(\mathbf{P}^\pi - \mathbf{I})\right\|_1 \le \frac{1}{M}\epsilon.$$

It also leads to

$$\begin{aligned}
\mathbb{E}\left[(\boldsymbol{\nu}^\pi - \boldsymbol{\lambda}^\epsilon)^\top\mathbf{r}^\pi\right] &= \mathbb{E}\left[(\boldsymbol{\nu}^\pi - \boldsymbol{\lambda}^\epsilon)^\top(\mathbf{r}^\pi - (\langle\mathbf{r}^\pi, \boldsymbol{\nu}^\pi\rangle)\mathbf{1})\right]\\
&= \mathbb{E}\left[(\boldsymbol{\nu}^\pi - \boldsymbol{\lambda}^\epsilon)^\top\left(\mathbf{I} - \mathbf{P}^\pi + \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\right)\left(\mathbf{I} - \mathbf{P}^\pi + \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\right)^{-1}(\mathbf{r}^\pi - (\langle\mathbf{r}^\pi, \boldsymbol{\nu}^\pi\rangle)\mathbf{1})\right]\\
&\le \mathbb{E}\left\|(\boldsymbol{\nu}^\pi - \boldsymbol{\lambda}^\epsilon)^\top\left(\mathbf{I} - \mathbf{P}^\pi + \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\right)\right\|_1\left\|\left(\mathbf{I} - \mathbf{P}^\pi + \mathbf{1}(\boldsymbol{\nu}^\pi)^\top\right)^{-1}(\mathbf{r}^\pi - (\langle\mathbf{r}^\pi, \boldsymbol{\nu}^\pi\rangle)\mathbf{1})\right\|_\infty\\
&\le M\cdot\mathbb{E}\left\|(\boldsymbol{\nu}^\pi - \boldsymbol{\lambda}^\epsilon)^\top(\mathbf{I} - \mathbf{P}^\pi)\right\|_1 \le \epsilon,
\end{aligned}$$

where for the last but one inequality we use the definition of $M = 5t_{\mathrm{mix}}/3$ and Lemma 7.

Note now the average reward under policy π satisfies

$$\begin{aligned}\mathbb{E}\bar{v}^\pi =&\mathbb{E}\left[(\boldsymbol{\nu}^\pi)^\top\mathbf{r}^\pi\right] = \mathbb{E}\left[\boldsymbol{\nu}^{\pi\top}(\mathbf{P}^\pi-\mathbf{I})\mathbf{v}^* + (\boldsymbol{\nu}^\pi)^\top\mathbf{r}^\pi\right]\\ =&\mathbb{E}\left[(\boldsymbol{\nu}^\pi-\boldsymbol{\lambda}^\epsilon)^\top\left[(\mathbf{P}^\pi-\mathbf{I})\mathbf{v}^*+\mathbf{r}^\pi\right]\right] + \mathbb{E}\left[\boldsymbol{\lambda}^{\epsilon\top}[(\mathbf{P}^\pi-\mathbf{I})\mathbf{v}^*+\mathbf{r}^\pi]\right]\\ \overset{(i)}{\ge}&\mathbb{E}\left[(\boldsymbol{\nu}^\pi-\boldsymbol{\lambda}^\epsilon)^\top(\mathbf{P}^\pi-\mathbf{I})\mathbf{v}^*\right] + \mathbb{E}\left[(\boldsymbol{\nu}^\pi-\boldsymbol{\lambda}^\epsilon)^\top\mathbf{r}^\pi\right] + \bar{v}^* - \epsilon\\ \overset{(ii)}{\ge}& \bar{v}^* - \mathbb{E}\|(\boldsymbol{\nu}^\pi-\boldsymbol{\lambda}^\epsilon)^\top(\mathbf{P}^\pi-\mathbf{I})\|_1\left\|\mathbf{v}^*\right\|_\infty - \mathbb{E}\left[(\boldsymbol{\nu}^\pi-\boldsymbol{\lambda}^\epsilon)^\top\mathbf{r}^\pi\right]-\epsilon\\ \overset{(iii)}{\ge}& \bar{v}^* - \frac{1}{M}\epsilon\cdot M - (\epsilon\cdot 1) - \epsilon = \bar{v}^* - 3\epsilon\end{aligned}$$

where we use (i) the optimality relation stated in Lemma 5, (ii) Cauchy-Schwarz inequality and (iii) conditions on ℓ_1 bounds of $(\boldsymbol{\lambda}^\epsilon-\boldsymbol{\nu}^\pi)^\top(\mathbf{P}^\pi-\mathbf{I})$ and $(\boldsymbol{\lambda}^\epsilon-\boldsymbol{\nu}^\pi)^\top\mathbf{r}^\pi$ we prove earlier. □

Proof of Theorem 1. Given mixing MDP tuple $\mathcal{M}=(\mathcal{S},\mathcal{A},\mathcal{P},\mathbf{r})$, let $\epsilon\in(0,1)$, one can construct an approximate policy π^ϵ using Algorithm 1 with accuracy level set to $\epsilon'=\frac{1}{3}\epsilon$ such that by Lemma 6,

$$\mathbb{E}\bar{v}^{\pi^\epsilon} \ge \bar{v}^* - \epsilon.$$

It follows from Corollary 4 straightforwardly to see the sample complexity is bounded by

$$O\left(\frac{t_{\mathrm{mix}}^2\mathrm{A}_{\mathrm{tot}}\log(\mathrm{A}_{\mathrm{tot}})}{\epsilon^2}\right).$$

□

E.2 Additional Proofs for Section 5.2

Proof of Lemma 8. Note by definition

$$\epsilon \ge \mathbb{E}\mathrm{Gap}(\mathbf{v}^\epsilon,\boldsymbol{\mu}^\epsilon) \coloneqq \mathbb{E}\max_{\hat{\mathbf{v}},\hat{\boldsymbol{\mu}}}\left[(1-\gamma)\mathbf{q}^\top\mathbf{v}^\epsilon + \hat{\boldsymbol{\mu}}^\top((\gamma\mathbf{P}-\hat{\mathbf{I}})\mathbf{v}^\epsilon+\mathbf{r}) - (1-\gamma)\mathbf{q}^\top\hat{\mathbf{v}} - \boldsymbol{\mu}^{\epsilon\top}((\gamma\mathbf{P}-\hat{\mathbf{I}})\hat{\mathbf{v}}+\mathbf{r})\right].$$

When picking $\hat{\mathbf{v}}=\mathbf{v}^*,\hat{\boldsymbol{\mu}}=\boldsymbol{\mu}^*$ optimizers of the minimax problem, this inequality becomes

$$\begin{aligned}\epsilon\ge&\,\mathbb{E}\left[(1-\gamma)\mathbf{q}^\top\mathbf{v}^\epsilon + \boldsymbol{\mu}^{*\top}((\gamma\mathbf{P}-\hat{\mathbf{I}})\mathbf{v}^\epsilon+\mathbf{r}) - (1-\gamma)\mathbf{q}^\top\mathbf{v}^* - \boldsymbol{\mu}^{\epsilon\top}((\gamma\mathbf{P}-\hat{\mathbf{I}})\mathbf{v}^*+\mathbf{r})\right]\\ \overset{(i)}{=}&\,\boldsymbol{\mu}^{*\top}\mathbf{r} - (1-\gamma)\mathbf{q}^\top\mathbf{v}^* - \mathbb{E}\left[\boldsymbol{\mu}^{\epsilon\top}((\gamma\mathbf{P}-\hat{\mathbf{I}})\mathbf{v}^*+\mathbf{r})\right]\\ \overset{(ii)}{=}&\,\mathbb{E}\left[\boldsymbol{\mu}^{\epsilon\top}\left((\hat{\mathbf{I}}-\gamma\mathbf{P})\mathbf{v}^*-\mathbf{r}\right)\right],\end{aligned}$$

where we use (i) the fact that $\boldsymbol{\mu}^{*\top}(\mathbf{I}-\gamma\mathbf{P})=(1-\gamma)\mathbf{q}^\top$ by dual feasibility and (ii) $(1-\gamma)\mathbf{q}^\top\mathbf{v}^*=\boldsymbol{\mu}^{*\top}\mathbf{r}$ by strong duality theory of linear programming. □

Proof of Lemma 9. Without loss of generality we reparametrize $(\mathbf{v}^\epsilon,\boldsymbol{\mu}^\epsilon)$ as an ϵ-optimal solution in the form $\mu^\epsilon_{i,a_i}=\lambda^\epsilon_i\pi^\epsilon_{i,a_i}$, for some $\boldsymbol{\lambda}^\epsilon,\pi^\epsilon$. For simplicity we still denote the induced policy as π and correspondingly probability transition matrix $\mathbf{P}^\pi$ and expected $\mathbf{r}^\pi$.

Given the optimality condition, we have

$$\mathbb{E}\left[f(\mathbf{v}^*,\boldsymbol{\mu}^\epsilon) - \min_{\mathbf{v}\in\mathbb{B}^{\mathcal{S}}_{2M}} f(\mathbf{v},\boldsymbol{\mu}^\epsilon)\right]\le\epsilon,$$

which is also equivalent to

$$\mathbb{E}\max_{\mathbf{v}\in\mathbb{B}^{\mathcal{S}}_{2M}}\left[(1-\gamma)\mathbf{q}^\top+\boldsymbol{\lambda}^{\epsilon\top}(\gamma\mathbf{P}^\pi-\mathbf{I})\right](\mathbf{v}^*-\mathbf{v})\le\epsilon.$$

By optimality conditions and Lemma 15 one has $\|\mathbf{v}^*\|_\infty = \left\|(\mathbf{I}-\gamma\mathbf{P}^*)^{-1}\mathbf{r}^*\right\|_\infty \le M\cdot 1 = M = (1-\gamma^{-1})$. Now notice $\mathbf{v}\in\mathbb{B}^{\mathcal{S}}_{2M}$, we have $\|(1-\gamma)\mathbf{q}+\boldsymbol{\lambda}^{\epsilon\top}(\gamma\mathbf{P}^\pi-\mathbf{I})\|_1 \le \frac{\epsilon}{M}$ as a consequence of

$$
\begin{aligned}
& 2M\cdot\mathbb{E}\left\|(1-\gamma)\mathbf{q}+\boldsymbol{\lambda}^{\epsilon\top}(\gamma\mathbf{P}^\pi-\mathbf{I})\right\|_1\\
=&\mathbb{E}\left[\max_{v\in\mathbb{B}^{\mathcal{S}}_{2M}}\left[(1-\gamma)\mathbf{q}+\boldsymbol{\lambda}^{\epsilon\top}(\gamma\mathbf{P}^\pi-\mathbf{I})\right](-\mathbf{v})\right]\\
=&\mathbb{E}\left[\max_{v\in\mathbb{B}^{\mathcal{S}}_{2M}}\left[(1-\gamma)\mathbf{q}+\boldsymbol{\lambda}^{\epsilon\top}(\gamma\mathbf{P}^\pi-\mathbf{I})\right](\mathbf{v}^*-\mathbf{v})-\left[(1-\gamma)\mathbf{q}+\boldsymbol{\lambda}^{\epsilon\top}(\gamma\mathbf{P}^\pi-\mathbf{I})\right]\mathbf{v}^*\right]\\
\le&\epsilon+\mathbb{E}\left\|(1-\gamma)\mathbf{q}+\boldsymbol{\lambda}^{\epsilon\top}(\gamma\mathbf{P}^\pi-\mathbf{I})\right\|_1\|\mathbf{v}^*\|_\infty\le\epsilon+M\cdot\mathbb{E}\left\|(1-\gamma)\mathbf{q}+\boldsymbol{\lambda}^{\epsilon\top}(\gamma\mathbf{P}^\pi-\mathbf{I})\right\|_1.
\end{aligned}
$$

Now by definition of $\boldsymbol{\nu}^\pi$ as the dual feasible solution under policy $\pi := \pi^\epsilon$,

$$
\mathbb{E}\left[(1-\gamma)\mathbf{q}^\top+\boldsymbol{\nu}^{\pi\top}(\gamma\mathbf{P}^\pi-\mathbf{I})\right]=0.
$$

Combining the two this gives

$$
\mathbb{E}\left\|(\boldsymbol{\lambda}^\epsilon-\boldsymbol{\nu}^\pi)^\top(\gamma\mathbf{P}^\pi-\mathbf{I})\right\|_1\le\frac{\epsilon}{M},
$$

and consequently

$$
\begin{aligned}
\mathbb{E}\left\|\boldsymbol{\lambda}^\epsilon-\boldsymbol{\nu}^\pi\right\|_1 &= \mathbb{E}\left\|(\gamma\mathbf{P}^\pi-\mathbf{I})^{-\top}(\gamma\mathbf{P}-\mathbf{I})^\top(\boldsymbol{\lambda}^\epsilon-\boldsymbol{\nu}^\pi)\right\|_1\\
&\le\mathbb{E}\left\|(\gamma\mathbf{P}^\pi-\mathbf{I})^{-\top}\right\|_1\left\|(\gamma\mathbf{P}-\mathbf{I})^\top(\boldsymbol{\lambda}^\epsilon-\boldsymbol{\nu}^\pi)\right\|_1\le\frac{M}{M}\epsilon=\epsilon,
\end{aligned}
$$

where the last but one inequality follows from the norm equality that $\left\|(\gamma\mathbf{P}^\pi-\mathbf{I})^{-\top}\right\|_1=\left\|(\gamma\mathbf{P}^\pi-\mathbf{I})^{-1}\right\|_\infty$ and Lemma 15. Note now the discounted reward under policy π satisfies

$$
\begin{aligned}
\mathbb{E}(1-\gamma)\bar{v}^\pi=\mathbb{E}(\boldsymbol{\nu}^\pi)^\top\mathbf{r}^\pi=&\mathbb{E}\left[(1-\gamma)\mathbf{q}^\top+\boldsymbol{\nu}^{\pi\top}(\gamma\mathbf{P}^\pi-\mathbf{I})\right]\mathbf{v}^*+\mathbb{E}(\boldsymbol{\nu}^\pi)^\top\mathbf{r}^\pi\\
=&(1-\gamma)\mathbf{q}^\top\mathbf{v}^*+\mathbb{E}\left[\boldsymbol{\nu}^{\pi\top}\left[(\gamma\mathbf{P}^\pi-\mathbf{I})\mathbf{v}^*+\mathbf{r}^\pi\right]\right]\\
=&(1-\gamma)\mathbf{q}^\top\mathbf{v}^*+\mathbb{E}\left[(\boldsymbol{\nu}^\pi-\boldsymbol{\lambda}^\epsilon)^\top\left[(\gamma\mathbf{P}^\pi-\mathbf{I})\mathbf{v}^*+\mathbf{r}^\pi\right]\right]+\mathbb{E}\left[\boldsymbol{\lambda}^{\epsilon\top}\left[(\gamma\mathbf{P}^\pi-\mathbf{I})\mathbf{v}^*+\mathbf{r}^\pi\right]\right]\\
\overset{(i)}{\ge}&(1-\gamma)\mathbf{q}^\top\mathbf{v}^*+\mathbb{E}\left[(\boldsymbol{\nu}^\pi-\boldsymbol{\lambda}^\epsilon)^\top\left[(\gamma\mathbf{P}^\pi-\mathbf{I})\mathbf{v}^*+\mathbf{r}^\pi\right]\right]-\epsilon\\
\overset{(ii)}{\ge}&(1-\gamma)\bar{v}^*-\mathbb{E}\|(\boldsymbol{\nu}^\pi-\boldsymbol{\lambda}^\epsilon)^\top(\gamma\mathbf{P}^\pi-\mathbf{I})\|_1\|\mathbf{v}^*\|_\infty-\mathbb{E}\|\boldsymbol{\nu}^\pi-\boldsymbol{\lambda}^\epsilon\|_1\|\mathbf{r}^\pi\|_\infty-\epsilon\\
\overset{(iii)}{\ge}&(1-\gamma)\bar{v}^*-\frac{1}{M}\epsilon\cdot M-\frac{M}{M}\epsilon\cdot 1-\epsilon=(1-\gamma)\bar{v}^*-3\epsilon,
\end{aligned}
$$

where we use (i) the optimality relation stated in Lemma 8, (ii) Cauchy-Schwarz inequality and (iii) conditions on ℓ_1 bounds of $(\boldsymbol{\lambda}^\epsilon-\boldsymbol{\nu}^\pi)^\top(\gamma\mathbf{P}^\pi-\mathbf{I})$ and $\boldsymbol{\lambda}^\epsilon-\boldsymbol{\nu}^\pi$ we prove earlier. $\square$

F Detailed Discussions on Open Problems

This work reveals an interesting connection between structure of MDP problems with ℓ_∞ regression. Here we point out a few interesting directions and open problems and elaborate them a bit for future work:

Primal-dual methods with optimal sample-complexity for DMDPs: For DMDPs, the sample complexity of our method has $(1-\gamma)^{-1}$ gap with the known lower bound (Azar et al., 2012), which can be achieved by stochastic value-iteration (Sidford et al., 2018a) or Q-learning (Wainwright, 2019). If it is achievable using convex-optimization lies at the core of further understanding the power (or hardness) of convex optimization methods relative to standard value / policy-iteration methods.

High-precision methods. There have been recent high-precision stochastic value-iteration algorithms (Sidford et al., 2018b) that produce an ϵ-optimal strategy in runtime $\widetilde{O}(|\mathcal{S}|\mathrm{A}_{\mathrm{tot}} + (1-\gamma)^{-3}\mathrm{A}_{\mathrm{tot}})$ while only depend logarithmically on the accuracy paramter $1/\epsilon$. However the result would strictly shrinks the value domain in an ℓ_∞ ball and it is open to generalize our method to obtain this.

Lower bound for average-reward MDP. There has been established lower-bound on sample complexity needed for discounted MDP (Azar et al., 2012), however the lower bound for average-reward MDP is less understood. For mixing MDP, we ask the question of what the best possible sample complexity dependence on mixing time is, and what the hard cases are. For more general average-reward MDP, we ask if there is any lower-bound result depending on problem parameters other than mixing time bound.

Extension to more general classes of MDP. While average-reward MDP with bounded mixing time t_{mix} and discounted MDP with discount factor γ are two fundamentally important classes of MDP, there are many MDP instances that fall beyond the range. It is thus an interesting open direction to extend our framework for more general MDP instances and understand what problem parameters the sample complexity of SMD-like methods should depend on.

G More Details for Section 6

Much of the approach and proofs follow similarly as for the mixing AMDP (unconstrained) case. Therefore, here we focus on the few places that one need to change a bit to cater to the constrained structure.

By considering (equivalently) the relaxation of (13) with $\boldsymbol{\mu} \geq \mathbf{0}$, $\|\boldsymbol{\mu}\|_1 \leq 1$ instead of $\boldsymbol{\mu} \in \Delta^{\mathcal{A}}$, one can obtain the following primal form of the problem:

$$\text{(P)} \quad \min_{\mathbf{s}\geq\mathbf{0},\mathbf{v},t\geq 0} \quad t - \sum_k s_k$$
$$\text{subject to} \quad (\mathbf{P}-\hat{\mathbf{I}})\mathbf{v}+\mathbf{D}\mathbf{s} \leq t\mathbf{1}.$$

Now by our assumptions, strong duality and strict complementary slackness there exists some optimal $t^* > 0$. Thus we can safely consider the case when $t > 0$, and rescale all variables $\mathbf{s}$, $\mathbf{v}$, and t by $1/t$ without changing that optimal solution with $t^* > 0$ to obtain the following equivalent primal form of the problem:

$$\text{(P')} \quad \min_{\mathbf{s}\geq\mathbf{0},\mathbf{v}} \quad 1 - \sum_k s_k$$
$$\text{subject to} \quad (\mathbf{P}-\hat{\mathbf{I}})\mathbf{v}+\mathbf{D}\mathbf{s} \leq \mathbf{1}.$$

Discussion on constraints of minimax formulation (15).
For $\mathbf{s}^*$, due to the feasibility assumption and strong duality theory, we know the optimality must achieve when $1-\sum_k s_k^* = 0$, i.e. one can safely consider a domain of $\mathbf{s}$ as $\sum_k s_k \leq 2, \mathbf{s} \geq \mathbf{0}$. For the bound on $\mathbf{v}^*$, using a method similar as in Section 5 we know there exists some $\mathbf{v}^*$, optimal policy π, its corresponding stationary distribution $\boldsymbol{\nu}^\pi$ and probability transition matrix $\mathbf{P}^\pi$ satisfying

$$(\mathbf{P}^\pi - \hat{\mathbf{I}})\mathbf{v}^* + \mathbf{D}\mathbf{s}^* = \mathbf{r}^* \leq \mathbf{1},$$

which implies that

$$\exists \mathbf{v}^* \perp \boldsymbol{\nu}^\pi, \|\mathbf{v}^*\|_\infty = \left\|(\mathbf{I} - \mathbf{P}^\pi + \mathbf{1}(\boldsymbol{\nu}^\pi)^\top)^{-1}(\mathbf{D}\mathbf{s}^* - \mathbf{r}^*)\right\|_\infty \leq 3Dt_{\mathrm{mix}},$$

where the last inequality follows from Lemma 7.

Divergence terms and gradient estimators.
We set Bregman divergence unchanged with respect to $\boldsymbol{\mu}$ and $\mathbf{v}$, for $\mathbf{s}$, we consider a standard distance generating function for ℓ_1 setup defined as $r(\mathbf{s}) \coloneqq \sum_k s_k \log(s_k)$, note it induces a rescaled KL-divergence as $V_{\mathbf{s}}(\mathbf{s}') \coloneqq \sum_k s_k \log(s'_k/s'_k) - \|\mathbf{s}\|_1 + \|\mathbf{s}'\|_1$, which also satisfies the local-norm property that

$$\forall \mathbf{s}', \mathbf{s} \geq \mathbf{0}, \mathbf{1}^\top\mathbf{s} \leq 2, \mathbf{1}^\top\mathbf{s}' \leq 2, k \geq 6, \|\delta\|_\infty \leq 1; \langle \delta, \mathbf{s}' - \mathbf{s}\rangle - V_{\mathbf{s}'}(\mathbf{s}) \leq \sum_{k\in[K]} s_k\delta_k^2.$$

Now for the primal side, the gradient mapping is $g^{\mathsf{v}}(\mathbf{v},\mathbf{s},\boldsymbol{\mu}) = (\hat{\mathbf{I}}-\mathbf{P})^\top\boldsymbol{\mu}$, $g^{\mathsf{s}}(\mathbf{v},\mathbf{s},\boldsymbol{\mu}) = \mathbf{D}^\top\boldsymbol{\mu}-\mathbf{1}$, we can define gradient estimators correspondingly as

$$\begin{aligned}&\text{Sample } (i,a_i)\sim[\boldsymbol{\mu}]_{i,a_i}, j\sim p_{ij}(a_i), &&\text{set } \tilde{g}^{\mathsf{v}}(\mathbf{v},\mathbf{s},\boldsymbol{\mu}) = \mathbf{e}_j-\mathbf{e}_i.\\ &\text{Sample } (i,a_i)\sim[\boldsymbol{\mu}]_{i,a_i}, k\sim 1/K, &&\text{set } \tilde{g}^{\mathsf{s}}(\mathbf{v},\mathbf{s},\boldsymbol{\mu}) = K[d_k]_{i,a_i}\mathbf{e}_k-\mathbf{1}.\end{aligned} \tag{23}$$

These are bounded gradient estimator for the primal side respectively.

Lemma 16. *$\tilde{g}^{\mathsf{v}}$ defined in (23) is a $(2,\|\cdot\|_2)$-bounded estimator, and $\tilde{g}^{\mathsf{s}}$ defined in (23) is a $(KD+2, 2KD^2+2, \|\cdot\|_{\Delta^K})$-bounded estimator.*

For the dual side, $g^{\mu}(\mathbf{v},\mathbf{s},\boldsymbol{\mu}) = (\hat{\mathbf{I}}-\mathbf{P})\mathbf{v}+\mathbf{D}\mathbf{s}$, with its gradient estimator

$$\begin{aligned}&\text{Sample } (i,a_i)\sim 1/\mathrm{A}_{\mathrm{tot}}, j\sim p_{ij}(a_i), k\sim s_k/\|\mathbf{s}\|_1 \quad ;\\ &\text{set } \tilde{g}^{\mu}(\mathbf{v},\mathbf{s},\boldsymbol{\mu}) = \mathrm{A}_{\mathrm{tot}}(v_i-\gamma v_j - r_{i,a_i}+[d_k]_{i,a_i}\|\mathbf{s}\|_1)\mathbf{e}_{i,a_i}.\end{aligned} \tag{24}$$

This is a bounded gradient estimator for the dual side with the following property.

Lemma 17. *$\tilde{g}^{\mu}$ defined in (24) is a $((2M+1+2D)\mathrm{A}_{\mathrm{tot}}, 2(2M+1+2D)^2\mathrm{A}_{\mathrm{tot}}, \|\cdot\|_{\Delta^{\mathcal{A}}})$-bounded estimator.*

Instead of directly using Theorem 3, for the primal updates we consider a slightly generalized variant where we take a standard gradient step at $\mathbf{v}$ using $\tilde{g}^{\mathsf{v}}$, and an entropic step at $\mathbf{s}$ using $\tilde{g}^{\mathsf{s}}$. The regret bound for the primal side following Lemma 10 and 11 becomes

$$\sum_{t\in[T]}\langle\tilde{g}^{\mathsf{v}}_t,\mathbf{v}_t-\mathbf{v}\rangle+\sum_{t\in[T]}\langle\tilde{g}^{\mathsf{s}}_t,\mathbf{s}_t-\mathbf{s}\rangle \le \frac{V_{\mathbf{v}_0}(\mathbf{v})}{\eta^{\mathsf{v}}}+\frac{\sum_{t=0}^{T}\eta^{\mathsf{v}}\|\tilde{g}^{\mathsf{v}}_t\|_2^2}{2}+\frac{V_{\mathbf{s}_0}(\mathbf{s})}{\eta^{\mathsf{s}}}+\frac{\sum_{t=0}^{T}\eta^{\mathsf{s}}\|\tilde{g}^{\mathsf{s}}_t\|_{\mathbf{s}_t}^2}{2},$$

together also with a similar ghost-iterate analysis for $\mathbf{s}$ part of the primal side.

Given the guarantees of designed gradient estimators in Lemma 16, 17 and choice of $M=3Dt_{\mathrm{mix}}$ this gives Corollary 4, and 5.

Proof of Corollary 5

Proof of Corollary 5. Following the similar rounding technique as in Section 5, one can consider the policy induced by the ϵ-approximate solution of MDP π^ϵ from the unique decomposition of $\mu^\epsilon_{i,a_i} = \lambda^\epsilon_i\cdot\pi^\epsilon_{i,a_i}$, for all $i\in\mathcal{S}, a_i\in\mathcal{A}_i$.

Given the optimality condition, we have

$$\mathbb{E}\left[f(\mathbf{v}^*,\mathbf{s}^\epsilon,\boldsymbol{\mu}^\epsilon)-\min_{\mathbf{v}\in\mathbb{B}^{\mathcal{S}}_{2M}}f(\mathbf{v},\mathbf{s}^\epsilon,\boldsymbol{\mu}^\epsilon)\right]\le\epsilon,$$

which is also equivalent to (denoting $\pi:=\pi^\epsilon$, and $\boldsymbol{\nu}^\pi$ as stationary distribution under it)

$$\mathbb{E}\left[\max_{\mathbf{v}\in\mathbb{B}^{\mathcal{S}}_{2M}}\boldsymbol{\lambda}^{\epsilon\top}(\mathbf{I}-\mathbf{P}^\pi)(\mathbf{v}^*-\mathbf{v})\right]\le\epsilon,$$

thus implying that $\|(\boldsymbol{\lambda}^\epsilon)^\top(\mathbf{I}-\mathbf{P}^\pi)\|_1\le\frac{1}{M}\epsilon$, $\|(\boldsymbol{\lambda}^\epsilon-\boldsymbol{\nu}^\pi)^\top(\mathbf{I}-\mathbf{P}^\pi-\mathbf{1}(\boldsymbol{\nu}^\pi)^\top)\|_1\le\frac{1}{M}\epsilon=O(\frac{1}{t_{\mathrm{mix}}D}\epsilon)$ hold in expectation.

Now consider $\boldsymbol{\mu}$ constructed from $\mu_{i,a_i}=\nu^\epsilon_i\cdot\pi^\epsilon_{i,a_i}$, by definition of $\boldsymbol{\nu}$ it holds that $\boldsymbol{\mu}(\hat{\mathbf{I}}-\mathbf{P})=0$.

For the second inequality of problem (13), similarly in light of primal-dual optimality

$$\mathbb{E}\left[f(\mathbf{v}^\epsilon,\mathbf{s}^*,\boldsymbol{\mu}^\epsilon)-\min_{\mathbf{s}\ge\mathbf{0}:\sum_k s_k\le 2}f(\mathbf{v}^\epsilon,\mathbf{s},\boldsymbol{\mu}^\epsilon)\right]\le\epsilon \quad\Leftrightarrow\quad \mathbb{E}\left[\max_{\mathbf{s}\ge\mathbf{0}:\sum_k s_k\le 2}\left[(\boldsymbol{\mu}^\epsilon)^\top\mathbf{D}-\mathbf{1}^\top\right](\mathbf{s}^*-\mathbf{s})\right]\le\epsilon,$$

which implies that $\mathbf{D}^\top \boldsymbol{\mu}^\epsilon \geq \mathbf{e} - \epsilon \mathbf{1}$ hold in expectation given $\mathbf{s}^* \in \Delta^K$.

Consequently, we can bound the quality of dual variable $\boldsymbol{\mu}$

$$\begin{aligned}
\mathbf{D}^\top \boldsymbol{\mu} &= \mathbf{D}^\top \boldsymbol{\mu}^\epsilon + \mathbf{D}^\top (\boldsymbol{\mu} - \boldsymbol{\mu}^\epsilon) = \mathbf{D}^\top \boldsymbol{\mu}^\epsilon + \mathbf{D}^\top \Pi^{\epsilon\top} (\boldsymbol{\nu}^\pi - \boldsymbol{\lambda}^\epsilon) \\
&= \mathbf{D}^\top \boldsymbol{\mu}^\epsilon + \mathbf{D}^\top \Pi^{\epsilon\top} (\mathbf{I} - (\mathbf{P}^\pi)^\top + \boldsymbol{\nu}^\pi \mathbf{1}^\top)^{-1} (\mathbf{I} - (\mathbf{P}^\pi)^\top + \boldsymbol{\nu}^\pi \mathbf{1}^\top)(\boldsymbol{\nu}^\pi - \boldsymbol{\lambda}^\epsilon) \\
&\geq \mathbf{e} - \epsilon \mathbf{1} - \left\| \mathbf{D}^\top \Pi^{\epsilon\top} (\mathbf{I} - (\mathbf{P}^\pi)^\top + \boldsymbol{\nu}^\pi \mathbf{1}^\top)^{-1} (\mathbf{I} - (\mathbf{P}^\pi)^\top + \boldsymbol{\nu}^\pi \mathbf{1}^\top)(\boldsymbol{\nu}^\pi - \boldsymbol{\lambda}^\epsilon) \right\|_\infty \cdot \mathbf{1} \\
&\geq \mathbf{e} - \epsilon \mathbf{1} - \max_k \left\| \mathbf{d}_k^\top \Pi^{\epsilon\top} (\mathbf{I} - (\mathbf{P}^\pi)^\top + \boldsymbol{\nu}^\pi \mathbf{1}^\top)^{-1} \right\|_\infty \left\| (\mathbf{I} - (\mathbf{P}^\pi)^\top + \boldsymbol{\nu}^\pi \mathbf{1}^\top)(\boldsymbol{\nu}^\pi - \boldsymbol{\lambda}^\epsilon) \right\|_1 \cdot \mathbf{1} \\
&\geq \mathbf{e} - O(\epsilon) \mathbf{1},
\end{aligned}$$

where the last inequality follows from definition of D, Π^ϵ, Lemma 7 and the fact that

$$\|(\boldsymbol{\lambda}^\epsilon - \boldsymbol{\nu}^\pi)^\top (\mathbf{I} - \mathbf{P}^\pi - \mathbf{1}(\boldsymbol{\nu}^\pi)^\top)\|_1 \leq O(\tfrac{\epsilon}{t_{\mathrm{mix}} D}).$$

From above we have shown that assuming the stationary distribution under π^ϵ is $\boldsymbol{\nu}^\epsilon$, it satisfies $\|\boldsymbol{\nu}^\epsilon - \boldsymbol{\lambda}^\epsilon\|_1 \leq O(\frac{\epsilon}{D t_{\mathrm{mix}}})$, thus giving an approximate solution $\boldsymbol{\mu} = \boldsymbol{\nu}^\epsilon \cdot \pi^\epsilon$ satisfying $\boldsymbol{\mu}(\hat{\mathbf{I}} - \mathbf{P}) = 0$, $\mathbf{D}^\top \boldsymbol{\mu} \geq \mathbf{e} - O(\epsilon)$ and consequently for the original problem (13) an approximately optimal policy π^ϵ. $\square$