Faster saddle-point optimization for solving large-scale Markov decision processes

Joan Bas-Serrano JOANBASSERRANO@GMAIL.COM
Universitat Pompeu Fabra, Barcelona, Spain
Gergely Neu GERGELY.NEU@GMAIL.COM
Universitat Pompeu Fabra, Barcelona, Spain

Abstract

We consider the problem of computing optimal policies in average-reward Markov decision processes. This classical problem can be formulated as a linear program directly amenable to saddle-point optimization methods, albeit with a number of variables that is linear in the number of states. To address this issue, recent work has considered a linearly relaxed version of the resulting saddle-point problem. Our work aims at achieving a better understanding of this relaxed optimization problem by characterizing the conditions necessary for convergence to the optimal policy, and designing an optimization algorithm enjoying fast convergence rates that are independent of the size of the state space. Notably, our characterization points out some potential issues with previous work.

1. Introduction

Computing optimal policies in Markov decision processes (MDPs) is one of the most important problems in sequential decision making and control (Puterman, 1994). Arguably, the most classical approach to solve this task is through the method of *dynamic programming*, understood in this context as computing fixed points of certain operators (Bellman, 1957; Howard, 1960; Bertsekas, 2007). The use and influence of dynamic-programming methods like value and policy iteration extend well beyond the world of decision and control theory, as the underlying ideas serve as foundations for most algorithms for *learning* optimal policies in unknown MDPs: the setting of *reinforcement learning* (Szepesvári, 2010; Sutton and Barto, 2018). While being hugely successful, DP-based methods have the downside of being somewhat incompatible with classical machine-learning tools that are rooted in convex optimization. Indeed, most of the popular reductions of dynamic programming to (non-)convex optimization are based on heuristics that are not directly motivated by theory. Examples include the celebrated DQN approach of Mnih et al. (2015) that reduces value-function estimation to minimizing the "squared Bellman error", or the TRPO algorithm of Schulman et al. (2015) that reduces policy updates to minimizing a "regularized surrogate objective". While these methods can be justified to a certain extent, it is technically unknown if solving the resulting optimization problems actually leads to a desirable solution to the original sequential decision-making problem.

In this paper, we explore a family of methods that reduce MDP optimization to a form of convex optimization in a theoretically grounded way. Our starting point is an alternative approach based on linear programming (LP), first proposed roughly at the same time as the DP methods of Bellman (1957); Howard (1960): the idea of LP-based methods for sequential decision-making goes back to the works of de Ghellinck (1960); Manne (1960); Denardo (1970). While LP-based methods seem to be more obscure in present day than DP methods, they have the clear advantage that they lead to an objective function directly amenable to modern large-scale optimization methods. Recent reinforcement-learning methods inspired by the LP perspective include policy-gradient and actor-critic methods (Sutton et al., 1999; Konda and Tsitsiklis, 1999) and various "entropy-regularized" learning algorithms (e.g., Peters et al., 2010; Zimin and Neu, 2013; Neu et al., 2017). While these methods promise to directly tackle the policy-optimization problem through

solving the underlying linear program, most of them still require the computation of certain value functions through dynamic programming.

In the present work, we argue for the viability of a method fully based on a form of convex optimization, rooted in the LP approach. Our approach is based on a *bilinear saddle-point* formulation of the linear program, building on a well-known general equivalence between the two optimization problems. One particular advantage of this formulation is that it enables a straightforward form of dimensionality reduction of the original problem through a linear parametrization of the optimization variables, which provides a natural framework for studying effects of "function approximation" in the underlying policy optimization problem. Our main contribution regarding this setting lies in characterizing a set of assumptions that allow a reduced-order saddle-point representation of the optimal policy. These include a realizability assumption and a newly identified *coherence assumption* about the subspaces used for approximation. Our main positive result is showing that these conditions are sufficient for constructing an algorithm that outputs an ε-optimal policy with runtime guarantees of $\widetilde{\mathcal{O}}\left(\tau_{\text{mix}}^2 N^3/\varepsilon\right)$, where N is the number of variables in the relaxed optimization problem, and τ_{mix} is a notion of mixing time. Our approach is based on the celebrated Mirror Prox algorithm of Nemirovski (2004) (see also Korpelevich, 1976). We complement our positive results by showing that our newly defined coherence assumption is necessary for the relaxed saddle-point approach to be viable: we construct a simple example violating the assumption, where achieving full optimality on the relaxed problem leads to a suboptimal policy.

We are not the first to consider saddle-point methods for optimization in Markov decision processes. Wang (2017) proposed variants of Mirror Descent to solve the original saddle-point problem without relaxations and provide runtime guarantees of $\widetilde{\mathcal{O}}\left((\alpha\tau_{\text{mix}})^2 |\mathcal{X}||\mathcal{A}|/\varepsilon^2\right)$, where $\mathcal{X}$ and $\mathcal{A}$ are the finite state and action spaces, and α is a parameter that characterizes the uniformity of the stationary distributions of every policy. Specifically, their assumption implies[1] that for the stationary distribution d_π any policy π, one has $\frac{\max_x d_\pi(x)}{\min_{x'} d_\pi(x')} \leq \alpha$. In most cases of practical interest, this ratio is at least as large as $|\mathcal{X}|$ (e.g., when there are states that some policies visit with constant probability), and can easily be exponentially large in $|\mathcal{X}|$, or even infinite if the underlying MDP has transient states. When specialized to this setting, our bounds replace α^2 by the much more manageable $|\mathcal{X}|$ and also improve the dependence on ε from $1/\varepsilon^2$ to $1/\varepsilon$. One downside of our method is that we need full access to the transition probabilities of the MDP, whereas the algorithm of Wang (2017) only requires a generative model.

The linearly relaxed saddle-point problem we consider was first studied by Lakshminarayanan et al. (2018) and Chen et al. (2018). Our runtime guarantees improve over the ones claimed by Chen et al. (2018) in a similar way as our first set of results improve over those of Wang (2017). Notably, their results still feature a factor of α^2, which generally depends on the size of the original state space rather than the number of features, rendering these guarantees void of meaning in very large state spaces. In contrast, our bounds replace this factor by the number of features N. Furthermore, our characterization highlighting the importance of the coherence assumption discussed above hints at some potential technical issues with the results of Chen, Li, and Wang (2018), who claimed convergence to the optimal policy *without the coherence assumption*.

The rest of the paper is organized as follows. After providing background on the saddle-point formulation of MDP optimization in Section 2, we describe the relaxed saddle-point problem in Section 3. Section 4 presents our algorithm and its performance guarantees, and Section 5 provides a sketch of the proofs. We conclude by providing a simple numerical illustration of our method in Section 6 and discuss our results in Section 7.

Notation. Inner products over vector spaces will be denoted by $\langle\cdot,\cdot\rangle$. We use $\Delta_{\mathcal{S}}$ to denote the set of probability distributions on the finite set $\mathcal{S}$: $\Delta_{\mathcal{S}} = \left\{p \in \mathbb{R}_+^{\mathcal{S}} : \sum_{s\in\mathcal{S}} p(s) = 1\right\}$. Sums spanning over the spaces $x \in \mathcal{X}$ and $a \in \mathcal{A}$ will be simply denoted by $\sum_x$ or $\sum_a$.

1. The actual assumption made by Wang (2017) is even more restrictive.

2. Preliminaries

Consider an undiscounted Markov decision process $M = (\mathcal{X}, \mathcal{A}, P, r)$, where $\mathcal{X}$ is the finite state space, $\mathcal{A}$ is the finite action space, P is the transition function with $P(x'|x,a)$ denoting the probability of moving to state $x' \in \mathcal{X}$ from state $x \in \mathcal{X}$ when taking action $a \in \mathcal{A}$ and r is the reward function mapping state-action pairs to rewards with $r(x,a)$ denoting the reward of being in state x and taking action a. We assume that $r(x,a) \in [0,1]$ for all x, a. In each round t, the learner observes state $x_t \in \mathcal{X}$, selects action $a_t \in \mathcal{A}$, moves to the next state $x_{t+1} \sim P(\cdot|x_t, a_t)$, and obtains reward $r(x_t, a_t)$.

In this paper we focus on the infinite-horizon average-reward scenario where the goal of the learner is to select its actions a_t in a way that maximizes the average reward per time step, $\liminf_{t\to\infty} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^{T} r_t(x_t, a_t)\right]$. We will work with randomized stationary policies with $\pi(a|x)$ denoting the probability of taking action a in state x. Under technical assumptions discussed shortly, each such policy π generates a unique stationary state distribution $d_\pi \in \Delta_{\mathcal{X}}$ over the state space satisfying $d_\pi(x) = \lim_{t\to\infty} \mathbb{P}[x_t = x]$ for all x when the trajectory $(x_t)_t$ is generated by following policy π. Similarly, each policy π generates a stationary state-action distribution $\mu_\pi \in \Delta_{\mathcal{X}\times\mathcal{A}}$ satisfying $\mu_\pi(x,a) = \lim_{t\to\infty} \mathbb{P}[x_t = x, a_t = a] = d_\pi(x)\pi(a|x)$. Given these definitions, it can be easily shown that the average-reward of a policy π can be written as

$$\rho_\pi = \lim_{t\to\infty} \inf \mathbb{E}_\pi\left[\frac{1}{T}\sum_{t=1}^{T} r_t(x_t, a_t)\right] = \sum_{x,a} \mu(x,a) r(x,a),$$

where the notation $\mathbb{E}_\pi[\cdot]$ indicates that the trajectory $(x_t, a_t)_t$ was generated by following policy π: $a_t \sim \pi(\cdot|x_t)$ and $x_{t+1} \sim P(\cdot|x_t, a_t)$. Under our assumptions, the optimal policy can be shown to be a stationary one; we will denote its average reward as $\rho^* = \max_\pi \rho_\pi$. Thus, one can show that finding the optimal policy is equivalent to solving the following linear program:

$$\begin{aligned} \text{maximize} \quad & \sum_{x,a} \mu(x,a) r(x,a) \\ \text{s.t.} \quad & \mu \in \Delta_{\mathcal{X}\times\mathcal{A}}, \qquad \sum_{a'} \mu(x',a') = \sum_{x,a} P(x'|x,a)\mu(x,a) \quad (\forall x' \in \mathcal{X}). \end{aligned}$$

To simplify our notation, we will represent μ and r by $|\mathcal{X}\times\mathcal{A}|$-dimensional vectors and also define the $|\mathcal{X}\times\mathcal{A}|\times|\mathcal{X}|$-dimensional matrix Q with entries $Q_{(x,a),x'} = P(x'|x,a) - \mathbb{I}_{\{x'=x\}}$. Then, one can easily see[2] that solving the linear program stated above is equivalent to finding the following *saddle point*:

$$\min_{v\in\mathbb{R}^{|\mathcal{X}|}} \max_{\mu\in\Delta} \mathcal{L}(v,\mu) = \min_{v\in\mathbb{R}^{|\mathcal{X}|}} \max_{\mu\in\Delta} \langle \mu, Qv\rangle + \langle \mu, r\rangle. \tag{1}$$

Here, we introduced the *Lagrangian function* $\mathcal{L}$ and the shorthand $\Delta = \Delta_{\mathcal{X}\times\mathcal{A}}$. Optimal solutions (v^*, μ^*) to the above saddle-point problem are easily seen to correspond to the stationary distribution μ^* of the optimal policy and the *optimal differential value function* v^* (also known as the optimal bias function, cf. Puterman, 1994). Besides the full saddle-point optimization problem, we will consider a relaxed version based on the introduction feature maps. Details on this variant are provided in Section 3.

We will make two structural assumptions about the underlying Markov decision process. The first of these guarantees the existence of stationary distributions for all policies.

Assumption 1 (Uniform ergodicity) *Every policy π generates an ergodic Markov chain. Specifically, letting P_π be the transition operator of π defined as the matrix with elements $P_\pi(x'|x) = \sum_a \pi(a|x)P(x'|x,a)$, and d, d' be any two distributions over $\mathcal{X}$, the following inequality is satisfied for some $C, \tau > 0$ and for all k:*

$$\left\| (d - d') P_\pi^k \right\|_1 \le C e^{-k/\tau} \left\| d - d' \right\|_1.$$

We say that our MDP is *uniformly ergodic* if it satisfies Assumption 1. Notice that this assumption is significantly weaker than the 1-step mixing assumption often made in the related literature (Even-Dar

2. This can be seen, e.g., by introducing the KKT multipliers for the constraints in the linear program.

et al., 2009; Neu et al., 2014). It is easily shown to hold when all policies induce aperiodic and irreducible Markov chains—see Theorem 4.9 in Levin et al. (2017) for a proof. Clearly, this assumption immediately implies that every policy admits a unique stationary distribution as required in the discussion above. In what follows below, we will often use the notation $\tau_{\text{mix}} = 2C(\tau+1)$ and refer to this quantity as the *mixing time* of the MDP[3].

Given this assumption and the above definitions, we can establish a number of useful facts about the optimal solutions (v^*, μ^*) to the saddle-point problem (1). We first note that an optimal policy π^* can be extracted from μ^* in the states where $\mu^*(x,\cdot) > 0$ as $\pi^*(a|x) = \frac{\mu^*(x,a)}{\sum_{a'}\mu^*(x,a')}$. Regarding v^*, the following proposition summarizes some of its most important properties:

Proposition 1 *Let (v^*, μ^*) be a solution of the problem (1). Then, v^* satisfies the following properties:*

- *v^* satisfies the Bellman optimality equations $v^*(x) = r(x) - \rho^* + \sum_{x'} P(x'|x,a)v^*(x')$ for all x; for any $c \in \mathbb{R}$, $v^* + c$ is also a solution to (1);*
- *for any x, x', $|v^*(x) - v^*(x')| \le \tau_{mix} = 2C(\tau+1)$.*

All of these properties can be proven by standard arguments; we refer the reader to Lemma 1 in Wang (2017) for a proof of the first item and Lemma 3 in Neu et al. (2014) for a proof of the second one.

3. The linearly relaxed saddle-point problem

While one can directly derive optimization algorithms to solve the saddle-point problem (1), such a direct approach would suffer from serious scalability issues due to the sheer number of variables involved in the problem: the size of the objects of interest μ and v are linear in the size of the state space, which results in prohibitive memory and computation costs for most algorithms. To address this issue, we study a *linearly relaxed* version of the full saddle-point problem that reduces the order of the original optimization problem by linearly parametrizing the variables v and μ through two sets of *feature maps*. Formally, we consider the matrices F of size $|\mathcal{X}| \times N$ and W of size $M \times |\mathcal{X} \times \mathcal{A}|$, introduce the new optimization variables $y \in \mathbb{R}^M$ and $u \in \mathbb{R}^N$, and use these to (hopefully) approximate the solutions to (1) as $\mu^* \approx yW$ and $v^* \approx Fu$. For a tractable problem formulation, we will assume that the rows of W are non-negative and sum to one: $W_{m,x} \ge 0$ for all x, m and $\sum_x W_{m,x} = 1$ for all m. We will also assume that all entries of F are bounded by 1 in absolute value. These conditions enable us to optimize y over the probability simplex $\widetilde{\Delta} = \Delta_{[M]}$ and to formulate our relaxed saddle-point problem as

$$\min_{u\in\mathbb{R}^N}\max_{y\in\widetilde{\Delta}} \widetilde{\mathcal{L}}(u,y) = \min_{u\in\mathbb{R}^N}\max_{y\in\widetilde{\Delta}} \left\langle W^\mathsf{T} y, QFu\right\rangle + \left\langle W^\mathsf{T} y, r\right\rangle. \tag{2}$$

The relaxed optimization problem above has been studied before by Lakshminarayanan and Bhatnagar (2015); Lakshminarayanan et al. (2018), and Chen et al. (2018). Lakshminarayanan and Bhatnagar (2015); Lakshminarayanan et al. (2018) studied the relaxed linear program underlying (2) as a natural extension of the classic relaxed LP analyzed by de Farias and Van Roy (2003), and have focused on understanding the discrepancies between the optimal value function and the relaxed value function attaining the minimum in the above expression. On the other hand, Chen et al. (2018) focused on proposing stochastic optimization algorithms and analyzing the rate of convergence to the optimum, but provide little insight about the quality of the optimal solution of the relaxed problem.

One of our main goals in the present paper is to obtain a better understanding of the effects of approximation on the policies that can be obtained through approximately solving the the relaxed saddle-point problem (2). One peculiar challenge associated with our setting is that it is not enough to ensure that the values of $\widetilde{\mathcal{L}}$ and $\mathcal{L}$ are close at their respective saddle points, but we rather need to understand the

3. Note that this is just one of many possible definitions of a mixing time, see, e.g., Seneta (2006); Levin et al. (2017).

performance of the policy extracted from the optimal solution y^*. Precisely, defining the policy extracted from y as

$$\pi_y(a|x) = \frac{(W^\mathsf{T} y)(x,a)}{\sum_{a'}(W^\mathsf{T} y)(x,a')}$$

for all x, a, and the corresponding stationary distribution as μ_y induced in the original MDP, we are interested in the suboptimality gap $\langle \mu^* - \mu_{y^*}, r \rangle$. In the present paper, we focus on identifying assumptions on the feature maps that allow the computation of true optimal policies with (almost) zero suboptimality gap. Specifically, we will show that the following two assumptions have a decisive role in making this gap small:

Assumption 2 (Realizability) *The optimal solution is realizable by the feature maps: there exists* (u^*, y^*) *such that* $v^* = Fu^*$ *and* $\mu^* = W^\mathsf{T} y^*$. *Additionally,* $\|u^*\|_\infty \le U\tau_{mix}$ *holds for some* $U > 0$.

Assumption 3 (Coherence) *The image of the set* $\widetilde{\Delta}$ *under the map* $Q^\mathsf{T} W^\mathsf{T}$ *is included the column space of* F*: for all* $y \in \widetilde{\Delta}$ *such that* $Q^\mathsf{T} W^\mathsf{T} y \neq 0$, *there exists a* $u \in \mathbb{R}^N$ *such that* $\langle Q^\mathsf{T} W^\mathsf{T} y, Fu \rangle \neq 0$. *Additionally, for all* $v \in \mathbb{R}^{|\mathcal{X}|}$ *with* $\|v\|_\infty \le 1$, *there exists a* $u \in \mathbb{R}^N$ *with* $\|u\|_\infty \le U$ *such that* $\langle Q^\mathsf{T} W^\mathsf{T} y, Fu \rangle = \langle Q^\mathsf{T} W^\mathsf{T} y, v \rangle$.

The second condition of each assumption is to ensure that the columns of F are well-conditioned and are satisfied if the columns form an orthonormal basis. While realizability may already seem sufficient for the relaxed problem to be a good enough approximation of the original one, we argue that the second assumption is also necessary for the relaxation scheme to be reliable. Specifically, the following theorem shows that in the absence of the coherence assumption, near-optimal solutions to the relaxed saddle-point problem (2) can still lead to suboptimal policies in the original MDP.

Theorem 1 *For any* $\varepsilon > 0$, *there exists an MDP with relaxations* W, F *satisfying Assumption 2 and violating Assumption 3, and a solution* $(\widehat{u}, \widehat{y}_\varepsilon)$ *simultaneously satisfying*

$$\mathcal{L}(F\widehat{u}, \mu^*) - \mathcal{L}(v^*, W^\mathsf{T}\widehat{y}_\varepsilon) = \varepsilon$$

and

$$\langle \mu^* - \mu_{\widehat{y}_\varepsilon}, r \rangle = 2/3.$$

Proof The proof is based on constructing an MDP with three states x_1 (left), x_2 (middle) and x_3 (right) and two actions a_l and a_r corresponding to moving "left" or "right", respectively. The transition probabilities and rewards are as shown on Figure 1. It is easy to see that the optimal policy is to take action a_r in state x_2, which yields the optimal stationary state-action distribution

$$\mu^* = \big(\mu(x_1,a_r), \mu(x_2,a_l), \mu(x_2,a_r), \mu(x_3,a_l)\big)^\mathsf{T} = \left(0, 0, \frac{1}{3}, \frac{2}{3}\right)^\mathsf{T}$$

and the optimal average reward $\rho^* = 1$. The optimal value function can be shown to be $v^* = (-1,-1,1)^\mathsf{T}$. For the relaxation, define $F = v^*$ and W as the identity map so that the realizability assumption is clearly fulfilled with $y^* = \mu^*$ and $u^* = 1$. Now, choosing $\widehat{y} = (1,0,0,0)^\mathsf{T}$ results in

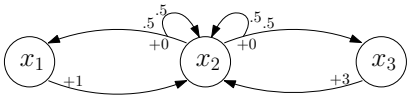


Figure 1: Three-state MDP for illustrating the necessity of the coherence assumption. Transitions from x_2 are stochastic with probability 1/2 of staying in x_2 and moving to x_1 and x_3 otherwise, depending on the chosen action. All other transitions are deterministic. Rewards are given as a function of the state as $r(x_1) = 1$, $r(x_2) = 0$ and $r(x_3) = 3$.

$$\langle W^\mathsf{T}\widehat{y}, QFu \rangle = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} -1 & 1 & 0 \\ 1/2 & -1/2 & 0 \\ 0 & -1/2 & 1/2 \\ 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} -1 \\ -1 \\ 1 \end{pmatrix} u = \begin{pmatrix} 1 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 2 \end{pmatrix} u = 0 \cdot u$$

for any u. Observing that taking $v = (-1,1,0)^\mathsf{T}$ gives $\langle W^\mathsf{T}\widehat{y}, Qv\rangle = 2$, we see that the coherence assumption is violated since there exists no u such that the condition $\langle W^\mathsf{T}\widehat{y}, Qv\rangle = \langle W^\mathsf{T}\widehat{y}, QFu\rangle$ is satisfied. Furthermore, it is easy to see that $(\widehat{y}, u)$ for any u is an optimal solution to (2) with value $\rho^* = 1$ since

$$\widetilde{\mathcal{L}}(u,\widehat{y}) = \widehat{y}^\mathsf{T} WQFu + \widehat{y}^\mathsf{T} Wr = \begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix}\begin{pmatrix}1\\0\\0\\3\end{pmatrix} = 1.$$

showing that $(\widehat{y}, u)$ with any u is also an optimal solution to the relaxed saddle-point problem (2). The resulting optimal state-action distribution $\widehat{\mu} = \widehat{y}W = \widehat{y}$ is clearly not a stationary distribution.

To conclude the proof, fix any ε and consider $\widehat{y}_\varepsilon = (1-\varepsilon, \varepsilon, 0, 0)^\mathsf{T}$ and any $\widehat{u}$. Noticing that $\langle W^\mathsf{T}\widehat{y}_\varepsilon, QFu\rangle = 0$ holds for all u, the duality gap associated with $(\widehat{u}, \widehat{y}_\varepsilon)$ can be seen to be

$$\mathcal{L}(F\widehat{u}, \mu^*) - \mathcal{L}(v^*, W^\mathsf{T}\widehat{y}_\varepsilon) = \begin{pmatrix}0 & 0 & 2/3 & 1/3\end{pmatrix}\begin{pmatrix}1\\0\\0\\3\end{pmatrix} - \begin{pmatrix}1-\varepsilon & \varepsilon & 0 & 0\end{pmatrix}\begin{pmatrix}1\\0\\0\\3\end{pmatrix} = 1 - (1-\varepsilon) = \varepsilon.$$

The policy $\pi_{\widehat{y}_\varepsilon}$ extracted from the state-action distribution $\widehat{y}_\varepsilon$ takes action a_l in state x_2, which results in an average reward of $2/3$. These two statements together prove the theorem. ■

4. Algorithm and main results

In this section, we provide our main positive results: deriving strong performance guarantees for policies derived from approximate solutions of (2) under Assumptions 2 and 3. Our algorithm attaining these guarantees is based on the Optimistic Mirror Descent framework proposed by Rakhlin and Sridharan (2013a,b), and more specifically on its variant known as Mirror Prox due to Nemirovski (2004) (see also Sections 4.5 and 5.2.3 in Bubeck (2015) for an easily accessible overview of this method).

For a generic description of Mirror Prox on a convex set $\mathcal{Z}$, we let $G : \mathcal{Z} \to \mathbb{R}$ be a monotone operator satisfying $\langle G(z) - G(z'), z - z'\rangle \geq 0$ for all $z, z' \in \mathcal{Z}$, and let $\Phi : \mathcal{Z} \to \mathbb{R}$ be a σ-strongly convex regularization function under some norm $\|\cdot\|$ with its corresponding Bergman divergence $D_\Phi(z\|z') = \Phi(z) - \Phi(z') - \langle \nabla\Phi(z'), z - z'\rangle$. Mirror Prox computes a sequence of iterates with $z_1 \in \arg\min \Phi(z)$ and

$$\begin{aligned}\widehat{z}_{t+1} &= \arg\min_{\mathcal{Z}} \eta\langle G(z_t), z\rangle + D_\Phi(z, z_t)\\ z_{t+1} &= \arg\min_{\mathcal{Z}} \eta\langle G(\widehat{z}_{t+1}), z\rangle + D_\Phi(z, z_t).\end{aligned} \tag{3}$$

The first of these steps is often referred to as an *extrapolation step*. A simpler version of this algorithm not involving such an extrapolation step is commonly known as Mirror Descent (Nemirovski and Yudin, 1983; Beck and Teboulle, 2003; Bubeck, 2015). This step serves to enhance the stability of the algorithm, and indeed Mirror Prox can be shown to enjoy favorable convergence properties in the problem setting described above.

We instantiate the Mirror Prox method to address the relaxed saddle-point problem as follows. Our optimization variables will be $z = (u, y)$ and the monotone operator G will be chosen as

$$G(z) = \begin{pmatrix}\nabla_v \widetilde{\mathcal{L}}\\ -\nabla_\mu \widetilde{\mathcal{L}}\end{pmatrix} = \begin{pmatrix}F^\mathsf{T} Q^\mathsf{T} W^\mathsf{T} y\\ -Wr - WQFu\end{pmatrix}. \tag{4}$$

We will use the regularization function

$$\Phi(z) = \frac{1}{2}\|u\|_2^2 + \sum_{j=1}^{M} y_j \log y_j,$$

that is, a linear combination of the squared 2-norm of the value-function parameters u and the Shannon entropy of the distribution y. Clearly, Φ is 1-strongly convex on $\mathcal{Z}$ with respect to the norm $\|z\|^2 = \|u\|_2^2 + \|y\|_1^2$. Given the above specifications, the updates of our algorithm can be written as

$$\widehat{u}_{t+1} = u_t - \eta F^\top Q^\top W^\top y_t, \quad \widehat{y}_{t+1,i} \propto y_{t,i} e^{\eta((Wr)_i + (WQFu_t)_i)} \tag{5}$$

$$u_{t+1} = u_t - \eta F^\top Q^\top W^\top \widetilde{y}_{t+1}, \quad y_{t+1,i} \propto y_{t,i} e^{\eta((Wr)_i + (WQF\widehat{u}_{t+1})_i)}, \tag{6}$$

where we used the notation "$\propto$" to signify that $\widehat{y}_{t+1}$ and y_{t+1} are normalized multiplicatively after each update so that $\sum_j y_{t+1,j} = 1$ is satisfied. Also introducing the notations $\overline{y}_T = \frac{1}{T}\sum_{t=1}^T y_t$ and $\overline{u}_T = \frac{1}{T}\sum_{t=1}^T \widehat{u}_t$, the algorithm outputs the policy extracted from the distribution $\overline{y}_T$: $\pi_T = \pi_{\overline{y}_T}$. Letting $d_T = d_{\pi_T}$ be the stationary distribution associated with π_T, the corresponding average reward can be written as $\rho_T = \sum_{x,a} d_T(x)\pi_T(a|x)r(x,a)$. The following theorem presents our main result regarding the suboptimality of the resulting policy in terms of its average reward.

Theorem 2 *Suppose that Assumptions 1, 2 and 3 hold and $\eta \le 1/4N$. Then, the average reward ρ_T output by the algorithm satisfies*

$$\rho^* - \rho_T \le \frac{11\tau_{mix}^2 U^2 N + 7\log M}{\eta T}.$$

In particular, setting $\eta = 1/4N$, the bound becomes $\rho^* - \rho_T = \mathcal{O}\left(\frac{\tau_{mix}^2 N^2 U^2}{T}\right)$.

We note that this result can be tightened by a factor of N if we further assume that the rows of F are chosen as probability distributions. In the special case where F and W are the identity maps, the relaxed saddle-point problem becomes the original problem (1), and our Assumptions 2 and 3 are clearly satisfied with $U = 1$. In this case, our algorithm satisfies the following bound:

Corollary 3 *Suppose that Assumption 1 holds, W and F are the identity maps, and $\eta \le 1/4$. Then, the average reward ρ_T of the policy output by our algorithm satisfies*

$$\rho^* - \rho_T \le \frac{11\tau_{mix}^2 |\mathcal{X}| + 7\log(|\mathcal{X}||\mathcal{A}|)}{\eta T}.$$

In particular, setting $\eta = 1/4$, the bound becomes $\rho^* - \rho_T = \widetilde{\mathcal{O}}\left(\frac{\tau_{mix}^2 |\mathcal{X}|}{T}\right)$.

A brief inspection of Equations (5)-(6) suggests that each update of our algorithm can be computed in $\mathcal{O}(MN)$ time, the most expensive operation being computing the matrix-vector products $WQFu$ and $y^\top WQF$. While this suggests that the algorithm may have runtime and memory complexity independent of the size of the state space, we note that exact computation of the matrix WQF can still take $\mathcal{O}\left(|\mathcal{X}|^2|\mathcal{A}|\right)$ time in the worst case. This can be improved to $\mathcal{O}(K)$ when assuming that only K entries of the transition matrix P are nonzero, which can be of order $|\mathcal{X}||\mathcal{A}|$ in many interesting problems where the support of $P(\cdot|x,a)$ is of size $\mathcal{O}(1)$ for all x, a. We stress however that the matrix WQF only needs to be computed *once* as an initialization step of our algorithm. In contrast, a general algorithm like value iteration needs at least $\Theta(K) = \Theta(|\mathcal{X}||\mathcal{A}|)$ for computing *each update*, showing a clear computational advantage of our method. Further discussion of computational issues is deferred to Section 7.

5. Analysis

This section provides an outline of the analysis of our algorithm. At a high level, our analysis builds on some well-known results regarding the performance of Mirror Prox, including a classical bound on the *duality gap* of the obtained solutions. The crucial challenge posed by our setting is connecting the duality gap on the saddle-point problem to a suboptimality gap of the extracted policies. The key innovation in our

analysis is providing a new technique to connect these quentities through exploiting further properties of Mirror Prox. In what follows, we first provide some general tools that will be helpful throughout the proofs, and then provide the proof outline for Theorem 2. Full proofs are provided in Appendix A.

A central piece of our our analysis is the following useful lemma regarding the iterates computed by Mirror Prox:

Lemma 4 *Let Φ be σ-strongly convex and F be L-Lipschitz. Then, for all t, Mirror Prox guarantees*

$$\eta \langle \widehat{z}_{t+1} - z, G(\widehat{z}_{t+1}) \rangle \leq D_\Phi(z \| z_t) - D_\Phi(z \| z_{t+1}) - \frac{\sigma - \eta L}{4} \| z_{t+1} - z_t \|^2 .$$

holds for every $z \in \mathcal{Z}$ and $t > 0$.

The proof is based on standard arguments, see, for instance, Lemma 1 of Rakhlin and Sridharan (2013b). We include it in Appendix A.1 for completeness. This lemma has two important corollaries that we will crucially use throughout the analysis. The first one shows that the iterates remain bounded during the optimization procedure.

Corollary 5 *Let $z^* = (u^*, y^*)$ be any solution to $\max_y \min_u \widetilde{\mathcal{L}}(u, y)$ and suppose that the conditions of Lemma 4 hold. Then, for all t, Mirror Prox guarantees*

$$D_\Phi(z^* \| z_t) \leq D_\Phi(z^* \| z_0) .$$

The proof follows from noticing that z^*, being an optimal solution to the saddle-point problem, satisfies the variational inequality $\langle \widehat{z}_{t+1} - z^*, G(\widehat{z}_{t+1}) \rangle \geq 0$. The second corollary establishes a bound on the *duality gap* evaluated at $(\overline{u}_T, \overline{y}_T)$:

Corollary 6 *Let $z = (u, y) \in \mathcal{Z}$ be arbitrary and assume that $\eta \leq \frac{\sigma}{2L}$. Then, Mirror Prox guarantees the following bound on the duality gap:*

$$\mathcal{L}(\overline{u}_T, y) - \mathcal{L}(u, \overline{y}_T) \leq \frac{D_\Phi(z, z_0)}{\eta T} .$$

The proof easily follows by noticing that $\langle \widehat{z}_{t+1} - z, G(\widehat{z}_{t+1}) \rangle$ equals the duality gap evaluated at $(\widehat{u}_{t+1}, \widehat{y}_{t+1})$, and summing the bound given in Lemma 4.

In order to apply the above tools to our problem, we first need to confirm that our objective is indeed smooth with respect to the norm $\|z\|^2 = \|u\|_2^2 + \|y\|_1^2$. The following lemma establishes this property.

Lemma 7 *Let $K = \max_x \|F_{x,\cdot}\|_1$. Then, the function $\widetilde{\mathcal{L}}$ is $2K$-smooth with respect to $\|\cdot\|$.*

The proof is provided in Appendix A.3. Notably, this lemma implies that the $\widetilde{\mathcal{L}}$ is 2-smooth when the rows of F form probability distributions. In the worst case, however, when we only assume that the entries of F are bounded in absolute value by 1, the smoothness constant can be as large as $2N$. In what follows, we will assume that $\eta \leq 1/(4K)$.

We proceed by appealing to the realizability assumption to choose $x = (u^*, y^*)$ such that $z = (v^*, \mu^*) = (Fu^*, W^\top y^*)$, and observe that

$$\widetilde{\mathcal{L}}(\overline{u}_T, y^*) - \widetilde{\mathcal{L}}(u^*, \overline{y}_T) = \langle \mu^*, QF\overline{u}_T + r \rangle - \langle W^\top \overline{y}_T, Qv^* + r \rangle \leq \frac{D_\Phi(z^*, z_0)}{\eta T}$$

holds by virtue of Corollary 6 and the choice of η. Observing that $Q^\top \mu^* = 0$ holds due to the stationarity of μ^* and reordering gives

$$\langle \mu^* - W^\top \overline{y}_T, r \rangle \leq \frac{D_\Phi(z^*, z_0)}{\eta T} + \langle Q^\top W^\top \overline{\mu}_T, v^* \rangle . \tag{7}$$

The remaining key question is how to relate $\langle W^\top \overline{y}_T, r \rangle$ to the true average reward ρ_T associated with the extracted policy. This is done with the help of the following lemma, one of our key results:

Lemma 8 *Suppose that Assumption 1 holds. Let μ be an arbitrary distribution over $\mathcal{X} \times \mathcal{A}$ and let π_μ be the policy extracted from μ. Then, the average reward ρ_μ of π_μ satisfies $\langle \mu, r \rangle - \rho_\mu \leq \tau_{mix} \left\| Q^\mathsf{T} \mu \right\|_1$.*

The proof is provided in Appendix A.2. Combining this result with the bound of Equation 7 and using that $\|v^*\|_\infty \leq \tau_{\text{mix}}$, we obtain

$$\rho^* - \rho_T \leq \frac{D_\Phi(z^*, z_0)}{\eta T} + 2\tau_{\text{mix}} \left\| Q^\mathsf{T} W^\mathsf{T} \overline{y}_T \right\|_1 . \tag{8}$$

Thus, it only remains to bound $\left\| Q^\mathsf{T} W^\mathsf{T} \overline{y}_T \right\|_1$. In order to do this, we crucially use Assumption 3 that guarantees the coherence of the feature maps to prove the following result:

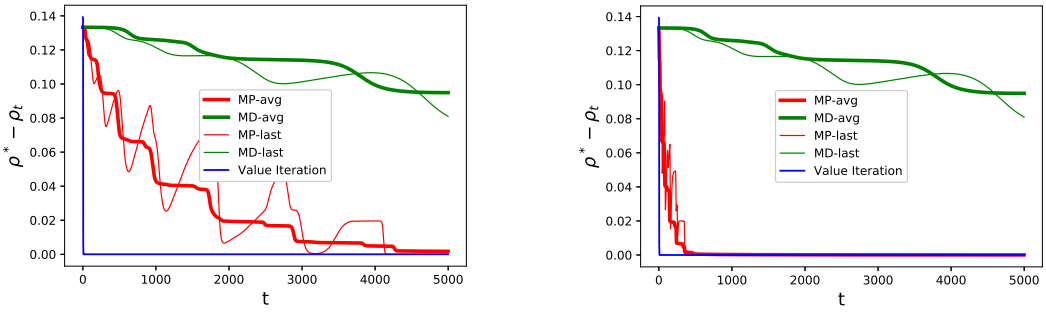
Lemma 9 *Suppose that Assumptions 2 and 3 hold. Then,*

$$\tau_{mix} \left\| Q^\mathsf{T} W^\mathsf{T} \bar{y}_T \right\|_1 \leq \frac{5\tau_{mix}^2 U^2 N + 3 \log M}{\eta T}$$

The proof of this lemma is provided in Appendix A.4. Combining the bound of this lemma with Equation (8) and using $D_\Phi(z^* \| z_0) \leq \tau_{\text{mix}}^2 N + \log(M)$ concludes our proof of Theorem 3.

6. Numerical illustration

In this section, we provide preliminary empirical results on a simple Markov decision process in order to illustrate our theoretical results, and specifically compare the performance of our algorithm with that of Mirror Descent and the classic value iteration algorithm. We consider a rectangular $s \times s$ gridworld with one nonzero reward placed in state x_r, so that $r(x,a) = \mathbb{I}_{x=x_r}$. Once the agent arrives to x_r, it is randomly teleported to any of the other states with equal probability. In any other state, the agent can decide to move to a neighboring cell in any direction. The attempt to move in the desired direction is successful with probability p, otherwise the agent moves in the opposite direction with probability $1-p$. If the agent is in an edge of the grid and it makes an step in the direction of the edge, it appears in the opposite edge.



(a) $p = 0.9$, and $\eta = \frac{1}{4}$.

(b) $p = 0.9$, and $\eta = 3$.

Figure 2: Regret as a function of the number of iterations of MP, MD, and value iteration in a grid world example.

Figure 2 shows some results on a grid of side $s = 10$. We observe that the convergence of Mirror Prox is much faster than that of Mirror Descent, and that the last iterate of MP converges very quickly to the optimum, achieving it after *finitely many iterations*. We also note that for higher values of η than the ones found to be safe in our bounds (at most $1/4$), the algorithm is still stable and can lead to faster convergence to the optimum. Explaining the fast convergence of the last iterate of Mirror Prox from a theoretical perspective remains an interesting open problem. Similarly, we are not aware of any performance guarantees for value iteration in the setting that we consider (average reward MDPs under the relatively weak Assumption 1),

nevertheless we observe that this classic algorithm performs spectacularly well on the simple example we consider. On the other hand, we stress that the advantage of our methodology is being able to directly address large-scale problems through the linearly relaxed saddle-point formulation—a structure that value iteration cannot exploit.

7. Discussion

Our most important contributions concern the relaxed saddle-point problem (2), most notably including our discussion on the necessity and sufficience of the coherence assumption (Assumption 3). As we've mentioned earlier, several relaxation schemes similar to ours have been studied in the literature. In fact, relaxing the linear program underlying (1) through the introduction of the feature map F for approximating the value function v^* is one of the oldest ideas in approximate dynamic programming, originally introduced by Schweitzer and Seidman (1985). The effects of this approximation were studied by de Farias and Van Roy (2003) in the context of discounted Markov decision processes. A relaxation scheme involving both the feature maps F and W was considered by Lakshminarayanan and Bhatnagar (2015); Lakshminarayanan et al. (2018). Both sets of authors carefully observed that introducing relaxations may make the linear program unbounded, and proposed algorithmic steps and structural assumptions of F and W to fight this issue. The results of these works are incomparable to ours since they focus on controlling the errors in approximating the optimal value function v^* rather than controlling the suboptimality of the policies output by the algorithm. Interestingly, the widely popular REPS algorithm of Peters et al. (2010) is also originally derived from the relaxed linear program analyzed by de Farias and Van Roy (2003), even if this connection has not been pointed out by the authors.

The work of Chen et al. (2018) is very close to ours in spirit. Chen et al. consider a variation of the relaxed saddle-point problem (2) with W being block-diagonal with F^T in each of its blocks, and claim convergence results for their algorithm to the optimal policy under only a realizability assumption. Unfortunately, their choice of W does not necessarily ensure that the coherence assumption holds, which raises concerns regarding the generality of their guarantees. Indeed, the results of Chen et al. require an additional assumption that implies that $\frac{\max_x d_\pi(x)}{\min_{x'} d_\pi(x')}$ remains bounded by a constant for any policy π, which is extremely difficult to ensure in problems of practical interest. In fact, this ratio is already exponentially large in $|\mathcal{X}|$ in very simple problems like the one we consider in our experiments. Additionally, the analysis of Chen et al. is based on the potentially erroneous claim that under the realizability assumption, the representation (u^*, y^*) of the original optimal solution $(v^*, \mu^*) = (Fu^*, W^\mathsf{T} y^*)$ always remains an optimal solution to the relaxed saddle-point problem. It is currently unclear if this claim is indeed true, or to what extent their condition regarding the boundedness of stationary distribution can be relaxed.

In any case, we believe that our coherence assumption is more fundamental than the previously considered conditions, and it enables a much more transparent analysis of optimization algorithms addressing the relaxed saddle-point problem (2). Beyond this particular positive result, our work also cleans the slate for further theoretical work on approximate optimization in Markov decision processes. Indeed, the form of our coherence assumption naturally invites the question: can we compute good approximate solutions to the original problem when our assumptions are only satisified approximately? Similar questions are not without precedent in the reinforcement-learning literature. Translated to our notation, classical results concerning the performance of (least-squares) temporal difference learning algorithms imply that the approximation errors are controlled by the projection error of $QFu^* + r$ to the column space of F (Tsitsiklis and Van Roy, 1997; Bradtke and Barto, 1996; Lazaric et al., 2010). When using more general function classes to approximate v^*, Munos and Szepesvári (2008) show that the approximation errors are controlled by the *inherent Bellman error* of the function class, which captures an approximation property related to our coherence condition. Whether or not we can generalize our techniques to construct provably efficient algorithms under such milder assumptions remains an exciting open problem that we leave open for future research.

References

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

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

D. P. Bertsekas. *Dynamic Programming and Optimal Control*, volume 2. Athena Scientific, Belmont, MA, 3 edition, 2007.

S. J. Bradtke and A. G. Barto. Linear least-squares algorithms for temporal difference learning. *Machine Learning*, 22:33–57, 1996.

S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015.

Y. Chen, L. Li, and M. Wang. Scalable bilinear π learning using state and action features. In *International Conference on Machine Learning*, pages 833–842, 2018.

D. P. de Farias and B. Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

G. de Ghellinck. Les problèmes de décisions séquentielles. *Cahiers du Centre d'Études de Recherche Opérationnelle*, 2:161–179, 1960.

E. V. Denardo. On linear programming in a markov decision problem. *Management Science*, 16(5):281–288, 1970.

E. Even-Dar, S. M. Kakade, and Y. Mansour. Online Markov decision processes. *Mathematics of Operations Research*, 34(3):726–736, 2009.

R. A. Howard. *Dynamic Programming and Markov Processes*. The MIT Press, Cambridge, MA, 1960.

V. R. Konda and J. N. Tsitsiklis. Actor-critic algorithms. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 1008–1014, Cambridge, MA, USA, 1999. MIT Press.

G. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12: 747–756, 1976.

C. Lakshminarayanan and S. Bhatnagar. A generalized reduced linear program for Markov decision processes. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, pages 2722–2728. AAAI Press, 2015.

C. Lakshminarayanan, S. Bhatnagar, and r. Szepesvári. A linearly relaxed approximate linear program for Markov decision processes. *IEEE Transactions on Automatic control*, 2018.

A. Lazaric, M. Ghavamzadeh, and R. Munos. Finite-sample analysis of LSTD. In J. Fürnkranz and T. Joachims, editors, *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, pages 615–622. Omnipress, 2010.

D. A. Levin, Y. Peres, and E. L. Wilmer. Markov chains and mixing times. 2nd edition. 2017.

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

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, and G. Ostrovski. Human-level control through deep reinforcement learning. *Nature*, 518 (7540):529–533, 2015.

R. Munos and C. Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(May):815–857, 2008.

A. Nemirovski. 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, 2004.

A. Nemirovski and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley Interscience, 1983.

G. Neu, A. György, Cs. Szepesvári, and A. Antos. Online Markov decision processes under bandit feedback. *IEEE Transactions on Automatic Control*, 59:676–691, 2014.

G. Neu, A. Jonsson, and V. Gómez. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

J. Peters, K. Mülling, and Y. Altun. Relative entropy policy search. In *Proceedings of the 25th AAAI Conference on Artificial Intelligence (AAAI-10)*, pages 1607–1612, Menlo Park, CA, USA, 2010. AAAI Press.

M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley-Interscience, April 1994.

A. Rakhlin and K. Sridharan. Online learning with predictable sequences. In *Conference on Learning Theory*, pages 993–1019, 2013a.

A. Rakhlin and K. Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems*, pages 3066–3074, 2013b.

J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1889–1897, 2015.

P. Schweitzer and A. Seidman. Generalized polynomial approximations in Markovian decision processes. *J. of Math. Anal. and Appl.*, 110:568–582, 1985.

E. Seneta. *Non-negative matrices and Markov chains*. Springer Science & Business Media, 2006.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction. 2nd edition*. 2018.

R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In S. Solla, T. Leen, and K. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 1057–1063, Cambridge, MA, USA, 1999. MIT Press.

Cs. Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

J. N. Tsitsiklis and B. Van Roy. An analysis of temporal difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42:674–690, 1997.

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

A. Zimin and G. Neu. Online learning in episodic Markovian decision processes by relative entropy policy search. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1583–1591, 2013.

Appendix A. Ommitted proofs

A.1. The proof of Lemma 4

The proof will rely on repeatedly using the so-called three-points identity that can easily be shown to hold for all points $x, y, z \in \mathcal{Z}$:

$$D_\Phi\big(x\big\|y\big) = D_\Phi(x\|z) + D_\Phi\big(z\big\|y\big) + \big\langle \nabla\Phi(y) - \nabla\Phi(z), z - x\big\rangle.$$

We first use it to show

$$\begin{aligned} D_\Phi(z\|z_{t+1}) &= D_\Phi(z\|z_t) - D_\Phi(z_{t+1}\|z_t) + \eta\,\langle z - z_{t+1}, \nabla\Phi(z_{t+1}) - \nabla\Phi(z_t)\rangle \\ &\le D_\Phi(z\|z_t) - D_\Phi(z_{t+1}\|z_t) + \eta\,\langle z - z_{t+1}, G(\widehat{z}_{t+1})\rangle\,, \end{aligned}$$

where we also used the first-order optimality condition for z_{t+1} in the second step:

$$\big\langle \nabla\Phi(z_t) - \nabla\Phi(z_{t+1}) - \eta G(\widehat{z}_{t+1}), z_{t+1} - z\big\rangle \ge 0.$$

Furthermore, we have

$$\langle z - z_{t+1}, G(\widehat{z}_{t+1})\rangle = \langle z - \widehat{z}_{t+1}, G(\widehat{z}_{t+1})\rangle + \langle \widehat{z}_t - z_{t+1}, G(\widehat{z}_{t+1})\rangle\,.$$

Using this bound together with the three-points identity

$$D_\Phi(z_{t+1}\|z_t) = D_\Phi(z_{t+1}\|\widehat{z}_{t+1}) + D_\Phi(\widehat{z}_{t+1}\|z_t) + \langle \nabla\Phi(z_t) - \nabla\Phi(\widehat{z}_{t+1}), \widehat{z}_{t+1} - z_{t+1}\rangle\,,$$

we obtain

$$\begin{aligned} D_\Phi(z\|z_{t+1}) &\le D_\Phi(z\|z_t) - D_\Phi(z_{t+1}\|z_t) + \eta\,\langle \widehat{z}_{t+1} - z_{t+1}, G(\widehat{z}_{t+1})\rangle + \eta\,\langle z - \widehat{z}_{t+1}, G(\widehat{z}_{t+1})\rangle \\ &= D_\Phi(z\|z_t) - D_\Phi(z_{t+1}\|\widehat{z}_{t+1}) - D_\Phi(\widehat{z}_{t+1}\|z_t) + \eta\,\langle z - \widehat{z}_{t+1}, G(\widehat{z}_{t+1})\rangle \\ &\qquad + \big\langle \nabla\Phi(z_t) - \nabla\Phi(\widehat{z}_{t+1}) - \eta G(\widehat{z}_{t+1}), z_{t+1} - \widehat{z}_{t+1}\big\rangle \\ &= D_\Phi(z\|z_t) - D_\Phi(z_{t+1}\|\widehat{z}_{t+1}) - D_\Phi(\widehat{z}_{t+1}\|z_t) \\ &\qquad + \big\langle \nabla\Phi(z_t) - \nabla\Phi(\widehat{z}_{t+1}) - \eta G(z_t), z_{t+1} - \widehat{z}_{t+1}\big\rangle + \eta\,\langle G(z_t) - G(\widehat{z}_{t+1}), z_{t+1} - \widehat{z}_{t+1}\rangle \\ &\qquad + \eta\,\langle z - \widehat{z}_{t+1}, G(\widehat{z}_{t+1})\rangle \\ &\le D_\Phi(z\|z_t) - D_\Phi(z_{t+1}\|\widehat{z}_{t+1}) - D_\Phi(\widehat{z}_{t+1}\|z_t) + \eta\,\langle G(z_t) - G(\widehat{z}_{t+1}), z_{t+1} - \widehat{z}_{t+1}\rangle \\ &\qquad + \eta\,\langle z - \widehat{z}_{t+1}, G(\widehat{z}_{t+1})\rangle\,, \end{aligned}$$

where the last step follows from the fact that $\widehat{z}_{t+1}$ satisfies the first-order optimality condition

$$\big\langle \nabla\Phi(z_t) - \nabla\Phi(\widehat{z}_{t+1}) - \eta G(z_t), z_{t+1} - \widehat{z}_t\big\rangle \le 0.$$

Now, using the σ-strong convexity of Φ and the L-Lipschitz continuity of F, we obtain

$$\begin{aligned} D_\Phi(z\|z_{t+1}) &\le D_\Phi(z\|z_t) - D_\Phi(z_{t+1}\|\widehat{z}_{t+1}) - D_\Phi(\widehat{z}_{t+1}\|z_t) + \eta\,\langle G(z_t) - G(\widehat{z}_{t+1}), z_{t+1} - \widehat{z}_{t+1}\rangle \\ &\qquad + \eta\,\langle z - \widehat{z}_{t+1}, G(\widehat{z}_{t+1})\rangle \\ &\le D_\Phi(z\|z_t) - \frac{\sigma}{2}\,\|z_{t+1} - \widehat{z}_{t+1}\|_2^2 - \frac{\sigma}{2}\,\|\widehat{z}_{t+1} - z_t\|_2^2 + \eta L\,\|z_t - \widehat{z}_{t+1}\|_2\,\|z_{t+1} - \widehat{z}_{t+1}\|_2 \\ &\qquad + \eta\,\langle z - \widehat{z}_{t+1}, G(\widehat{z}_{t+1})\rangle \\ &\le D_\Phi(z\|z_t) - \frac{\sigma - \eta L}{2}\,\big(\|z_{t+1} - \widehat{z}_{t+1}\|_2^2 + \|\widehat{z}_{t+1} - z_t\|_2^2\big) \\ &\qquad + \eta\,\langle z - \widehat{z}_{t+1}, G(\widehat{z}_{t+1})\rangle \\ &\le D_\Phi(z\|z_t) - \frac{\sigma - \eta L}{4}\,\|z_{t+1} - z_t\|_2^2 + \eta\,\langle z - \widehat{z}_{t+1}, G(\widehat{z}_{t+1})\rangle\,, \end{aligned}$$

where we also used the elementary inequalities $2ab \le a^2 + b^2$ and $(a+b)^2 \le 2a^2 + 2b^2$ in the last two steps, respectively. ∎

A.2. The proof of Lemma 8

To enhance readability of the proof, we will omit explicit references to T below, and will simply use π, ρ and $\overline{\mu}$ to refer to π_T, ρ_T and $\overline{\mu}_T$, respectively. Defining $\overline{d}(x) = \sum_a \overline{\mu}(x,a)$ for all x, we start by noticing that

$$\langle \overline{\mu}, r \rangle - \rho = \sum_{x,a} \left(\overline{d}(x) - d(x)\right) \pi(a|x) r(x,a) \leq \left\| \overline{d} - d \right\|_1,$$

so all we are left with is bounding the total variation distance between d and $\overline{d}$. To do this, we start by fixing an arbitrary $k > 0$ and observing that

$$\begin{aligned} \left\| \left(\overline{d} - d\right) P_\pi^k \right\|_1 &\leq C e^{-k/\tau} \left\| \overline{d} - d \right\|_1 \\ &\leq C e^{-k/\tau} \left(\left\| \overline{d} - \overline{d} P_\pi^k \right\|_1 + \left\| \overline{d} P_\pi^k - d \right\|_1 \right), \end{aligned} \tag{9}$$

where we used Assumption 1 in the first step and the triangle inequality in the second one. Regarding the first term in the parentheses, we repeatedly use the triangle inequality to obtain

$$\begin{aligned} \left\| \overline{d} - \overline{d} P_\pi^k \right\|_1 &\leq \left\| \overline{d} - \overline{d} P_\pi \right\|_1 + \left\| \overline{d} P_\pi - \overline{d} P_\pi^2 \right\|_1 + \cdots + \left\| \overline{d} P_\pi^{k-1} - \overline{d} P_\pi^k \right\|_1 \\ &= \left\| \overline{d} - \overline{d} P_\pi \right\|_1 + \left\| \left(\overline{d} - \overline{d} P_\pi\right) P_\pi \right\|_1 + \cdots + \left\| \left(\overline{d} - \overline{d} P_\pi\right) P_\pi^{k-1} \right\|_1 \\ &\leq \left\| \overline{d} - \overline{d} P_\pi \right\|_1 + C e^{-1/\tau} \left\| \overline{d} - \overline{d} P_\pi \right\|_1 + \cdots + C e^{-(k-1)/\tau} \left\| \overline{d} - \overline{d} P_\pi \right\|_1 \\ &\leq C \left\| \overline{d} - \overline{d} P_\pi \right\|_1 \sum_{i=0}^{k-1} e^{-i/\tau} \leq \frac{C}{1 - e^{-1/\tau}} \left\| \overline{d} - \overline{d} P_\pi \right\|_1. \end{aligned}$$

Plugging this bound into Equation 9 and observing that $\overline{d} P_\pi^k - d = \left(\overline{d} - d\right) P_\pi^k$ due to stationarity of d, we get

$$\left\| \left(\overline{d} - d\right) P_\pi^k \right\|_1 \leq C e^{-k/\tau} \left(\frac{C}{1 - e^{-1/\tau}} \left\| \overline{d} - \overline{d} P_\pi^k \right\|_1 + \left\| \left(\overline{d} - d\right) P_\pi^k \right\|_1 \right).$$

Reordering gives

$$\left\| \left(\overline{d} - d\right) P_\pi^k \right\|_1 \leq \frac{C e^{-k/\tau}}{1 - C e^{-k/\tau}} \cdot \frac{C}{1 - e^{-1/\tau}} \left\| \overline{d} - \overline{d} P_\pi \right\|_1.$$

Thus, using the triangle inequality again yields

$$\begin{aligned} \left\| \overline{d} - d \right\|_1 &\leq \left\| \overline{d} - \overline{d} P_\pi^k \right\|_1 + \left\| \overline{d} P_\pi^k - d \right\|_1 \\ &\leq \left(1 + \frac{C e^{-k/\tau}}{1 - C e^{-k/\tau}} \right) \frac{C}{1 - e^{-1/\tau}} \left\| \overline{d} - \overline{d} P_\pi \right\|_1. \end{aligned}$$

Now, choosing any $k \geq \tau \log(2C)$ and using the elementary inequality $1/(1 - e^{-1/\tau}) \leq \tau + 1$ concludes the proof. ■

A.3. The proof of Lemma 7

We start by noticing that the dual norm of $\|z\|^2 = \|u\|_2^2 + \|y\|_1^2$ evaluated at $x = (w, q)$ is $\|x\|_*^2 = \|w\|_2^2 + \|q\|_\infty^2$. Recalling that the smoothness of $\widetilde{\mathcal{L}}$ with respect to $\|\cdot\|$ is equivalent to the Lipschitzness of G with respect to $\|\cdot\|_*$, we will prove that $\left\|G(z) - G(z')\right\|_*^2 \leq 2\left\|z - z'\right\|^2$. Using the definition of $G(z)$, we have for any $z = (u, y)$ and $z' = (u', y')$ that

$$\left\|G(z) - G(z')\right\|_*^2 = \left\|Q^\top W^\top \left(y - y'\right)\right\|_2^2 + \left\|QF(u - u')\right\|_\infty^2$$

The first term can be bounded as

$$\left\|Q^{\mathsf{T}}W^{\mathsf{T}}(y-y')\right\|_2 \le \left\|Q^{\mathsf{T}}W^{\mathsf{T}}(y-y')\right\|_1 \le \left\|W^{\mathsf{T}}(y-y')\right\|_1 + \left\|P^{\mathsf{T}}W^{\mathsf{T}}(y-y')\right\|_1 \le 2\left\|y-y'\right\|_1,$$

where we used the fact that the rows of W form probability distributions. To bound the last term, we observe that

$$\begin{aligned}
\left\|QF(u-u')\right\|_\infty^2 &= \max_{x,a}\left|\sum_{x'}\left(\mathbb{I}_{\{x=x'\}} - P(x'|x,a)\right)\sum_i F_{x',i}\left(u_i-u_i'\right)\right|^2 \\
&\le \max_{x,a}\left(\left\|\mathbb{I}_{\{x=\cdot\}} - P(\cdot|x,a)\right\|_1 \cdot \left\|F\left(u-u'\right)\right\|_\infty\right)^2 \\
&\le 2\max_x\left|\sum_i F_{x,i}(u_i-u_i')\right|^2 \le 2\max_x\left\|F_{x,\cdot}\right\|_1^2\left\|u-u'\right\|_\infty^2 \\
&\le 2K\left\|u-u'\right\|_\infty^2 \le 2K\left\|u-u'\right\|_2^2.
\end{aligned}$$

This concludes the proof. ∎

A.4. The proof of Lemma 9

The statement is obvious when $Q^{\mathsf{T}}W^{\mathsf{T}}\overline{y}_T = 0$, so we will assume that the contrary holds below. Let us define

$$w = \tau_{\mathrm{mix}}\cdot \operatorname*{arg\,max}_{v:\|v\|_\infty=1}\left\langle Q^{\mathsf{T}}W^{\mathsf{T}}\overline{y}_T, v\right\rangle,$$

noting that $\left\langle Q^{\mathsf{T}}W^{\mathsf{T}}\overline{y}_T, w\right\rangle = \tau_{\mathrm{mix}}\left\|Q^{\mathsf{T}}W^{\mathsf{T}}\overline{y}_T\right\|_1 > 0$. By using this fact and Assumption 3, we crucially observe that there exists a $\widetilde{u}$ such that $\left\langle Q^{\mathsf{T}}W^{\mathsf{T}}\overline{y}_T, w\right\rangle = \left\langle Q^{\mathsf{T}}W^{\mathsf{T}}\overline{y}_T, F\widetilde{u}\right\rangle$ and $\|\widetilde{u}\|_\infty \le \tau_{\mathrm{mix}}U$. This implies that we can apply Corollary 6 with $z = (F\overline{u}_T - F\widetilde{u}, W^{\mathsf{T}}\overline{y}_T)$ to obtain the bound

$$\left\langle Q^{\mathsf{T}}W^{\mathsf{T}}\overline{y}_T, w\right\rangle = \left\langle Q^{\mathsf{T}}W^{\mathsf{T}}\overline{y}_T, F\overline{u}_T\right\rangle + \left\langle W^{\mathsf{T}}\overline{y}_T, r\right\rangle - \left\langle Q^{\mathsf{T}}W^{\mathsf{T}}\overline{y}_T, F\left(\overline{u}_T-\widetilde{u}\right)\right\rangle - \left\langle W^{\mathsf{T}}\overline{y}_T, r\right\rangle \le \frac{D_\Phi(z\|z_0)}{\eta T}.$$

Plugging in the definition of w and the Bregman divergence D_Φ, we obtain

$$\left\|Q^{\mathsf{T}}W^{\mathsf{T}}\overline{y}_T\right\|_1 \le \frac{\frac{1}{2}\left\|\widetilde{u}-\overline{u}_T\right\|_2^2 + \log M}{\eta\tau_{\mathrm{mix}}T}.$$

Due to Assumption 2 and our assumption on F stated before Theorem 3, we can choose an optimal solution u^* satisfying $Fu^* = v^*$ and $\|u^*\|_\infty \le \tau_{\mathrm{mix}}U$ and write

$$\begin{aligned}
\left\|\widetilde{u}-\overline{u}_T\right\|_2^2 &\le 2\left\|\widetilde{u}-u^*\right\|_2^2 + 2\left\|\overline{u}_T-u^*\right\|_2^2 \le 4\left\|\widetilde{u}\right\|_2^2 + 4\left\|u^*\right\|_2^2 + 4D_\Phi\left(z^*\middle\|\overline{z}_T\right) \\
&\le 4N\left\|\widetilde{u}\right\|_\infty^2 + 4N\left\|u^*\right\|_\infty^2 + 4D_\Phi\left(z^*\middle\|z_0\right) \\
&\le 10\tau_{\mathrm{mix}}^2U^2N + 4\log M,
\end{aligned}$$

where in the second line we have used Corollary 5 that implies $D_\Phi\left(z^*\middle\|\overline{z}_T\right) \le D_\Phi(z^*\|z_0)$. ∎