Dealing with unbounded gradients in stochastic saddle-point optimization

Gergely Neu [1] **Nneka Okolo** [1]

Abstract

We study the performance of stochastic first-order methods for finding saddle points of convex-concave functions. A notorious challenge faced by such methods is that the gradients can grow arbitrarily large during optimization, which may result in instability and divergence. In this paper, we propose a simple and effective regularization technique that stabilizes the iterates and yields meaningful performance guarantees even if the domain and the gradient noise scales linearly with the size of the iterates (and is thus potentially unbounded). Besides providing a set of general results, we also apply our algorithm to a specific problem in reinforcement learning, where it leads to performance guarantees for finding near-optimal policies in an average-reward MDP without prior knowledge of the bias span.

1. Introduction

We study the performance of stochastic optimization algorithms for solving convex-concave saddle-point problems of the form $\min_{\boldsymbol{x}\in\mathcal{X}} \max_{\boldsymbol{y}\in\mathcal{Y}} f(\boldsymbol{x},\boldsymbol{y})$. The algorithms we consider aim to approximate saddle points via running two stochastic convex optimization methods against each other, one aiming to minimize the objective function and the other aiming to maximize. Both players have access to noisy gradient evaluations at individual points $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ of the primal and dual domains $\mathcal{X}$ and $\mathcal{Y}$, and typically compute their updates via gradient-descent-like procedures. Due to the complicated interaction between the two concurrent procedures, it is notoriously difficult to ensure convergence of these methods towards the desired saddle points, and in fact even guaranteeing their stability is far from trivial. One common way to make sure that the iterates do not diverge is projecting them to bounded sets around the initial point. While this idea does the job, it gives rise to a dilemma: how should one pick the size of these constraint sets to make sure both that the optimum remains in there while keeping the optimization process reasonably efficient? In this paper, we propose a method that addresses this question and provides as good guarantees as the best known projection-based method, but without having to commit to a specific projection radius.

It is well known that simply running gradient descent for both the minimizing and maximizing players can easily result in divergence, even when having access to exact gradients without noise (Goodfellow, 2016; Mertikopoulos et al., 2018). While the average of the iterates may converge in such cases, their rate of convergence is typically affected by the magnitude of the gradients, which grows larger and larger as the iterates themselves diverge, thus resulting in arbitrarily slow convergence of the average. Numerous solutions have been proposed to this issue in the literature, most notably using some form of *gradient extrapolation* (Korpelevich, 1976; Popov, 1980; Gidel et al., 2018; Mertikopoulos et al., 2018). When these methods have access to noiseless gradients and are run on smooth objectives, these methods are remarkably stable: they can be shown to converge monotonically towards their limit. That said, convergence of such methods in the stochastic case is much less well-understood, unless the iterates are projected to a compact set (Juditsky et al., 2011; Gidel et al., 2018), or the boundedness of the gradients ensured by other assumptions (Mishchenko et al., 2020; Loizou et al., 2021; Sadiev et al., 2023). Indeed, unless projections are employed, the iterates of one player may grow large, which can result in large gradients observed by the opposite player, which in turn may result in large iterates for the second player—which effects may eventually cascade and result in instability and divergence. Our main contribution is proposing a stabilization technique that eliminates the risk of divergence.

For the sake of exposition, let us consider the special case of bilinear objectives

$$f(\boldsymbol{x},\boldsymbol{y}) = \boldsymbol{x}^\mathsf{T}\boldsymbol{M}\boldsymbol{y} + \boldsymbol{b}^\mathsf{T}\boldsymbol{x} - \boldsymbol{c}^\mathsf{T}\boldsymbol{y},$$

and primal-dual gradient descent starting from the initial point $\boldsymbol{x}_1 = 0$ and $\boldsymbol{y} = 0$ as a baseline:

$$\begin{aligned}\boldsymbol{x}_{t+1} &= \boldsymbol{x}_t - \eta\nabla_x f(\boldsymbol{x}_t,\boldsymbol{y}_t) = \boldsymbol{x}_t - \eta\boldsymbol{g}_x(t)\\ \boldsymbol{y}_{t+1} &= \boldsymbol{y}_t - \eta\nabla_y f(\boldsymbol{x}_t,\boldsymbol{y}_t) = \boldsymbol{y}_t - \eta\boldsymbol{g}_y(t),\end{aligned}$$

[1] Universitat Pompeu Fabra, Barcelona, Spain. Correspondence to: Gergely Neu <gergely.neu@gmail.com>, Nneka Okolo <nnekamaureen.okolo@upf.edu>.

where we also introduced the shorthand notations $\boldsymbol{g}_x(t) = \nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t)$ and $\boldsymbol{g}_y(t) = \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)$. Using standard tools (that we will explain in detail below), the average of the first T iterates produced by the above procedure can be shown to satisfy the following guarantee on the *duality gap*:

$$\begin{aligned} G(\boldsymbol{x}^*, \boldsymbol{y}^*) &= f\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{x}_t, \boldsymbol{y}^*\right) - f\left(\boldsymbol{x}^*, \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{y}_t\right) \\ &\leq \frac{\|\boldsymbol{x}^*\|_2^2 + \|\boldsymbol{y}^*\|_2^2}{2\eta T} + \frac{\eta}{2T}\sum_{t=1}^{T}\left(\|\boldsymbol{g}_x(t)\|_2^2 + \|\boldsymbol{g}_y(t)\|_2^2\right). \end{aligned}$$

If one can ensure that the gradients $\boldsymbol{g}_x(t)$ and $\boldsymbol{g}_y(t)$ remain bounded by a constant $G > 0$, one can set $\eta \sim 1/(G\sqrt{T})$ and obtain a convergence rate of order $\frac{G(\|\boldsymbol{x}^*\|_2^2 + \|\boldsymbol{y}^*\|_2^2)}{\sqrt{T}}$. However, notice that *there is no way to make sure that the gradients actually remain bounded* while executing the algorithm! Indeed, notice that $\nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t) = \boldsymbol{M}\boldsymbol{y}_t + \boldsymbol{b}$, which grows large as $\boldsymbol{y}_t$ grows large. A natural idea is to project the iterates to balls of respective sizes $D_x,\ D_y > 0$, which guarantees that the iterates and thus the gradients remain bounded. However, convergence to the saddle point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ is now only possible whenever the respective norms satisfy $\|\boldsymbol{x}^*\| \leq D_x$ and $\|\boldsymbol{y}^*\| \leq D_y$, otherwise the optimal solution is excluded from the feasible set. Unfortunately, in many applications, it is impossible to pick the constants D_x and D_y appropriately due of lack of prior knowledge of the solution norms. We give an important example of such a situation at the end of this section.

In this paper, we propose a method that guarantees upper bounds on the duality gap of the following form (when specialized to the setting described above):

$$G(\boldsymbol{x}^*, \boldsymbol{y}^*) = \mathcal{O}\left(\frac{\|\boldsymbol{x}^* - \boldsymbol{x}_1\|_2^2 + \|\boldsymbol{y}^* - \boldsymbol{y}_1\|_2^2 + 1}{\sqrt{T}}\right),$$

where the big-O notation hides some problem-dependent constants related to the objective function f (which will be made explicit in our main theorem). Notably, our method requires no prior knowledge of the norms $\|\boldsymbol{x}^* - \boldsymbol{x}_1\|_2^2$ and $\|\boldsymbol{y}^* - \boldsymbol{y}_1\|_2^2$ whatsoever, and in particular performs no projections to make sure that the iterates remain bounded, thus addressing the challenge outlined above. As we will show, this guarantee holds even when the gradients are subject to *multiplicative noise* that can scale with the magnitude of the gradient itself. Our main technical tool is augmenting the objective with a well-chosen regularization term which allows us to eliminate the terms $\|\boldsymbol{g}_x(t)\|_2^2 + \|\boldsymbol{g}_y(t)\|_2^2$ appearing in the guarantee of standard primal-dual gradient descent, and replace them with an upper bound of the gradients of the objective evaluated at the initial point $(\boldsymbol{x}_1, \boldsymbol{y}_1)$. These bounds have the appealing property of being *initialization-dependent*, in that they guarantee improved performance when we pick the initial points $(\boldsymbol{x}_1, \boldsymbol{y}_1)$ close to $(\boldsymbol{x}^*, \boldsymbol{y}^*)$.

Our contributions are closely related to the work of Liu & Orabona (2022), who propose algorithms that are guaranteed to achieve an initialization-dependent convergence rate of the order $G(\boldsymbol{x}^*, \boldsymbol{y}^*) = \widetilde{\mathcal{O}}\left(\frac{\|\boldsymbol{x}^* - \boldsymbol{x}_1\|_2 + \|\boldsymbol{y}^* - \boldsymbol{y}_1\|_2}{\sqrt{T}}\right)$, where $\widetilde{\mathcal{O}}(\cdot)$ hides polylogarithmic factors of T and the solution norms. When compared to our main result, this bound demonstrates an improved scaling with the initialization error. However, their result is proved under the condition that all gradients remain bounded: $\|\boldsymbol{g}_x(t)\|_2^2 \leq 1$ and $\|\boldsymbol{g}_y(t)\|_2^2 \leq 1$. As per the above discussion, this is only possible in general when projecting the iterates to a bounded subset of the domain, which requires prior knowledge of the norms $\|\boldsymbol{x}^*\|$ and $\|\boldsymbol{y}^*\|$. When accounting for this issue and adapting the method of Liu & Orabona (2022) to our setting, their bounds end up scaling with $(D_x + D_y)(\|\boldsymbol{x}^* - \boldsymbol{x}_1\|_2 + \|\boldsymbol{y}^* - \boldsymbol{y}_1\|_2)$, which is worse than the scaling of our bound. Furthermore, the results of Liu & Orabona (2022) are only proved for noiseless gradients (although we believe that this specific restriction may not be hard to address). On the positive side, this assumption allows their algorithm to achieve very fast convergence when initialized very close to the optimum—which is generally not possible in the stochastic case we consider. Taking all this into account, we regard our guarantees as significantly different from their results, and in many senses a strict improvement.

More broadly speaking, our work contributes to the line of work on *parameter-free* optimization methods that are able to adapt to problem complexity without prior knowledge of the relevant problem parameters. In the context of online convex optimization (OCO), several effective parameter-free algorithms are known to achieve guarantees scaling optimally with the initialization error $\|\boldsymbol{x}^* - \boldsymbol{x}_1\|$, without requiring prior knowledge thereof (Streeter & McMahan, 2012; Orabona, 2013; van der Hoeven, 2019). (Cutkosky & Boahen, 2016; Cutkosky, 2019; Mhammedi & Koolen, 2020) improve these guarantees by providing initialization-adaptive bounds for OCO in unconstrained domains without prior knowledge of both the size of the domain or subgradients of the loss. One would think that this would make their method suitable for solving the problem we study in this paper—however, their bound depends on the maximum norm of the observed sub-gradients, which is problematic for the reasons we have discussed extensively above.

Unconstrained saddle-point problems have many important applications. Perhaps the most well-known such applications is in optimizing dual representations of convex functions (Bubeck et al., 2015, Section 5.2,Shalev-Shwartz & Singer, 2006; Wang & Abernethy, 2018; Wang et al., 2023). Our original motivation during the development of this work has been to develop primal-dual methods for solving average-reward Markov decision processes (MDPs): this problem can be formulated as a linear program with primal variables that are of unknown scale. In the simpler setting of discounted Markov decision processes, previous work has provided efficient planning methods based on saddle-point optimization (Wang, 2017; Jin & Sidford, 2020; Cheng et al., 2020). While in this simple setting the primal variables (called *value functions* in this setting) are known to be uniformly bounded, this is not the case in the more challenging average-reward setting we consider here: in this case, the value functions can have arbitrarily large norm depending on the program structure. As it is well-known in the reinforcement-learning literature, estimating this parameter is as hard as solving the original problem, and learning optimal policies without its prior knowledge has been widely conjectured to be impossible (Bartlett & Tewari, 2009; Fruit et al., 2018b;a; Zhang & Ji, 2019). Using our techniques developed in the present paper, we make progress on this important problem by proposing a planning algorithm that is guaranteed to produce a near-optimal policy without having prior knowledge of the scale of the value functions after a polynomial number of queries made to a simulator of the environment.

Notations. For an integer T, we use $[T] = 1, 2, \cdots, T$. We denote as $\mathbf{1}$ the vector with all one entries in $\mathbb{R}^m$ and represent the positive orthant as $\mathbb{R}^n_+$. Let $\mathcal{X} \subseteq \mathbb{R}^m$ and $f : \mathcal{X} \to \mathbb{R}$ differentiable. For vectors $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ we define their inner product as $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle = \sum_{i=1}^m \boldsymbol{x}_i \boldsymbol{x}'_i$ and the *Bregman divergence* of $\boldsymbol{x}$ at $\boldsymbol{x}'$ induced by $f : \mathcal{X} \to \mathbb{R}$ as

$$\mathcal{D}_x(\boldsymbol{x} \| \boldsymbol{x}') = f(\boldsymbol{x}) - f(\boldsymbol{x}') - \langle \nabla f(\boldsymbol{x}'), \boldsymbol{x} - \boldsymbol{x}' \rangle .$$

2. Preliminaries

We now formally define our problem setup and objectives. First, we recall some standard definitions.

Definition 2.1. **(convex-concave function)** Let $\mathcal{X} \subseteq \mathbb{R}^m, \mathcal{Y} \subseteq \mathbb{R}^n$ be convex sets. A function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is said to be *convex-concave* if it is convex in the first argument and concave in the second. That is, f is convex-concave if for any $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$, $\boldsymbol{y}, \boldsymbol{y}' \in \mathcal{Y}$ and $\lambda \in [0, 1]$, we have

$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{x}', \boldsymbol{y}) \leq \lambda f(\boldsymbol{x}, \boldsymbol{y}) + (1-\lambda) f(\boldsymbol{x}', \boldsymbol{y}),$$

and

$$f(\boldsymbol{x}, \lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{y}') \geq \lambda f(\boldsymbol{x}, \boldsymbol{y}) + (1-\lambda) f(\boldsymbol{x}, \boldsymbol{y}').$$

Definition 2.2. **(subgradient and subdifferential)** Let $\mathcal{X}^*$ denote the dual space of $\mathcal{X}$. For a function $h : \mathcal{X} \to \mathbb{R}$, $\boldsymbol{g} \in \mathcal{X}^*$ is a subgradient of h at $\boldsymbol{x} \in \mathcal{X}$ if for all $\boldsymbol{x}' \in \mathcal{X}$,

$$h(\boldsymbol{x}) - h(\boldsymbol{x}') \leq \langle \boldsymbol{g}, \boldsymbol{x} - \boldsymbol{x}' \rangle .$$

The set of all subgradients of a function h at $\boldsymbol{x}$ is called the *subdifferential*, and is denoted by $\partial h(\boldsymbol{x})$.

We recall that when h is convex and differentiable, then $\nabla h(\boldsymbol{x}) \in \partial h(\boldsymbol{x})$ holds for all $\boldsymbol{x} \in \mathcal{X}$, and additionally $\partial h(\boldsymbol{x}) = \{\nabla h(\boldsymbol{x})\}$ holds whenever $\boldsymbol{x}$ is in the interior of the domain $\mathcal{X}$.

Definition 2.3. **(strong convexity)** For $\gamma \geq 0$, a function $h : \mathcal{X} \to \mathbb{R}$ is γ-strongly convex with respect to the norm $\|\cdot\|$ if and only if for all $\boldsymbol{x}, \boldsymbol{x}' \in \operatorname{dom}(h)$, $\boldsymbol{g} \in \partial h(\boldsymbol{x})$:

$$h(\boldsymbol{x}') - h(\boldsymbol{x}) \geq \langle \boldsymbol{g}, \boldsymbol{x}' - \boldsymbol{x} \rangle + \frac{\gamma}{2} \|\boldsymbol{x}' - \boldsymbol{x}\|^2 .$$

We consider the problem of finding (approximate) saddle points of convex-concave functions on the potentially unbounded convex domains $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^m \times \mathbb{R}^n$:

$$\inf_{\boldsymbol{x} \in \mathcal{X}} \sup_{\boldsymbol{y} \in \mathcal{Y}} f(\boldsymbol{x}, \boldsymbol{y}), \tag{1}$$

where $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is assumed to be convex-concave in the sense of Definition 2.1. We focus on the classic stochastic first-order oracle model where algorithms can only access noisy estimates of the subgradients at individual points in $\mathcal{X} \times \mathcal{Y}$. Specifically, we will consider incremental algorithms that produce a sequence of iterates $(\boldsymbol{x}_t, \boldsymbol{y}_t)_{t=1}^T$ by running two concurrent online learning methods for choosing the two sequences $\{\boldsymbol{x}_t\}_{t=1}^T$ and $\{\boldsymbol{y}_t\}_{t=1}^T$. The algorithm picking the sequence $\{\boldsymbol{x}_t\}_{t=1}^T$ aims to minimize the sequence of losses $\{f(\cdot, \boldsymbol{y}_t)\}_{t=1}^T$ and is referred to as the *min player*, and the algorithm picking $\{\boldsymbol{y}_t\}_{t=1}^T$ that aims to minimize $\{-f(\boldsymbol{x}_t, \cdot)\}_{t=1}^T$ is called the *max player*. In each round t, the two players have access to a stochastic first-order oracle that provides the following noisy estimates of a pair of subgradients $\boldsymbol{g}_x(t) \in \partial_x f(\boldsymbol{x}_t, \boldsymbol{y}_t)$ and $\boldsymbol{g}_y(t) \in -\partial_y(-f(\boldsymbol{x}_t, \boldsymbol{y}_t))$, with the noisy estimates written as

$$\widetilde{\boldsymbol{g}}_x(t) = \boldsymbol{g}_x(t) + \xi_x(t)$$
$$\widetilde{\boldsymbol{g}}_y(t) = \boldsymbol{g}_y(t) + \xi_y(t).$$

Here, $\boldsymbol{\xi}_x(t) \in \mathbb{R}^m$ and $\boldsymbol{\xi}_y(t) \in \mathbb{R}^n$ are zero-mean noise vectors generated in round $t \in [T]$ from some unknown distributions, independently of the interaction history $\mathcal{F}_{t-1}$. Using the notation $\mathbb{E}_t[\boldsymbol{x}_t] = \mathbb{E}[\boldsymbol{x}_t | \mathcal{F}_{t-1}] = \boldsymbol{x}_t$ to denote expectations conditioned on the history of observations up to the end of time t, we can write the above conditions as $\mathbb{E}_t[\widetilde{\boldsymbol{g}}_x(t)] = \boldsymbol{g}_x(t)$ and $\mathbb{E}_t[\widetilde{\boldsymbol{g}}_y(t)] = \boldsymbol{g}_y(t)$. Note that when the objective is differentiable we can simply set $\boldsymbol{g}_x(t) = \nabla_x f(\boldsymbol{x}_t, \boldsymbol{y}_t)$ and $\boldsymbol{g}_y(t) = \nabla_y f(\boldsymbol{x}_t, \boldsymbol{y}_t)$.

In a good part of this work, we focus on the important class of *bilinear* objective functions that take the following form:

$$f(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^\mathsf{T} \boldsymbol{M} \boldsymbol{y} + \boldsymbol{b}^\mathsf{T} \boldsymbol{x} - \boldsymbol{c}^\mathsf{T} \boldsymbol{y}.$$

Here, $\boldsymbol{M} \in \mathbb{R}^{m \times n}$, $\boldsymbol{b} \in \mathbb{R}^m$ and $\boldsymbol{c} \in \mathbb{R}^n$ and the domains for the optimization variables are $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{Y} = \mathbb{R}^n$. This objective is clearly differentiable, and its gradients with respect to $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ are given as $\boldsymbol{g}_x(t) = \boldsymbol{M}\boldsymbol{y}_t + \boldsymbol{b}$ and $\boldsymbol{g}_y(t) = \boldsymbol{M}^\intercal\boldsymbol{x}_t - \boldsymbol{c}$. In the context of bilinear games, we will consider a natural noise model where in each round t, we have access to noisy versions of the matrices and vectors necessary for computing the gradients. Specifically, we have $\widehat{\boldsymbol{M}}(t) = \boldsymbol{M} + \boldsymbol{\xi}_{\boldsymbol{M}}(t)$, $\widehat{\boldsymbol{b}}(t) = \boldsymbol{b} + \boldsymbol{\xi}_b(t)$, and $\widehat{\boldsymbol{c}}(t) = \boldsymbol{c} + \boldsymbol{\xi}_c(t)$ where $\boldsymbol{\xi}_{\boldsymbol{M}}(t), \boldsymbol{\xi}_b(t), \boldsymbol{\xi}_c(t)$ are $i.i.d$, zero-mean random matrices and vectors generated from unknown distributions. We then use these observations to build the following estimators for the gradients:

$$\widetilde{\boldsymbol{g}}_x(t) = \widehat{\boldsymbol{M}}(t)\boldsymbol{y}_t + \widehat{\boldsymbol{b}}(t)$$
$$\widetilde{\boldsymbol{g}}_y(t) = \widehat{\boldsymbol{M}}(t)^\intercal\boldsymbol{x}_t - \widehat{\boldsymbol{c}}(t).$$

This fits into the generic noise model defined earlier with $\boldsymbol{\xi}_x(t) = \boldsymbol{\xi}_{\boldsymbol{M}}(t)\boldsymbol{y}_t + \boldsymbol{\xi}_b(t)$ and $\boldsymbol{\xi}_y(t) = \boldsymbol{\xi}_M(t)^\intercal\boldsymbol{x}_t - \boldsymbol{\xi}_c(t)$. Regarding the magnitude of the noise, we will make the assumption that there exists constants L_M, L_b and L_c such that $\mathbb{E}_t\left[\|\widehat{\boldsymbol{b}}_t\|^2\right] \le L_b$, $\mathbb{E}_t\left[\|\widehat{\boldsymbol{c}}_t\|_2^2\right] \le L_c$, and

$$\mathbb{E}_t\left[\left\|\widehat{\boldsymbol{M}}(t)\boldsymbol{y}\right\|_2^2\right] \le L_M^2\|\boldsymbol{y}\|_2^2$$
$$\mathbb{E}_t\left[\left\|\widehat{\boldsymbol{M}}(t)^\intercal\boldsymbol{x}\right\|_2^2\right] \le L_M^2\|\boldsymbol{x}\|_2^2,$$

holds for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X} \times \mathcal{Y}$. Note that this latter assumption is satisfied whenever the operator norm of each $\widehat{\boldsymbol{M}}(t)$ is upper bounded by L_M with probability one.

This noise model is often more realistic than simply assuming that $\boldsymbol{\xi}_x(t)$ and $\boldsymbol{\xi}_y(t)$ have uniformly bounded norms, and is much more challenging to work with: notably, these noise variables scale with the iterates $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$, and may thus grow uncontrollably as the iterates grow large. In particular, the noise is precisely of this form in our application to reinforcement learning presented in Section 4.

The final output of the algorithm will be denoted as $(\overline{\boldsymbol{x}}_T, \overline{\boldsymbol{y}}_T)$, and due to noise in the gradients, its quality will be measured in terms of the *expected duality gap*, defined with respect to a comparator point $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ as

$$\mathbb{E}\left[G(\boldsymbol{x}^*, \boldsymbol{y}^*)\right] = \mathbb{E}\left[f(\overline{\boldsymbol{x}}_T, \boldsymbol{y}^*) - f(\boldsymbol{x}^*, \overline{\boldsymbol{y}}_T)\right].$$

Here, it is typical to choose as comparator point a saddle point of f that satisfies the inequalities

$$f(\boldsymbol{x}^*, \boldsymbol{y}) \le f(\boldsymbol{x}^*, \boldsymbol{y}^*) \le f(\boldsymbol{x}, \boldsymbol{y}^*). \tag{2}$$

for all $\boldsymbol{x} \in \mathcal{X}, \boldsymbol{y} \in \mathcal{Y}$. Our analysis will allow proving guarantees against arbitrary comparator points, which is useful for certain applications (cf. Section 4).

3. Algorithm and main results

We now present our algorithmic approach and provide its performance guarantees. For didactic purposes, we start with the special case of bilinear games and Euclidean geometries, and then later provide an extension to sub-bilinear objectives and more general geometries in Section 3.2.

3.1. Unconstrained bilinear games

As a gentle start, we first describe our approach for bilinear games as defined in Section 2 where the domains are $\mathcal{X} = \mathbb{R}^m$ and $\mathcal{Y} = \mathbb{R}^n$, and distances are measured in terms of the Euclidean distances in the respective spaces. For this case, the core idea of our approach is to run stochastic gradient descent/ascent to compute the iterates of the two players. As discussed before, this procedure may diverge and produce large gradients when run on the original objective, unless the iterates are projected to a bounded set. Our key idea is to replace the projection set with an appropriately chosen regularization term added to the objective. Precisely, we introduce the regularization functions $\mathcal{H}_x(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_1\|_2^2$ and $\mathcal{H}_y(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{y}_1\|_2^2$, and define our algorithm via the following recursive updates:

$$\boldsymbol{x}_{t+1} = \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathbb{R}^m}\left\{\left\langle\boldsymbol{x}, \widetilde{\boldsymbol{g}}_x(t)\right\rangle + \varrho_x\mathcal{H}_x(\boldsymbol{x}) + \frac{1}{\eta_x}\|\boldsymbol{x} - \boldsymbol{x}_t\|_2^2\right\}$$
$$\boldsymbol{y}_{t+1} = \operatorname*{arg\,min}_{\boldsymbol{y}\in\mathbb{R}^n}\left\{-\left\langle\boldsymbol{y}, \widetilde{\boldsymbol{g}}_y(t)\right\rangle + \varrho_y\mathcal{H}_y(\boldsymbol{y}) + \frac{1}{\eta_y}\|\boldsymbol{y} - \boldsymbol{y}_t\|_2^2\right\}.$$

For each player, the update rules can be recognized as an instance of Composite Objective MIrror Descent (**COMID**, Duchi et al., 2010), and accordingly we refer to the resulting algorithm as Composite Objective Gradient Descent-Ascent (**COGDA**). The updates can be written in closed form as

$$\begin{aligned}\boldsymbol{x}_{t+1} &= \frac{\boldsymbol{x}_t - \eta_x\widetilde{\boldsymbol{g}}_x(t)}{1 + \varrho_x\eta_x} + \frac{\varrho_x\eta_x\boldsymbol{x}_1}{1 + \varrho_x\eta_x} \\ \boldsymbol{y}_{t+1} &= \frac{\boldsymbol{y}_t + \eta_y\widetilde{\boldsymbol{g}}_y(t)}{1 + \varrho_y\eta_y} + \frac{\varrho_y\eta_y\boldsymbol{y}_1}{1 + \varrho_y\eta_y}.\end{aligned} \tag{3}$$

This expression has a clear intuitive interpretation: for the min-player, it is a convex combination of the standard SGD update $\boldsymbol{x}_t - \eta_x\widetilde{\boldsymbol{g}}_x(t)$ and the initial point $\boldsymbol{x}_1$, with weights that depend on the regularization parameter ϱ_x. Setting $\varrho_x = 0$ recovers the standard SGD update and makes the algorithm vulnerable to divergence issues. The overall method closely resembles the *stabilized online mirror descent* method of Fang et al. (2022), and we will accordingly refer to the effect of the newly introduced regularization term as *stabilization*.

After running the above iterations for T steps, the algorithm outputs $\overline{\boldsymbol{x}}_T = \frac{1}{T}\sum_{t=1}^T \boldsymbol{x}_t$ and $\overline{\boldsymbol{y}}_T = \frac{1}{T}\sum_{t=1}^T \boldsymbol{y}_t$. The following theorem is our main result regarding the performance of this algorithm.

Theorem 3.1. *Let $\varrho_y = 2\eta_x L_M^2$ and $\varrho_x = 2\eta_y L_M^2$. Then, the duality gap achieved by* **COGDA** *satisfies the following* *bound against any comparator* $(\boldsymbol{x}^*, \boldsymbol{y}^*) \in \mathbb{R}^m \times \mathbb{R}^n$*:*

$$\begin{aligned}\mathbb{E}\left[G(\boldsymbol{x}^*;\boldsymbol{y}^*)\right] \leq{}& \left(\frac{1}{2\eta_y T} + \eta_x L_M^2\right)\left\|\boldsymbol{y}^* - \boldsymbol{y}_1\right\|_2^2 \\ &+ \left(\frac{1}{2\eta_x T} + \eta_y L_M^2\right)\left\|\boldsymbol{x}^* - \boldsymbol{x}_1\right\|_2^2 \\ &+ \frac{\eta_y}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\widehat{\boldsymbol{M}}(t)^{\mathsf{T}}\boldsymbol{x}_1 - \widehat{\boldsymbol{c}}(t)\right\|_2^2\right] \\ &+ \frac{\eta_x}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\widehat{\boldsymbol{M}}(t)\boldsymbol{y}_1 + \widehat{\boldsymbol{b}}(t)\right\|_2^2\right]\end{aligned}$$

In particular, setting $\boldsymbol{x}_1 = 0$ *and* $\boldsymbol{y}_1 = 0$ *and* $\eta_x = 1/L_M\sqrt{T}$ *and* $\eta_y = 1/L_M\sqrt{T}$*, the duality gap is upper bounded as*

$$\mathbb{E}\left[G(\boldsymbol{x}^*, \boldsymbol{y}^*)\right] = \mathcal{O}\left(\frac{L_M^2\left(\left\|\boldsymbol{y}^*\right\|_2^2 + \left\|\boldsymbol{x}^*\right\|_2^2\right) + L_b^2 + L_c^2}{L_M\sqrt{T}}\right).$$

It is insightful to compare this bound side by side with the one we would get by running primal-dual stochastic gradient without regularization. By standard arguments (see, e.g., Liu & Orabona, 2022; Abernethy et al., 2018; Zinkevich, 2003), the following bound is easy to prove:

$$\begin{aligned}\mathbb{E}\left[G(\boldsymbol{x}^*;\boldsymbol{y}^*)\right] \leq{}& \frac{\left\|\boldsymbol{x}^* - \boldsymbol{x}_1\right\|_2^2}{\eta_x T} + \frac{\left\|\boldsymbol{y}^* - \boldsymbol{y}_1\right\|_2^2}{2\eta_y T} \\ &+ \frac{\eta_x}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\widehat{\boldsymbol{M}}(t)\boldsymbol{y}_t + \widehat{\boldsymbol{b}}(t)\right\|_2^2\right] \\ &+ \frac{\eta_y}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\widehat{\boldsymbol{M}}(t)^{\mathsf{T}}\boldsymbol{x}_t - \widehat{\boldsymbol{c}}(t)\right\|_2^2\right].\end{aligned}$$

A major problem with this bound is that it features the squared stochastic gradient norms evaluated at $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$, which are generally unbounded, which makes this guarantee void of meaning without projecting the updates. Our own guarantee stated above replaces these gradient norms with the norms of the gradients evaluated at *the initial point* $\boldsymbol{x}_1, \boldsymbol{y}_1$, which is *always bounded* irrespective of how large the actual iterates $\boldsymbol{x}_t, \boldsymbol{y}_t$ get.

At first, it may seem surprising that such an improvement is possible to achieve by such a simple regularization trick. To provide some insight about how regularization helps us achieve our goal, we provide the brief proof sketch of the above statement here.

Proof sketch of Theorem 3.1. Fix $(\boldsymbol{x}^*, \boldsymbol{y}^*) \in \mathbb{R}^m \times \mathbb{R}^n$. As the first step, we introduce the notation $f^{(\text{reg})}(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}, \boldsymbol{y}) + \frac{\varrho_x}{2}\left\|\boldsymbol{x} - \boldsymbol{x}_1\right\|_2^2 - \frac{\varrho_y}{2}\left\|\boldsymbol{y} - \boldsymbol{y}_1\right\|_2^2$ and rewrite the expected duality gap as

$$\begin{aligned}\mathbb{E}\left[G(\boldsymbol{x}^*;\boldsymbol{y}^*)\right] \leq{}& \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f^{(\text{reg})}(\boldsymbol{x}_t, \boldsymbol{y}^*) - f^{(\text{reg})}(\boldsymbol{x}^*, \boldsymbol{y}_t)\right] \\ &+ \frac{\varrho_x}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{x}^* - \boldsymbol{x}_1\right\|_2^2 - \left\|\boldsymbol{x}_t - \boldsymbol{x}_1\right\|_2^2\right] \\ &+ \frac{\varrho_y}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{y}^* - \boldsymbol{y}_1\right\|_2^2 - \left\|\boldsymbol{y}_t - \boldsymbol{y}_1\right\|_2^2\right].\end{aligned}$$

The first term in this decomposition then can be further written as the sum of *regrets* of the min and the max players:

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f^{(\text{reg})}(\boldsymbol{x}_t, \boldsymbol{y}^*) - f^{(\text{reg})}(\boldsymbol{x}^*, \boldsymbol{y}_t)\right] \\ &\quad= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f^{(\text{reg})}(\boldsymbol{x}_t, \boldsymbol{y}^*) - f^{(\text{reg})}(\boldsymbol{x}_t, \boldsymbol{y}_t)\right] \\ &\quad\quad+ \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f^{(\text{reg})}(\boldsymbol{x}_t, \boldsymbol{y}_t) - f^{(\text{reg})}(\boldsymbol{x}^*, \boldsymbol{y}_t)\right]\end{aligned}$$

These terms can then be controlled via the standard regret analysis of **`COMID`** due to Duchi et al. (2010). In particular, a few lines of calculations (along the lines of the online gradient descent analysis of Zinkevich, 2003) yield the following bound on the sum of the two regrets:

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f^{(\text{reg})}(\boldsymbol{x}_t, \boldsymbol{y}^*) - f^{(\text{reg})}(\boldsymbol{x}^*, \boldsymbol{y}_t)\right] \\ &\quad\leq \frac{\left\|\boldsymbol{y}^* - \boldsymbol{y}_1\right\|_2^2}{2\eta_y T} + \frac{\eta_y}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_y(t)\right\|_2^2\right] \\ &\quad\quad+ \frac{\left\|\boldsymbol{x}^* - \boldsymbol{x}_1\right\|_2^2}{2\eta_x T} + \frac{\eta_x}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_x(t)\right\|_2^2\right]\end{aligned}$$

Recalling the form of the gradient estimators, we note that

$$\begin{aligned}&\mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_x(t)\right\|_2^2\right] = \mathbb{E}\left[\left\|\widehat{\boldsymbol{M}}(t)\boldsymbol{y}_t + \widehat{\boldsymbol{b}}(t)\right\|_2^2\right] \\ &\leq 2\mathbb{E}\left[\left\|\widehat{\boldsymbol{M}}(t)\left(\boldsymbol{y}_t - \boldsymbol{y}_1\right)\right\|_2^2\right] + 2\mathbb{E}\left[\left\|\widehat{\boldsymbol{M}}(t)\boldsymbol{y}_1 + \widehat{\boldsymbol{b}}(t)\right\|_2^2\right] \\ &\leq 2L_M^2\mathbb{E}\left[\left\|\boldsymbol{y}_t - \boldsymbol{y}_1\right\|_2^2\right] + 2\mathbb{E}\left[\left\|\widehat{\boldsymbol{M}}(t)\boldsymbol{y}_1 + \widehat{\boldsymbol{b}}(t)\right\|_2^2\right],\end{aligned}$$

and a similar bound can be shown for the norm of $\widetilde{\boldsymbol{g}}_y(t)$ as well. Putting the above bounds together and setting $\varrho_y \geq 2L_M\eta_x$ and $\varrho_x \geq 2L_M\eta_y$ gives the result. □

As can be seen from the proof, the role of the additional regularization term for the x-player is to eliminate the gradient norms appearing in the regret bound of the y-player.

This effect kicks in once the regularization parameter ϱ_x becomes large enough, so that the corresponding negative term in the regret bound of the first player can overpower the positive term appearing on the bound of the opposite player. The same story applies to the second player. Note that while the regularization pulls the iterates closer to the initial point $\boldsymbol{x}_1, \boldsymbol{y}_1$, it does not explicitly guarantee that they remain uniformly bounded at all times t, and in fact such claim seems impossible to show in general due to the noise in the gradient estimates. Remarkably, the analysis above works seamlessly for noisy gradient estimates, even though the gradient noise can grow proportionally with the size of the iterates.

3.2. Sub-bilinear games and general divergences

After setting the stage in the previous section, we are now ready to introduce our method in its full generality. Specifically, we are going to consider a somewhat broader class of objective functions, and provide mirror-descent style performance guarantees that measure distances in terms of Bregman divergences. We are going to take inspiration from Theorem 3.1 and its proof we have just presented: in short, the idea is to add appropriate regularization terms to the objective that will cancel some otherwise large positive terms in the regret analyses of the two players. The choice of the regularization terms will be somewhat more involved in this case, and will require taking the structure of the objective function into account.

We will let $\omega_x : \mathcal{X} \to \mathbb{R}$ and $\omega_y : \mathcal{Y} \to \mathbb{R}$ be two convex functions, to be called the *distance-generating functions* over $\mathcal{X}$ and $\mathcal{Y}$. We suppose that ω_x is γ_x-strongly convex with respect to the norm $\|\cdot\|_x$ and similarly that ω_y is γ_y-strongly convex with respect to $\|\cdot\|_y$. We will respectively denote the Bregman divergences induced by ω_x and ω_y as $\mathcal{D}_x(\cdot\|\cdot)$ and $\mathcal{D}_y(\cdot\|\cdot)$. We will assume that the objective function satisfies the following condition:

Definition 3.2. (**sub-bilinear function**) A convex-concave function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is said to be *l-sub-bilinear* for some $l > 0$ with respect to the norms $\|\cdot\|_x$ and $\|\cdot\|_y$, if its subgradients $\boldsymbol{g}_x \in \partial_x f(\boldsymbol{x}, \boldsymbol{y})$ and $\boldsymbol{g}_y \in \partial_y f(\boldsymbol{x}, \boldsymbol{y})$ satisfy the following conditions for all $\boldsymbol{x}, \boldsymbol{y}$:

$$\|\boldsymbol{g}_x\|_{x,*}^2 \le l^2 \left(\|\boldsymbol{y}\|_{x,*}^2 + 1\right),$$
$$\|\boldsymbol{g}_y\|_{y,*}^2 \le l^2 \left(\|\boldsymbol{x}\|_{y,*}^2 + 1\right).$$

This condition effectively states that, for a fixed $\boldsymbol{y}$ (resp. $\boldsymbol{x}$), the objective function $f(\boldsymbol{x}, \boldsymbol{y})$ is Lipschitz with respect to $\boldsymbol{x}$ (resp. $\boldsymbol{y}$) with a constant that grows at most as fast as $\|\boldsymbol{y}\|_{x,*}$ (resp. $\|\boldsymbol{x}\|_{y,*}$). Put differently, it means that f behaves like a bilinear function asymptotically as one approaches infinity in each direction, which justifies the name "sub-bilinear" (mirroring the notion of "sublinearity" or "subadditivity" in convex analysis, cf. Hiriart-Urruty & Lemaréchal, 2001, Section C.1). We will further suppose that the stochastic gradients themselves satisfy the following conditions for some $L > 0$:

$$\begin{aligned}\mathbb{E}_t\left[\|\widetilde{\boldsymbol{g}}_x(t)\|_{x,*}\right]^2 &\le L^2\left(\|\boldsymbol{y}_t - \boldsymbol{y}_1\|_{x,*}^2 + 1\right), \\ \mathbb{E}_t\left[\|\widetilde{\boldsymbol{g}}_y(t)\|_{y,*}\right]^2 &\le L^2\left(\|\boldsymbol{x}_t - \boldsymbol{x}_1\|_{y,*}^2 + 1\right).\end{aligned} \tag{4}$$

Supposing that the condition holds with norms respectively centered at $\boldsymbol{x}_1$ and $\boldsymbol{y}_1$ is without loss of generality, and in particular one can always verify $l^2\left(\|\boldsymbol{x}\|_{y,*}^2 + 1\right) \le L^2\left(\|\boldsymbol{x} - \boldsymbol{x}_1\|_{y,*}^2 + 1\right)$ at the price of replacing l by a larger factor L that may depend on $\|\boldsymbol{x}_1\|_{y,*}$.

For this setting, our algorithm is an adaptation of *composite-objective mirror descent* (**COMID**, Duchi et al., 2010), which itself is an adaptation of the classic mirror descent method of Nemirovski & Yudin (1983); Beck & Teboulle (2003), variants of which have been used broadly since the early days of numerical optimization (Rockafellar, 1976; Martinet, 1970; 1978). In particular, we introduce the additional regularization functions $\mathcal{H}_x : \mathcal{X} \to \mathbb{R}$ and $\mathcal{H}_y : \mathcal{Y} \to \mathbb{R}$ defined respectively for each $\boldsymbol{x}$ and $\boldsymbol{y}$ as $\mathcal{H}_x(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_1\|_{y,*}^2$ and $\mathcal{H}_y(\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{y}_1\|_{x,*}^2$, and use these as additional regularization terms to calculate the following sequence of updates in each round $t = 1, 2, \dots, T$:

$$\begin{aligned}\boldsymbol{x}_{t+1} &= \operatorname*{arg\,min}_{\boldsymbol{x} \in \mathcal{X}} \left\{ \langle \boldsymbol{x}, \widetilde{\boldsymbol{g}}_x(t)\rangle + \varrho_x \mathcal{H}_x(\boldsymbol{x}) + \frac{1}{\eta_x}\mathcal{D}_x(\boldsymbol{x}\|\boldsymbol{x}_t)\right\} \\ \boldsymbol{y}_{t+1} &= \operatorname*{arg\,max}_{\boldsymbol{y} \in \mathcal{Y}} \left\{ \langle \boldsymbol{y}, \widetilde{\boldsymbol{g}}_y(t)\rangle - \varrho_y \mathcal{H}_y(\boldsymbol{y}) - \frac{1}{\eta_y}\mathcal{D}_y(\boldsymbol{y}\|\boldsymbol{y}_t)\right\},\end{aligned}$$

We refer to this algorithm as Composite-Objective Mirror Descent-Ascent (**COMIDA**), and provide our main result regarding its performance.

Theorem 3.3. *Suppose that f is sub-bilinear and the stochastic gradients satisfy the conditions in Equation (4). Letting $\varrho_x = \frac{\eta_y L^2}{\gamma_y}$ and $\varrho_y = \frac{\eta_x L^2}{\gamma_x}$, and $\boldsymbol{x}^*, \boldsymbol{y}^*$ be arbitrary comparator points, the expected duality gap of* **COMIDA** *satisfies the following bound:*

$$\begin{aligned}\mathbb{E}\left[G(\boldsymbol{x}^*; \boldsymbol{y}^*)\right] \le\ & \frac{\mathcal{D}_y(\boldsymbol{y}^*\|\boldsymbol{y}_1)}{\eta_y T} + \frac{\varrho_y \|\boldsymbol{y}^* - \boldsymbol{y}_1\|_{x,*}^2}{2} \\ &+ \frac{\mathcal{D}_x(\boldsymbol{x}^*\|\boldsymbol{x}_1)}{\eta_x T} + \frac{\varrho_x \|\boldsymbol{x}^* - \boldsymbol{x}_1\|_{y,*}^2}{2} \\ &+ L^2\left(\frac{\eta_y}{2\gamma_y} + \frac{\eta_x}{2\gamma_x}\right).\end{aligned}$$

The most important special case of our setting is when the norms appearing in the statement are dual to each other, and in particular $\|\cdot\|_x = \|\cdot\|_{y,*}$ and $\|\cdot\|_y = \|\cdot\|_{x,*}$, so that ω_x

is strongly convex with respect to $\|\cdot\|_{y,*}$ and ω_y is strongly convex with respect to $\|\cdot\|_{x,*}$. This is the case for instance when $\mathcal{X} = \mathcal{Y} = \mathbb{R}^m$, $\omega_x = \frac{1}{2}\|\cdot\|^2_{\boldsymbol{A}}$ and $\omega_y = \frac{1}{2}\|\cdot\|^2_{\boldsymbol{A}^{-1}}$ for a symmetric positive definite matrix $\boldsymbol{A} \in \mathbb{R}^{m \times m}$. We state a specialized version of our statement to this setting below.

Corollary 3.4. *Suppose that f is sub-bilinear and the stochastic gradients satisfy the conditions in Equation (4), and suppose additionally that ω_x is γ_x-strongly convex with respect to $\|\cdot\|_{y,*}$ and ω_y is γ_y-strongly convex with respect to $\|\cdot\|_{x,*}$. Set the parameters as $\varrho_x = \frac{\eta_y L^2}{\gamma_y}$, $\varrho_y = \frac{\eta_x L^2}{\gamma_x}$, $\eta_x = \eta_y = \sqrt{\gamma_x \gamma_y / L^2 T}$. Then, letting $\boldsymbol{x}^*, \boldsymbol{y}^*$ be arbitrary comparator points, the duality gap of* **`COMIDA`** *satisfies the following bound:*

$$\mathbb{E}\left[G(\boldsymbol{x}^*; \boldsymbol{y}^*)\right] = \mathcal{O}\left(\frac{L\left(\mathcal{D}_x(\boldsymbol{x}^* \| \boldsymbol{x}_1) + \mathcal{D}_y(\boldsymbol{y}^* \| \boldsymbol{y}_1) + 1\right)}{\sqrt{\gamma_x \gamma_y T}}\right).$$

The proof simply follows from using the definition of strong convexity to upper bound $\gamma_y \|\boldsymbol{y}^* - \boldsymbol{y}_1\|^2_{x,*} \leq 2\mathcal{D}_y(\boldsymbol{y}^* \| \boldsymbol{y}_1)$ and $\gamma_x \|\boldsymbol{x}^* - \boldsymbol{x}_1\|^2_{y,*} \leq 2\mathcal{D}_y(\boldsymbol{x}^* \| \boldsymbol{x}_1)$.

The above results enjoy the same initialization-dependent property as the ones we have established earlier for bilinear games, with the upgrade that the result now holds in terms of general Bregman divergences and also slightly relaxes the conditions on the objective function.

4. Application to Average Reward Markov Decision Processes

In this section, we apply techniques from the previous section for computing near-optimal policies in average-reward Markov Decision Processes (AMDPs). As it is well-known, this task can be formulated as a linear program (LP), which in turn can be solved by finding a saddle point of the associated Lagrangian. Below, we will only describe the saddle-point optimization problem itself and give more context on the problem in Appendix B. For a full technical description of the LP formulation of optimal control in MDPs, we refer to Section 8.8 in the classic textbook of Puterman (1994).

We consider infinite-horizon AMDPs denoted as $(\mathcal{S}, \mathcal{A}, r, P)$ where $\mathcal{S}$ is a finite state space of cardinality S, $\mathcal{A}$ is a finite action space of cardinality A, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ a reward function and $P : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ a stochastic transition model. For ease of notation, we often refer to the reward vector $\boldsymbol{r} \in \mathbb{R}^{SA}$ with entries $\{r(s,a)\}_{(s,a) \in \mathcal{S} \times \mathcal{A}}$, and the transition matrix $\boldsymbol{P} \in \mathbb{R}^{SA \times \mathcal{S}}$ with rows $\boldsymbol{P}_{(s,a),\cdot} = P(\cdot|s,a) \in \Delta_S$ for $(s,a) \in \mathcal{S} \times \mathcal{A}$. We also define the matrix $\boldsymbol{E} \in \mathbb{R}^{SA \times S}$ with entries $\boldsymbol{E}_{(s,a),s'} = \mathbb{I}_{\{s = s'\}}$.

The agent-environment interaction in this MDP setting is described thus: for $k = 1, 2, \cdots, K$ steps, having observed the current state s_k of the environment, the agent takes action a_k according to some stochastic policy $\pi(\cdot|s_k)$. In consequence of this action, the agent receives an immediate reward $r_k = r(s_k, a_k)$, and moves to the next state $s_{k+1} \sim P(\cdot|s_k, a_k)$, from where the interaction continues. The performance of the policy π is measured in terms of the *long-term average reward* (or *gain*) $\rho^\pi = \limsup_{K \to \infty} \frac{1}{K} \mathbb{E}_\pi\left[\sum_{k=1}^K r(s_k, a_k)\right]$. The goal of the optimal control problem is to find an optimal policy π^* that achieves maximal average reward: $\pi^* = \arg\max_\pi \rho(\pi)$. We provide more details on the existence conditions of such optimal policies in the Appendix.

The Lagrangian associated with the optimal control problem is written as

$$\mathcal{L}(\boldsymbol{\mu}; \boldsymbol{v}) = \langle \boldsymbol{\mu}, \boldsymbol{r} \rangle + \langle \boldsymbol{v}, \boldsymbol{P}^\mathsf{T} \boldsymbol{\mu} - \boldsymbol{E}^\mathsf{T} \boldsymbol{\mu} \rangle\,.$$

Here, the primal variable $\boldsymbol{\mu} \in \Delta_{SA}$ is a probability distribution on the state-action space that we will refer to as an *occupancy measure* and the dual $\boldsymbol{v} \in \mathbb{R}^S$ is a real-valued function that we will refer to as a *value function*. The saddle point $(\boldsymbol{\mu}^*, \boldsymbol{v}^*)$ corresponds to the pair of the optimal occupancy measure $\boldsymbol{\mu}^*$ and the optimal value function $\boldsymbol{v}^*$. In most problems of practical interest, the scale of the value functions is unknown a priori, and consequently there is no tractable way of coming up with a bounded set $\mathcal{V} \subset \mathbb{R}^S$ that will include the optimal value function $\boldsymbol{v}^*$. Without such prior knowledge, one has to solve the *unconstrained* saddle-point optimization problem $\min_{\boldsymbol{v} \in \mathbb{R}^S} \max_{\boldsymbol{\mu} \in \Delta_{SA}} \mathcal{L}(\boldsymbol{\mu}; \boldsymbol{v})$ in order to find the optimal policy—which is precisely the subject of our paper.

We will employ a version of our stochastic primal-dual algorithm to solve the above unconstrained problem. We work in the well-studied setting of *planning with random-access models*, where we are given a *simulator* (or *generative model*) of the transition function P that we can query at any state-action pair (s,a) for an i.i.d. sample from $P(\cdot|s,a)$. We will use this simulator to build estimators of the gradients

$$\begin{aligned}\nabla_{\boldsymbol{v}} \mathcal{L}(\boldsymbol{\mu}; \boldsymbol{v}) &= \boldsymbol{P}^\mathsf{T} \boldsymbol{\mu} - \boldsymbol{E}^\mathsf{T} \boldsymbol{\mu} \\ \nabla_{\boldsymbol{\mu}} \mathcal{L}(\boldsymbol{\mu}; \boldsymbol{v}) &= \boldsymbol{r} + \boldsymbol{P}\boldsymbol{v} - \boldsymbol{E}\boldsymbol{v},\end{aligned}$$

with their stochastic estimators calculated for each t as

$$\begin{aligned}\widetilde{\boldsymbol{g}}_v(t) &= \boldsymbol{e}_{s'_t} - \boldsymbol{e}_{s_t} \\ \widetilde{\boldsymbol{g}}_\mu(t) &= \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} [r(s,a) + v_t(\overline{s}'_t) - v_t(s)]\boldsymbol{e}_{(s,a)},\end{aligned}$$

using i.i.d. samples $(s_t, a_t) \sim \boldsymbol{\mu}_t, s'_t \sim P(\cdot|s_t, a_t)$, also $\overline{s}'_t(s,a) \sim P(\cdot|s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. This makes for a total of $SA + 1$ queries per gradient computation.

Since in our setting only $\boldsymbol{v}$ is unconstrained, it will be enough to introduce the stabilizing regularization for these parameters. With that, our algorithm will initialize $\boldsymbol{v}_1 = 0$ and $\boldsymbol{\mu}_1$ arbitrarily, and then perform the following sequence of updates for all $t = 1, 2, \dots, T$:

$$\boldsymbol{v}_{t+1} = \underset{\boldsymbol{v}\in\mathbb{R}^S}{\arg\min} \left\{ \langle \boldsymbol{v}, \widetilde{\boldsymbol{g}}_v(t) \rangle + \frac{1}{2\eta_v} \left\| \boldsymbol{v} - \boldsymbol{v}_t \right\|_2^2 + \varrho_v \left\| \boldsymbol{v} \right\|_\infty^2 \right\},$$

$$\boldsymbol{\mu}_{t+1} = \underset{\boldsymbol{\mu}\in\Delta_{SA}}{\arg\min} \left\{ -\left\langle \boldsymbol{\mu}, \widetilde{\boldsymbol{g}}_\mu(t) \right\rangle + \frac{1}{\eta_\mu} \mathcal{D}_{\mathrm{KL}}\left(\boldsymbol{\mu} \| \boldsymbol{\mu}_t \right) \right\},$$

where $\mathcal{D}_{\mathrm{KL}}\left(\boldsymbol{\mu} \| \boldsymbol{\mu}' \right) = \sum_{s,a} \boldsymbol{\mu}(s,a) \log \frac{\boldsymbol{\mu}(s,a)}{\boldsymbol{\mu}'(s,a)}$ is the relative entropy (or Kullback–Leibler divergence) between $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$. We refer to the resulting algorithm as **`COMIDA-MDP`**.

The output of **`COMIDA-MDP`** is a policy $\overline{\pi}_T : \mathcal{S} \to \Delta_{\mathcal{A}}$, defined by first computing the average of the primal iterates $\overline{\mu}_T = \frac{1}{T} \sum_{t=1}^{T} \mu_t$, and then setting

$$\overline{\pi}_T(a|s) = \frac{\overline{\mu}_T(s,a)}{\sum_{a'\in\mathcal{A}} \overline{\mu}_T(s,a')}$$

for all s, a. Then, adapting a result from Cheng et al. (2020), we can relate the suboptimality of the output policy to the duality gap evaluated at a well-chosen pair of comparator points $(\boldsymbol{\mu}^*, \boldsymbol{v}^{\overline{\pi}_T})$:

$$\rho^{\pi^*} - \rho^{\overline{\pi}_T} = G(\boldsymbol{\mu}^{\pi^*}; \boldsymbol{v}^{\overline{\pi}_T}).$$

Notably, the size of the comparator point $\boldsymbol{v}^{\overline{\pi}_T}$ is unknown a priori, and additionally it depends on the interaction history which will necessitate some extra care in our analysis. We once again refer to Appendix B for more details regarding the choice of $\boldsymbol{v}^{\overline{\pi}_T}$ and the formal proof of the above claim.

Our main result in this section is the following.

Theorem 4.1. *Let $\varrho_v = 4\eta_\mu$. Then, the output of* **`COMIDA-MDP`** *satisfies the following bound:*

$$\mathbb{E}\left[\left\langle \boldsymbol{\mu}^{\pi^*} - \boldsymbol{\mu}^{\overline{\pi}_T}, \boldsymbol{r} \right\rangle \right] \le \frac{\mathcal{D}_{\mathrm{KL}}\left(\boldsymbol{\mu}^{\pi^*} \| \boldsymbol{\mu}_1 \right)}{\eta_\mu T} + \eta_\mu + 4\eta_v + \left(\frac{1}{\eta_v T} + 4\eta_\mu \right) \mathbb{E}\left[\left\| \boldsymbol{v}^{\overline{\pi}_T} \right\|_2^2 \right].$$

In particular, if the output policy satisfies $\left\| \boldsymbol{v}^{\overline{\pi}_T} \right\|_\infty \le B$ for some $B > 0$ and μ_1 is the uniform distribution over SA, and tuning the parameters as $\eta_\mu = \sqrt{\frac{\log(SA)}{ST}}$ and $\eta_v = \sqrt{SA/T}$, the bound becomes

$$\mathbb{E}\left[\left\langle \boldsymbol{\mu}^{\pi^*} - \boldsymbol{\mu}^{\overline{\pi}_T}, \boldsymbol{r} \right\rangle \right] = \mathcal{O}\left(\sqrt{\frac{B^4 SA \log(SA)}{T}} \right).$$

Thus, the iteration complexity of **`COMIDA-MDP`** for finding an ε-optimal policy is of the order $\frac{B^4 SA \log(SA)}{\varepsilon^2}$. As each iteration uses $SA + 1$ queries to the generative model, this makes for a total of $\frac{B^4 S^2 A^2 \log(SA)}{\varepsilon^2}$ query complexity, which is suboptimal in terms of its dependence on SA, but optimal in terms of ε. Most importantly, this guarantee constitutes the first one we are aware of in the literature that does not require prior knowledge of the so-called "bias span" B.

5. Discussion

Our work contributes to the rich literature on saddle-point optimization via incremental first-order methods, a subject studied at least since the works of Martinet (1970; 1978); Rockafellar (1976); Nemirovski & Yudin (1983). In the last few years, this topic has enjoyed a massive comeback within the context of optimization for machine learning models, and in particular generative adversarial networks (GANs, Goodfellow et al., 2014). The instability of standard gradient descent/ascent methods has been pointed out early on during this revival, which brought significant attention to a family of methods known extragradient methods, first proposed by Korpelevich (1976) and further developed by Popov (1980); Nemirovski (2004); Juditsky et al. (2011); Rakhlin & Sridharan (2013a;b). A wealth of recent works have contributed to a better understanding of these methods, and most notably established last-iterate convergence of extragradient-type methods for a variety of problem settings (Daskalakis et al., 2017; Gidel et al., 2018; Mertikopoulos et al., 2018; Mishchenko et al., 2020). The majority of these works assume access to either deterministic gradients or gradients with uniformly bounded noise and bounded domain. The assumption of bounded noise was more recently lifted in the works of Loizou et al. (2021) and Sadiev et al. (2023), but their assumptions on the noise and the objective function are ultimately incompatible with our setting.

We leave several interesting questions open for future work. The biggest of these questions is if the scaling with the initialization error $\|\boldsymbol{x}^* - \boldsymbol{x}_1\|_2^2 + \|\boldsymbol{y}^* - \boldsymbol{y}_1\|_2^2$ can be improved to $\|\boldsymbol{x}^* - \boldsymbol{x}_1\|_2 + \|\boldsymbol{y}^* - \boldsymbol{y}_1\|_2$. This is obviously possible when we have prior knowledge of these norms, and can tune the learning rate to fully optimize the first set of bounds in Theorem 3.1. Without prior knowledge, it is less clear if such improvement is possible, unlike in the case of convex minimization problems where there exist efficient algorithms that achieve such improved rates, at least up to log factors (Streeter & McMahan, 2012; Orabona, 2013; 2014). Another interesting line of investigation is to find out if it is possible to extend our methodology to go significantly beyond bilinear objectives.

We close by recalling that our approach bears some significant similarity with the stabilized online mirror descent method of Fang et al. (2022): their approach introduces a similar regularization term to address issues faced by OMD in unconstrained convex minimization problems. This idea was adapted to equilibrium finding in multiplayer games by Hsieh et al. (2021), but their results are once again not comparable to ours even in two-player zero-sum games (e.g., they consider noiseless gradients and bounded decision sets). We are curious to see if this stabilization trick can find further uses in the context of saddle-point optimization and game theory in the future.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 950180).

References

Abernethy, J., Lai, K. A., Levy, K. Y., and Wang, J.-K. Faster rates for convex-concave games. In *Conference On Learning Theory*, pp. 1595–1625. PMLR, 2018.

Bartlett, P. L. and Tewari, A. REGAL: A regularization based algorithm for reinforcement learning in weakly communicating MDPs. In *Uncertainty in Artificial Intelligence*, 2009.

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

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

Cheng, C.-A., Combes, R. T., Boots, B., and Gordon, G. A reduction from reinforcement learning to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3514–3524. PMLR, 2020.

Cutkosky, A. Artificial constraints and hints for unbounded online learning. In *Conference on Learning Theory*, pp. 874–894. PMLR, 2019.

Cutkosky, A. and Boahen, K. A. Online convex optimization with unconstrained domains and losses. *Advances in neural information processing systems*, 29, 2016.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Duchi, J. C., Shalev-Shwartz, S., Singer, Y., and Tewari, A. Composite objective mirror descent. In *Conference on Learning Theorz*, volume 10, pp. 14–26, 2010.

Fang, H., Harvey, N. J., Portella, V. S., and Friedlander, M. P. Online mirror descent and dual averaging: keeping pace in the dynamic case. *The Journal of Machine Learning Research*, 23(1):5271–5308, 2022.

Fruit, R., Pirotta, M., and Lazaric, A. Near optimal exploration-exploitation in non-communicating markov decision processes. *Advances in Neural Information Processing Systems*, 31, 2018a.

Fruit, R., Pirotta, M., Lazaric, A., and Ortner, R. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In *International Conference on Machine Learning*, pp. 1578–1586, 2018b.

Gidel, G., Berard, H., Vignoud, G., Vincent, P., and Lacoste-Julien, S. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

Goodfellow, I. NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Hiriart-Urruty, J.-B. and Lemaréchal, C. *Fundamentals of Convex Analysis*. Springer, 2001.

Hsieh, Y.-G., Antonakopoulos, K., and Mertikopoulos, P. Adaptive learning in continuous games: Optimal regret bounds and convergence to nash equilibrium. In *Conference on Learning Theory*, pp. 2388–2422. PMLR, 2021.

Jin, Y. and Sidford, A. Efficiently solving mdps with stochastic mirror descent. In *International Conference on Machine Learning*, pp. 4890–4900. PMLR, 2020.

Juditsky, A., Nemirovski, A., and Tauvel, C. Solving variational inequalities with stochastic mirror-prox algorithm. *Stochastic Systems*, 1(1):17–58, 2011.

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

Liu, M. and Orabona, F. On the initialization for convex-concave min-max problems. In *International Conference on Algorithmic Learning Theory*, pp. 743–767. PMLR, 2022.

Loizou, N., Berard, H., Gidel, G., Mitliagkas, I., and Lacoste-Julien, S. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34:19095–19108, 2021.

Martinet, B. Régularisation d'inéquations variationnelles par approximations successives. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation*

Mathématique et Analyse Numérique, 4(R3):154–158, 1970.

Martinet, B. Perturbation des méthodes d'optimisation. applications. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 12(2):153–171, 1978. URL http://eudml.org/doc/193317.

Mertikopoulos, P., Lecouat, B., Zenati, H., Foo, C.-S., Chandrasekhar, V., and Piliouras, G. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. In *International Conference on Learning Representations*, 2018.

Mhammedi, Z. and Koolen, W. M. Lipschitz and comparator-norm adaptivity in online learning. In *Conference on Learning Theory*, pp. 2858–2887. PMLR, 2020.

Mishchenko, K., Kovalev, D., Shulgin, E., Richtárik, P., and Malitsky, Y. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pp. 4573–4582. PMLR, 2020.

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, 2004.

Nemirovski, A. S. and Yudin, D. B. Problem complexity and method efficiency in optimization. 1983.

Orabona, F. Dimension-free exponentiated gradient. *Advances in Neural Information Processing Systems*, 26, 2013.

Orabona, F. Simultaneous model selection and optimization through parameter-free stochastic learning. *Advances in Neural Information Processing Systems*, 27, 2014.

Popov, L. D. A modification of the arrow-hurwicz method for search of saddle points. *Mathematical notes of the Academy of Sciences of the USSR*, 28:845–848, 1980.

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

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

Rakhlin, A. and Sridharan, K. On equivalence of martingale tail bounds and deterministic regret inequalities. In *Conference on Learning Theory*, pp. 1704–1722. PMLR, 2017.

Rakhlin, S. and Sridharan, K. Optimization, learning, and games with predictable sequences. *Advances in Neural Information Processing Systems*, 26, 2013b.

Rockafellar, R. T. Monotone Operators and the Proximal Point Algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

Sadiev, A., Danilova, M., Gorbunov, E., Horváth, S., Gidel, G., Dvurechensky, P., Gasnikov, A., and Richtárik, P. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. *arXiv preprint arXiv:2302.00999*, 2023.

Shalev-Shwartz, S. and Singer, Y. Convex repeated games and fenchel duality. *Advances in neural information processing systems*, 19, 2006.

Streeter, M. and McMahan, H. B. No-regret algorithms for unconstrained online convex optimization. *arXiv preprint arXiv:1211.2260*, 2012.

van der Hoeven, D. User-specified local differential privacy in unconstrained adaptive online learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Wang, J.-K. and Abernethy, J. D. Acceleration through optimistic no-regret dynamics. *Advances in Neural Information Processing Systems*, 31, 2018.

Wang, J.-K., Abernethy, J., and Levy, K. Y. No-regret dynamics in the fenchel game: A unified framework for algorithmic convex optimization. *Mathematical Programming*, pp. 1–66, 2023.

Wang, M. Primal-dual π learning: Sample complexity and sublinear run time for ergodic markov decision problems. *CoRR*, abs/1710.06100, 2017. URL http://arxiv.org/abs/1710.06100.

Zhang, Z. and Ji, X. Regret minimization for reinforcement learning by evaluating the optimal bias function. *Advances in Neural Information Processing Systems*, 32, 2019.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, 2003.

A. Proof of results in Section 3

In this section, we provide a detailed proof of claims, lemmas and theorems in Section 3 in the main text.

A.1. Complete proof of Theorem 3.1

We start by rewriting the expected duality gap evaluated at $(\boldsymbol{x}^*; \boldsymbol{y}^*)$ as follows:

$$\begin{aligned}
\mathbb{E}\left[G(\boldsymbol{x}^*; \boldsymbol{y}^*)\right] &= \mathbb{E}\left[f(\overline{\boldsymbol{x}}_T, \boldsymbol{y}^*) - f(\boldsymbol{x}^*, \overline{\boldsymbol{y}}_T)\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f(\boldsymbol{x}_t, \boldsymbol{y}^*) - f(\boldsymbol{x}^*, \boldsymbol{y}_t)\right] \\
&= \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[f^{(\mathrm{reg})}\left(\boldsymbol{x}_t, \boldsymbol{y}^*\right) - f^{(\mathrm{reg})}\left(\boldsymbol{x}^*, \boldsymbol{y}_t\right)\right] \\
&\quad + \frac{\varrho_y}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{y}^* - \boldsymbol{y}_1\right\|_2^2 - \left\|\boldsymbol{y}_t - \boldsymbol{y}_1\right\|_2^2\right] \\
&\quad + \frac{\varrho_x}{2T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{x}^* - \boldsymbol{x}_1\right\|_2^2 - \left\|\boldsymbol{x}_t - \boldsymbol{x}_1\right\|_2^2\right]. \qquad (5)
\end{aligned}$$

To control the first set of terms in the above expression, we apply the regret analysis of **COMIDA** in Appendix A.2.1. Precisely, with $\mathcal{D}_x(\boldsymbol{x}\|\boldsymbol{x}') = \frac{1}{2}\left\|\boldsymbol{x} - \boldsymbol{x}'\right\|_2^2$, $\mathcal{D}_y(\boldsymbol{y}\|\boldsymbol{y}') = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{y}'\right\|_2^2$ and, $\mathcal{H}_x(\boldsymbol{x}) = \frac{1}{2}\left\|\boldsymbol{x} - \boldsymbol{x}_1\right\|_2^2$, $\mathcal{H}_y(\boldsymbol{y}) = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{y}_1\right\|_2^2$, we get:

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[f^{(\mathrm{reg})}\left(\boldsymbol{x}_t, \boldsymbol{y}^*\right) - f^{(\mathrm{reg})}\left(\boldsymbol{x}^*, \boldsymbol{y}_t\right)\right] \\
&\qquad = \sum_{t=1}^{T}\mathbb{E}\left[f^{(\mathrm{reg})}\left(\boldsymbol{x}_t, \boldsymbol{y}^*\right) - f^{(\mathrm{reg})}\left(\boldsymbol{x}_t, \boldsymbol{y}_t\right)\right] + \sum_{t=1}^{T}\mathbb{E}\left[f^{(\mathrm{reg})}\left(\boldsymbol{x}_t, \boldsymbol{y}_t\right) - f^{(\mathrm{reg})}\left(\boldsymbol{x}^*, \boldsymbol{y}_t\right)\right] \\
&\qquad \leq \frac{\left\|\boldsymbol{y}^* - \boldsymbol{y}_1\right\|_2^2}{2\eta_y} + \frac{\eta_y}{2}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_y(t)\right\|_2^2\right] + \frac{\left\|\boldsymbol{x}^* - \boldsymbol{x}_1\right\|_2^2}{2\eta_x T} + \frac{\eta_x}{2}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_x(t)\right\|_2^2\right]. \qquad (6)
\end{aligned}$$

To proceed, we recall the assumptions we made on the gradient estimators on the main text, namely that the inequalities $\mathbb{E}_t\left[\left\|\widehat{\boldsymbol{M}}(t)\boldsymbol{y}\right\|_2^2\right] \leq L_M^2\left\|\boldsymbol{y}\right\|_2^2$ and $\mathbb{E}_t\left[\left\|\widehat{\boldsymbol{M}}(t)^\intercal\boldsymbol{x}\right\|_2^2\right] \leq L_M^2\left\|\boldsymbol{x}\right\|_2^2$ hold for all $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X} \times \mathcal{Y}$. Using this condition allows us to bound the gradient norms as

$$\begin{aligned}
\mathbb{E}_t\left[\left\|\widetilde{\boldsymbol{g}}_y(t)\right\|_2^2\right] &= \mathbb{E}_t\left[\left\|\widehat{\boldsymbol{M}}(t)^\intercal\boldsymbol{x}_t - \widehat{\boldsymbol{c}}(t)\right\|_2^2\right] \\
&= \mathbb{E}_t\left[\left\|\widehat{\boldsymbol{M}}(t)^\intercal\left(\boldsymbol{x}_t - \boldsymbol{x}_1\right) + \widehat{\boldsymbol{M}}(t)^\intercal\boldsymbol{x}_1 - \widehat{\boldsymbol{c}}(t)\right\|_2^2\right] \\
&\leq 2\mathbb{E}_t\left[\left\|\widehat{\boldsymbol{M}}(t)^\intercal\left(\boldsymbol{x}_t - \boldsymbol{x}_1\right)\right\|_2^2\right] + 2\mathbb{E}_t\left[\left\|\widehat{\boldsymbol{M}}(t)^\intercal\boldsymbol{x}_1 - \widehat{\boldsymbol{c}}(t)\right\|_2^2\right] \\
&\leq 2L_M^2\left\|\boldsymbol{x}_t - \boldsymbol{x}_1\right\|_2^2 + 2\mathbb{E}_t\left[\left\|\widehat{\boldsymbol{M}}(t)^\intercal\boldsymbol{x}_1 - \widehat{\boldsymbol{c}}(t)\right\|_2^2\right],
\end{aligned}$$

where the third line uses the triangle inequality and Cauchy–Schwarz, and the second follows from said assumption. Likewise, we can show

$$\mathbb{E}_t\left[\left\|\widetilde{\boldsymbol{g}}_x(t)\right\|_2^2\right] \leq 2L_M^2\left\|\boldsymbol{y}_t - \boldsymbol{y}_1\right\|_2^2 + 2\mathbb{E}_t\left[\left\|\widehat{\boldsymbol{M}}(t)\boldsymbol{y}_1 + \widehat{\boldsymbol{b}}(t)\right\|_2^2\right].$$

Therefore, by the tower rule and monotonicity of expectation,

$$\mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_y(t)\right\|_2^2\right] = \mathbb{E}\left[\mathbb{E}_t\left[\left\|\widetilde{\boldsymbol{g}}_y(t)\right\|_2^2\right]\right] \leq 2L_M^2\mathbb{E}\left[\left\|\boldsymbol{x}_t - \boldsymbol{x}_1\right\|_2^2\right] + 2\mathbb{E}\left[\left\|\widehat{\boldsymbol{M}}(t)^\intercal\boldsymbol{x}_1 - \widehat{\boldsymbol{c}}(t)\right\|_2^2\right],$$

and

$$\mathbb{E}\left[\|\widetilde{\boldsymbol{g}}_x(t)\|_2^2\right] = \mathbb{E}\left[\mathbb{E}_t\left[\|\widetilde{\boldsymbol{g}}_x(t)\|_2^2\right]\right] \leq 2L_M^2\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_1\|_2^2\right] + 2\mathbb{E}\left[\left\|\widehat{\boldsymbol{M}}(t)\boldsymbol{y}_1 + \widehat{\boldsymbol{b}}(t)\right\|_2^2\right].$$

Plugging these derivations into the bound of Equation Equation (6) and then combining the result with the bound of Equation Equation (5), we obtain

$$\begin{aligned}\mathbb{E}\left[G(\boldsymbol{x}^*;\boldsymbol{y}^*)\right] \leq &\left(\frac{1}{2\eta_y T} + \frac{\varrho_y}{2}\right)\|\boldsymbol{y}^* - \boldsymbol{y}_1\|_2^2 + \frac{\eta_y}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\widehat{\boldsymbol{M}}(t)^\intercal\boldsymbol{x}_1 - \widehat{\boldsymbol{c}}(t)\right\|_2^2\right] \\ &+ \left(\frac{1}{2\eta_x T} + \frac{\varrho_x}{2}\right)\|\boldsymbol{x}^* - \boldsymbol{x}_1\|_2^2 + \frac{\eta_x}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\widehat{\boldsymbol{M}}(t)\boldsymbol{y}_1 + \widehat{\boldsymbol{b}}(t)\right\|_2^2\right] \\ &+ \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_1\|_2^2\right]\left(\eta_x L_M^2 - \frac{\varrho_y}{2}\right) + \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\|\boldsymbol{x}_t - \boldsymbol{x}_1\|_2^2\right]\left(\eta_y L_M^2 - \frac{\varrho_x}{2}\right).\end{aligned}$$

By setting $\varrho_y = 2\eta_x L_M^2$ and $\varrho_x = 2\eta_y L_M^2$, we eliminate the last two terms in the bound above and arrive at the result stated in the theorem:

$$\begin{aligned}\mathbb{E}\left[G(\boldsymbol{x}^*;\boldsymbol{y}^*)\right] \leq &\left(\frac{1}{2\eta_y T} + \eta_x L_M^2\right)\|\boldsymbol{y}^* - \boldsymbol{y}_1\|_2^2 + \frac{\eta_y}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\widehat{\boldsymbol{M}}(t)^\intercal\boldsymbol{x}_1 - \widehat{\boldsymbol{c}}(t)\right\|_2^2\right] \\ &+ \left(\frac{1}{2\eta_x T} + \eta_y L_M^2\right)\|\boldsymbol{x}^* - \boldsymbol{x}_1\|_2^2 + \frac{\eta_x}{T}\sum_{t=1}^T \mathbb{E}\left[\left\|\widehat{\boldsymbol{M}}(t)\boldsymbol{y}_1 + \widehat{\boldsymbol{b}}(t)\right\|_2^2\right].\end{aligned}$$

□

A.2. Proof of Theorem 3.3

Consider the expected duality gap at arbitrary comparator points $(\boldsymbol{x}^*, \boldsymbol{y}^*)$:

$$\mathbb{E}\left[G(\boldsymbol{x}^*;\boldsymbol{y}^*)\right] = \mathbb{E}\left[f(\overline{\boldsymbol{x}}_T, \boldsymbol{y}^*) - f(\boldsymbol{x}^*, \overline{\boldsymbol{y}}_T)\right].$$

By the convex-concave property of f and straightforward derivations, we can rewrite the above gap in terms of the regret of a min-max optimization scheme and regularization terms as

$$\begin{aligned}\mathbb{E}\left[G(\boldsymbol{x}^*;\boldsymbol{y}^*)\right] &= \mathbb{E}\left[f(\overline{\boldsymbol{x}}_T, \boldsymbol{y}^*) - f(\boldsymbol{x}^*, \overline{\boldsymbol{y}}_T)\right] \\ &\leq \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[f\left(\boldsymbol{x}_t, \boldsymbol{y}^*\right) - f\left(\boldsymbol{x}^*, \boldsymbol{y}_t\right)\right] \\ &= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[f\left(\boldsymbol{x}_t, \boldsymbol{y}^*\right) - f\left(\boldsymbol{x}_t, \boldsymbol{y}_t\right)\right] + \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[f\left(\boldsymbol{x}_t, \boldsymbol{y}_t\right) - f\left(\boldsymbol{x}^*, \boldsymbol{y}_t\right)\right] \\ &= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[f^{(\mathrm{reg})}\left(\boldsymbol{x}_t, \boldsymbol{y}^*\right) - f^{(\mathrm{reg})}\left(\boldsymbol{x}_t, \boldsymbol{y}_t\right)\right] + \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[f^{(\mathrm{reg})}\left(\boldsymbol{x}_t, \boldsymbol{y}_t\right) - f^{(\mathrm{reg})}\left(\boldsymbol{x}^*, \boldsymbol{y}_t\right)\right] \\ &\quad + \frac{\varrho_y}{T}\sum_{t=1}^T \mathbb{E}\left[\mathcal{H}_y(\boldsymbol{y}^*) - \mathcal{H}_y(\boldsymbol{y}_t)\right] + \frac{\varrho_x}{T}\sum_{t=1}^T \mathbb{E}\left[\mathcal{H}_x(\boldsymbol{x}^*) - \mathcal{H}_x(\boldsymbol{x}_t)\right],\end{aligned} \tag{7}$$

where in this case,

$$f^{(\mathrm{reg})}(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}, \boldsymbol{y}) + \frac{\varrho_x}{2}\mathcal{H}_x(\boldsymbol{x}) - \frac{\varrho_y}{2}\mathcal{H}_y(\boldsymbol{y}).$$

The rest of the proof is split in two parts. First, we control regularized regret of the min and max players, corresponding to the first two sums appearing on the right-hand side of the above bound. Then, substituting the resulting bound back into Equation (7), we take advantage of the negative terms $\mathcal{H}_x(\boldsymbol{x}_t)$ and $\mathcal{H}_y(\boldsymbol{y}_t)$ appearing on the right hand side to cancel out some potentially large terms in the regret analysis, arriving at a bound that is robust to large iterates.

A.2.1. Regret Analysis of COMIDA on a Regularized Objective

This part of the proof is based on the regret analysis of Composite Mirror Descent (**COMID**) for stochastic convex optimization. The proof is a more-or-less standard exercise in convex analysis (appearing, e.g., as Theorem 8 of Duchi et al. (2010)), and we provide it for completeness as Lemma D.1 in Appendix D. In this section, we directly apply the implied guarantee on the regret of **COMID** against a fixed comparator in Corollary D.2 to control the regret of each player.

For the max player, we denote the loss in round t as $\ell_t(\boldsymbol{y}) = -f(\boldsymbol{x}_t, \boldsymbol{y})$ for $\boldsymbol{y} \in \mathcal{Y}$ and we define its regularized loss as $\ell_t^{(\text{reg})}(\boldsymbol{y}) = -f(\boldsymbol{x}_t, \boldsymbol{y}) + \varrho_y \mathcal{H}_y(\boldsymbol{y})$. Then, the total expected regret of the max player on the regularized objective can be rewritten as

$$\sum_{t=1}^{T} \mathbb{E}\left[f^{(\text{reg})}(\boldsymbol{x}_t, \boldsymbol{y}^*) - f^{(\text{reg})}(\boldsymbol{x}_t, \boldsymbol{y}_t)\right] = \sum_{t=1}^{T} \mathbb{E}\left[\ell_t^{(\text{reg})}(\boldsymbol{y}_t) - \ell_t^{(\text{reg})}(\boldsymbol{y}^*)\right].$$

Notice that $\ell_t^{(\text{reg})}(\cdot)$ is convex by the concave property of $f(\boldsymbol{x}_t, \cdot)$. We will bound the regret using Corollary D.2, with initial iterate $\boldsymbol{u}_1 = \boldsymbol{y}_1$, gradient estimates $\widetilde{\boldsymbol{g}}_u(t) = -\widetilde{\boldsymbol{g}}_y(t)$ and gradients $\boldsymbol{g}_{\boldsymbol{u}}(t) = -\boldsymbol{g}_y(t)$. Also, we will set $\mathcal{U} = \mathbb{R}^n$, $\omega_u = \omega_y$, $\eta_u = \eta_y$ and $\varrho_u = \varrho_y$. With $\boldsymbol{y}^*$ fixed and independent of the iterates, this gives

$$\sum_{t=1}^{T} \mathbb{E}\left[f^{(\text{reg})}(\boldsymbol{x}_t, \boldsymbol{y}^*) - f^{(\text{reg})}(\boldsymbol{x}_t, \boldsymbol{y}_t)\right] \le \frac{\mathcal{D}_y(\boldsymbol{y}^* \| \boldsymbol{y}_1)}{\eta_y} + \frac{\eta_y}{2\gamma_y} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_y(t)\right\|_{y,*}^2\right] + \varrho_y \mathbb{E}\left[\mathcal{H}_y(\boldsymbol{y}_1)\right].$$

Likewise, reusing previous notation we denote the loss of the min player as in round t as $\ell_t(\boldsymbol{x}) = f(\boldsymbol{x}, \boldsymbol{y}_t)$. Since $f(\cdot, \boldsymbol{y}_t)$ is convex, and by equivalence of the minimization step of **COMIDA** to that of Corollary D.2 when $\boldsymbol{u}_1 = \boldsymbol{x}_1$, $\widetilde{\boldsymbol{g}}_u(t) = \widetilde{\boldsymbol{g}}_x(t)$, $\boldsymbol{g}_{\boldsymbol{u}}(t) = \boldsymbol{g}_x(t)$, $\mathcal{U} = \mathbb{R}^m$, $\omega_u = \omega_x$, $\eta_u = \eta_x$ and $\varrho_u = \varrho_x$, we can bound the regret of the min player against a fixed comparator $\boldsymbol{x}^*$ as follows:

$$\sum_{t=1}^{T} \mathbb{E}\left[f^{(\text{reg})}(\boldsymbol{x}_t, \boldsymbol{y}_t) - f^{(\text{reg})}(\boldsymbol{x}^*, \boldsymbol{y}_t)\right] \le \frac{\mathcal{D}_x(\boldsymbol{x}^* \| \boldsymbol{x}_1)}{\eta_x} + \frac{\eta_x}{2\gamma_x} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_x(t)\right\|_{x,*}^2\right] + \varrho_x \mathbb{E}\left[\mathcal{H}_x(\boldsymbol{x}_1)\right].$$

Therefore, the total expected regret of **COMIDA** on the regularized objective is bounded above as follows:

$$\begin{aligned} &\sum_{t=1}^{T} \mathbb{E}\left[f^{(\text{reg})}(\boldsymbol{x}_t, \boldsymbol{y}^*) - f^{(\text{reg})}(\boldsymbol{x}_t, \boldsymbol{y}_t)\right] + \sum_{t=1}^{T} \mathbb{E}\left[f^{(\text{reg})}(\boldsymbol{x}_t, \boldsymbol{y}_t) - f^{(\text{reg})}(\boldsymbol{x}^*, \boldsymbol{y}_t)\right] \\ &\quad \le \frac{\mathcal{D}_y(\boldsymbol{y}^* \| \boldsymbol{y}_1)}{\eta_y} + \frac{\eta_y}{2\gamma_y} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_y(t)\right\|_{y,*}^2\right] + \varrho_y \mathbb{E}\left[\mathcal{H}_y(\boldsymbol{y}_1)\right] \\ &\qquad + \frac{\mathcal{D}_x(\boldsymbol{x}^* \| \boldsymbol{x}_1)}{\eta_x} + \frac{\eta_x}{2\gamma_x} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_x(t)\right\|_{x,*}^2\right] + \varrho_x \mathbb{E}\left[\mathcal{H}_x(\boldsymbol{x}_1)\right]. \end{aligned}$$

This completes the first part of the proof.

A.2.2. Eliminating the gradient norms

For the second part, we make use of our specific definition of the regularization function: $\mathcal{H}_x(\boldsymbol{x}) = \frac{1}{2}\left\|\boldsymbol{x} - \boldsymbol{x}_1\right\|_{y,*}^2$ and $\mathcal{H}_y(\boldsymbol{y}) = \frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{y}_1\right\|_{x,*}^2$. In this case $\mathcal{H}_x(\boldsymbol{x}_1) = \mathcal{H}_y(\boldsymbol{y}_1) = 0$. Then, plugging in the bounds from Appendix A.2.1 in the expected duality gap we have:

$$\begin{aligned} \mathbb{E}\left[G(\boldsymbol{x}^*; \boldsymbol{y}^*)\right] \le{}& \frac{\mathcal{D}_y(\boldsymbol{y}^* \| \boldsymbol{y}_1)}{\eta_y T} + \frac{\eta_y}{2\gamma_y T} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_y(t)\right\|_{y,*}^2\right] + \frac{\mathcal{D}_y(\boldsymbol{x}^* \| \boldsymbol{x}_1)}{\eta_x T} + \frac{\eta_x}{2\gamma_x T} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_x(t)\right\|_{x,*}^2\right] \\ &+ \frac{\varrho_y}{2T} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\boldsymbol{y}^* - \boldsymbol{y}_1\right\|_{x,*}^2 - \left\|\boldsymbol{y}_t - \boldsymbol{y}_1\right\|_{x,*}^2\right] + \frac{\varrho_x}{2T} \sum_{t=1}^{T} \mathbb{E}\left[\left\|\boldsymbol{x}^* - \boldsymbol{x}_1\right\|_{y,*}^2 - \left\|\boldsymbol{x}_t - \boldsymbol{x}_1\right\|_{y,*}^2\right]. \end{aligned} \tag{8}$$

To proceed, we make crucial use of our noise condition stated as Equation (4) in the main text so that we can bound the gradient norms as

$$\mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_y(t)\right\|_2^2\right] = \mathbb{E}\left[\mathbb{E}_t\left[\left\|\widetilde{\boldsymbol{g}}_y(t)\right\|_2^2\right]\right] \le \mathbb{E}\left[L^2\left(\left\|\boldsymbol{x}_t - \boldsymbol{x}_1\right\|_{y,*}^2 + 1\right)\right].$$

Also,

$$\mathbb{E}\left[\|\widetilde{\boldsymbol{g}}_x(t)\|_2^2\right] = \mathbb{E}\left[\mathbb{E}_t\left[\|\widetilde{\boldsymbol{g}}_x(t)\|_2^2\right]\right] \le \mathbb{E}\left[L^2\left(\|\boldsymbol{y}_t - \boldsymbol{y}_1\|_{x,*}^2 + 1\right)\right].$$

Plugging these into the bound of Equation (8) gives

$$\begin{aligned}\mathbb{E}\left[G(\boldsymbol{x}^*;\boldsymbol{y}^*)\right] \le\ & \frac{\mathcal{D}_y(\boldsymbol{y}^*\|\boldsymbol{y}_1)}{\eta_y T} + \frac{\eta_y}{2\gamma_y}L^2 + \frac{\varrho_y}{2}\mathbb{E}\left[\|\boldsymbol{y}^* - \boldsymbol{y}_1\|_{x,*}^2\right] \\ & + \frac{\mathcal{D}_y(\boldsymbol{x}^*\|\boldsymbol{x}_1)}{\eta_x T} + \frac{\eta_x}{2\gamma_x}L^2 + \frac{\varrho_x}{2}\mathbb{E}\left[\|\boldsymbol{x}^* - \boldsymbol{x}_1\|_{y,*}^2\right] \\ & + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{y}_t - \boldsymbol{y}_1\|_{x,*}^2\right]\left(\frac{\eta_x L^2}{2\gamma_x} - \frac{\varrho_y}{2}\right) + \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\boldsymbol{x}_t - \boldsymbol{x}_1\|_{y,*}^2\right]\left(\frac{\eta_y L^2}{2\gamma_y} - \frac{\varrho_x}{2}\right).\end{aligned}$$

Lastly, choosing $\varrho_y = \frac{\eta_x L^2}{\gamma_x}$ and $\varrho_x = \frac{\eta_y L^2}{\gamma_y}$ results in the bound stated in the theorem:

$$\begin{aligned}\mathbb{E}\left[G(\boldsymbol{x}^*;\boldsymbol{y}^*)\right] \le\ & \frac{\mathcal{D}_y(\boldsymbol{y}^*\|\boldsymbol{y}_1)}{\eta_y T} + \frac{\eta_y L^2}{2\gamma_y} + \frac{\varrho_y\|\boldsymbol{y}^* - \boldsymbol{y}_1\|_{x,*}^2}{2} \\ & + \frac{\mathcal{D}_x(\boldsymbol{x}^*\|\boldsymbol{x}_1)}{\eta_x T} + \frac{\eta_x L^2}{2\gamma_x} + \frac{\varrho_x\left\|\boldsymbol{x}^* - \boldsymbol{x}_1\right\|_{y,*}^2}{2}.\end{aligned}$$

□

B. Analysis for the Average-Reward MDP Setting

B.1. Problem setup

First we briefly recall some general concepts related to average-reward MDPs (and refer the reader to Chapter 8 of Puterman, 1994 for a more detailed introduction into the topic). Consider infinite-horizon AMDPs denoted as $(\mathcal{S}, \mathcal{A}, r, P)$ where $\mathcal{S}$ is a finite state space of cardinality S, $\mathcal{A}$ is a finite action space of cardinality A, $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$ a reward model and $P : \mathcal{S} \times \mathcal{A} \to \Delta_S$ a stochastic transition model. For ease of notation, we often refer to the reward vector $\boldsymbol{r} \in \mathbb{R}^{SA}$ with $\{r(s, a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ entries, and the transition matrix $\boldsymbol{P} \in \mathbb{R}^{SA\times\mathcal{S}}$ with $\boldsymbol{P}[s, a] = p(\cdot|s, a) \in \Delta_S$ for $(s, a) \in \mathcal{S} \times \mathcal{A}$.

In this work, we primarily focus on the class of AMDPs where each policy π has a well-defined unique stationary state distribution (or state-occupancy measure) $\nu^\pi : \mathcal{S} \to [0, 1]$, defined for each s as

$$\nu^\pi(s) = \lim_{K\to\infty} \frac{1}{K} \sum_{k=1}^{K} \mathbb{P}\left[x_k = x \,|\, \pi\right].$$

The stationary distribution can be seen to satisfy the linear system of equations $\nu^\pi(s) = \sum_{(s',a')} p(s|s', a')\pi(a'|s')\nu^\pi(s')$ for all $s \in \mathcal{S}$. Hence, the corresponding stationary state-action distribution (or *state-action occupancy measure*) $\mu^\pi(s, a) = \pi(a|s)\nu^\pi(s)$ for $(s, a) \in \mathcal{S} \times \mathcal{A}$ is also unique, and we can write the average-reward objective as $\rho^\pi = \langle \boldsymbol{\mu}^\pi, \boldsymbol{r}\rangle$. This compact representation of the reward criterion and occupancy measure inspires the linear programming approach to optimal control in MDPs, wherein we are interested in solving the linear program

$$\begin{aligned} \max_{\boldsymbol{\mu}\in\mathbb{R}^{SA}} \quad & \langle \boldsymbol{\mu}, \boldsymbol{r}\rangle \\ \text{subject to} \quad & \boldsymbol{E}^\mathsf{T}\boldsymbol{\mu} = \boldsymbol{P}^\mathsf{T}\boldsymbol{\mu} \\ & \langle \boldsymbol{\mu}, \mathbf{1}\rangle = 1 \\ & \boldsymbol{\mu} \geq 0. \end{aligned} \tag{9}$$

In the above expressions, the operator $\boldsymbol{E} : \mathbb{R}^{SA} \to \mathbb{R}^{\mathcal{S}}$ is defined as $(\boldsymbol{E}^\mathsf{T}\boldsymbol{\mu})(s) = \sum_a \mu(s, a)$ for $s \in \mathcal{S}$. This LP is motivated by the fact that the set of distributions $\boldsymbol{\mu}$ that satisfy the constraints exactly corresponds to the set of stationary state-action distributions that can be potentially induced by a stationary policy in the MDP.

We also define the value function (or bias function) of policy π as $v^\pi : \mathcal{S} \to \mathbb{R}$, taking the following value in each state $s \in \mathcal{S}$:

$$\begin{aligned} v^\pi(s) &= \lim_{K\to\infty} \mathbb{E}_\pi\left[\sum_{k=1}^{K} \left(r(s_k, a_k) - \rho^\pi\right) \middle| s_0 = s\right] \\ &= \sum_a \pi(a|s)\left[r(s, a) - \rho^\pi + \langle p(\cdot|s, a), \boldsymbol{v}^\pi\rangle\right]. \end{aligned} \tag{10}$$

Then, the value function of an optimal policy maximizing ρ^π can be shown to be an optimal solution of the dual of the LP (9), written as follows:

$$\begin{aligned} \min_{\rho\in\mathbb{R}, \boldsymbol{v}\in\mathcal{V}} \quad & \rho \\ \text{subject to} \quad & \boldsymbol{E}\boldsymbol{v} \geq \boldsymbol{r} + \boldsymbol{P}\boldsymbol{v} - \mathbf{1}\rho. \end{aligned} \tag{11}$$

Finding an optimal solution to either of the LPs can be equivalently phrased as solving the following bilinear game:

$$\min_{\boldsymbol{v}\in\mathcal{V}} \max_{\boldsymbol{\mu}\in\Delta_{SA}} \mathcal{L}(\boldsymbol{v}; \boldsymbol{\mu}), \tag{12}$$

with the Lagrangian associated with the LPs is defined as

$$\begin{aligned} \mathcal{L}(\boldsymbol{v}; \boldsymbol{\mu}) &= \langle \boldsymbol{\mu}, \boldsymbol{r}\rangle + \langle \boldsymbol{v}, \boldsymbol{P}^\mathsf{T}\boldsymbol{\mu} - \boldsymbol{E}^\mathsf{T}\boldsymbol{\mu}\rangle + \rho(1 - \langle \boldsymbol{\mu}, \mathbf{1}\rangle) \\ &= \langle \boldsymbol{\mu}, \boldsymbol{r}\rangle + \langle \boldsymbol{v}, \boldsymbol{P}^\mathsf{T}\boldsymbol{\mu} - \boldsymbol{E}^\mathsf{T}\boldsymbol{\mu}\rangle. \end{aligned}$$

The gradients of the above objective are respectively expressed as

$$\nabla_{\boldsymbol{v}}\mathcal{L}(\boldsymbol{v}; \boldsymbol{\mu}) = \boldsymbol{P}^\mathsf{T}\boldsymbol{\mu} - \boldsymbol{E}^\mathsf{T}\boldsymbol{\mu} \qquad \text{and} \qquad \nabla_{\boldsymbol{\mu}}\mathcal{L}(\boldsymbol{v}; \boldsymbol{\mu}) = \boldsymbol{r} + \boldsymbol{P}\boldsymbol{v} - \boldsymbol{E}\boldsymbol{v}.$$

Now in the context of planning, it is assumed that the transition model is unknown, hence the gradients cannot be computed exactly. Rather, we assume access to an accurate simulator which can be queried at any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ to obtain a sample next state $s' \sim p(\cdot|s, a)$. Indeed, with $\boldsymbol{v}_t, \boldsymbol{\mu}_t$ determined by the end of round $t-1$, we can compute unbiased estimates in round t as:

$$\widetilde{\boldsymbol{g}}_v(t) = \boldsymbol{e}_{s'_t} - \boldsymbol{e}_{s_t}$$
$$\widetilde{\boldsymbol{g}}_\mu(t) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \big(r(s,a) + v_t(\overline{s}'_t(s,a)) - v_t(s)\big)\boldsymbol{e}_{(s,a)},$$

using i.i.d samples $(s_t, a_t) \sim \boldsymbol{\mu}_t, s'_t \sim p(\cdot|s_t, a_t)$, also $\overline{s}'_t(s,a) \sim p(\cdot|s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Our aim is to find a near-optimal policy with a polynomial number of queries to the generative model, by running a version of gradient descent-ascent on the Lagrangian $\mathcal{L}$. In particular, we aim to derive a bound on the suboptimality of the output policy in terms of the optimization-error guarantee that we obtain by running our algorithm. To achieve this, a key quantity to study is the expected gap of the averaged iterates $(\overline{\boldsymbol{\mu}}_T, \overline{\boldsymbol{v}}_T) \in \Delta_{SA} \times \mathcal{V}$ against arbitrary comparators $(\boldsymbol{\mu}^*, \boldsymbol{v}^*) \in \Delta_{SA} \times \mathcal{V}$ denoted as

$$\mathbb{E}\left[G(\boldsymbol{\mu}^*; \boldsymbol{v}^*)\right] = \mathbb{E}\left[\mathcal{L}(\boldsymbol{\mu}^*; \overline{\boldsymbol{v}}_T) - \mathcal{L}(\overline{\boldsymbol{\mu}}_T; \boldsymbol{v}^*)\right], \tag{13}$$

where $(\overline{\boldsymbol{\mu}}_T, \overline{\boldsymbol{v}}_T) = \left(\frac{1}{T}\sum_{t=1}^T \boldsymbol{\mu}_t, \frac{1}{T}\sum_{t=1}^T \boldsymbol{v}_t\right)$ and $\overline{\pi}_T$ are as described in the main text. Then, a relationship between the duality gap and the policy can be established by choosing the comparators as $(\boldsymbol{\mu}^*, \boldsymbol{v}^*) = (\boldsymbol{\mu}^{\pi^*}, \boldsymbol{v}^{\overline{\pi}_T}) \in \Delta_{SA} \times \mathbb{R}^S$. Indeed, as we show in Lemma C.2 (a result adapted from Cheng et al., 2020), the two quantities under this choice can be related as

$$\mathbb{E}\left[G(\boldsymbol{\mu}^{\pi^*}; \boldsymbol{v}^{\overline{\pi}_T})\right] = \mathbb{E}\left[\left\langle \boldsymbol{\mu}^{\pi^*} - \boldsymbol{\mu}^{\overline{\pi}_T}, \boldsymbol{r}\right\rangle\right]. \tag{14}$$

B.2. Methodology

In order to apply standard OMD to solve Equation (12), previous LP-based approaches to planning in finite AMDPs (Wang, 2017; Jin & Sidford, 2020) required the domain $\mathcal{V}$ to cover $\boldsymbol{v}^*$, which requires prior knowledge of the properties of the MDP. To this end, they made the assumption that the value functions of all policies have bounded *span seminorm*: for all policies π, the value function $\boldsymbol{v}^\pi$ satisfies $\|\boldsymbol{v}^\pi\|_{\text{sp}} = \max_s \boldsymbol{v}^\pi(s) - \min_{s'} \boldsymbol{v}^\pi(s') \le B$ for some $B > 0$. We call this quantity the *worst-case bias span*. A simple way to make sure that the above assumption holds is to suppose that the Markov chains induced by each policy π have bounded *mixing time* t_{mix}, defined as

$$t_{\text{mix}} = \max_\pi \left[\arg\min_{t \le 1} \left\{\max_{\boldsymbol{\nu}\in\Delta_S} \left\|\boldsymbol{\nu}^\top (\boldsymbol{P}^\pi)^t - \boldsymbol{\nu}^\pi\right\|_1\right\}\right].$$

This ensures that the supremum norm of the value of any policy is bounded above with $\|\boldsymbol{v}^\pi\|_\infty \le 2t_{\text{mix}}$. Previous works of Wang (2017); Jin & Sidford (2020) assumed this mixing-time parameter to be known, and designed iterative algorithms that require projections to the set $\mathcal{V}_B = \left\{\boldsymbol{v} \in \mathbb{R}^S : \|\boldsymbol{v}\|_\infty \le 2t_{\text{mix}}\right\}$. Since this parameter is typically unknown and is hard to estimate, these algorithms are not fully satisfactory.

We are interested in near-optimal planning in general AMDPs for which the stationary state distribution is well defined and bias span is potentially unknown, and thus we have to set $\mathcal{V} = \mathbb{R}^S$. Since the primal variables are naturally restricted to the simplex domain, we only require the stabilization trick to control the actions of the min-player in the bound. Hence, we can bound the duality gap against arbitrary comparator points $(\boldsymbol{v}^*; \boldsymbol{\mu}^*)$ as follows:

$$\begin{aligned}
\mathbb{E}\left[G(\boldsymbol{v}^*; \boldsymbol{\mu}^*)\right] &= \mathbb{E}\left[\mathcal{L}(\overline{\boldsymbol{v}}_T; \boldsymbol{\mu}^*) - \mathcal{L}(\boldsymbol{v}^*; \overline{\boldsymbol{\mu}}_T)\right] \\
&\le \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\mathcal{L}\left(\boldsymbol{v}_t; \boldsymbol{\mu}^*\right) - \mathcal{L}\left(\boldsymbol{v}^*; \boldsymbol{\mu}_t\right)\right] \\
&= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\mathcal{L}\left(\boldsymbol{v}_t; \boldsymbol{\mu}^*\right) - \mathcal{L}\left(\boldsymbol{v}_t; \boldsymbol{\mu}_t\right)\right] + \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\mathcal{L}\left(\boldsymbol{v}_t; \boldsymbol{\mu}_t\right) - \mathcal{L}\left(\boldsymbol{v}^*; \boldsymbol{\mu}_t\right)\right] \\
&= \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\mathcal{L}\left(\boldsymbol{v}_t; \boldsymbol{\mu}^*\right) - \mathcal{L}\left(\boldsymbol{v}_t; \boldsymbol{\mu}_t\right)\right] + \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\mathcal{L}^{(\text{reg})}\left(\boldsymbol{v}_t; \boldsymbol{\mu}_t\right) - \mathcal{L}^{(\text{reg})}\left(\boldsymbol{v}^*; \boldsymbol{\mu}_t\right)\right]
\end{aligned}$$

$$+\frac{\varrho_v}{T}\sum_{t=1}^{T}\mathbb{E}\left[\mathcal{H}_v(\boldsymbol{v}^*)-\mathcal{H}_v(\boldsymbol{v}_t)\right], \tag{15}$$

where we have defined $\mathcal{L}^{(\text{reg})}\left(\boldsymbol{v};\boldsymbol{\mu}\right)=\mathcal{L}\left(\boldsymbol{v};\boldsymbol{\mu}\right)+\varrho_v\mathcal{H}_v(\boldsymbol{v})$.

Taking into account the new (unregularized) loss objective of the max-player and required projections to the simplex, our algorithm executes **COMID** to optimize $\boldsymbol{v}$ and standard OMD (which is same as **COMIDA** with $\varrho_\mu=0$) for $\boldsymbol{\mu}$. Precisely, the updates are calculated by solving

$$\boldsymbol{v}_{t+1}=\underset{\boldsymbol{v}\in\mathbb{R}^S}{\arg\min}\left\{\langle\boldsymbol{v},\widetilde{\boldsymbol{g}}_v(t)\rangle+\varrho_v\left\|\boldsymbol{v}\right\|_\infty^2+\frac{1}{2\eta_v}\left\|\boldsymbol{v}-\boldsymbol{v}_t\right\|_2^2\right\}$$
$$\boldsymbol{\mu}_{t+1}=\underset{\boldsymbol{\mu}\in\Delta_{SA}}{\arg\min}\left\{-\left\langle\boldsymbol{\mu},\widetilde{\boldsymbol{g}}_\mu(t)\right\rangle+\frac{1}{\eta_\mu}\mathcal{D}_{\text{KL}}\left(\boldsymbol{\mu}\|\boldsymbol{\mu}_t\right)\right\},$$

using the gradient estimators described in the main text. We present the complete pseudocode as Algorithm 1.

Algorithm 1 COMIDA-MDP

```
Input: Step sizes $\eta_v, \eta_\mu$, Regularization constants $\varrho_v$, Initial points $\boldsymbol{v}_1, \boldsymbol{\mu}_1$.
for $t = 1$ to $T$ do
  //Mirror Descent//
  Sample $(s_t, a_t) \sim \boldsymbol{\mu}_t$, $s'_t \sim p(\cdot|s_t, a_t)$
  Compute $\widetilde{\boldsymbol{g}}_v(t) = \boldsymbol{e}_{s'_t} - \boldsymbol{e}_{s_t}$
  Update
      $\boldsymbol{v}_{t+1} = \arg\min_{\boldsymbol{v}\in\mathbb{R}^S}\left\{\langle\boldsymbol{v},\widetilde{\boldsymbol{g}}_v(t)\rangle+\varrho_v\|\boldsymbol{v}\|_\infty^2+\frac{1}{2\eta_v}\|\boldsymbol{v}-\boldsymbol{v}_t\|_2^2\right\}$
  //Mirror Ascent//
  Sample $\overline{s}'_t \sim p(\cdot|s,a)$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$
  Compute $\widetilde{\boldsymbol{g}}_\mu(t) = \sum_{(s,a)}[r(s,a) + v_t(\overline{s}'_t) - v_t(s)]\boldsymbol{e}_{(s,a)}$
  Update
      $\boldsymbol{\mu}_{t+1} = \arg\min_{\boldsymbol{\mu}\in\Delta_{SA}}\left\{-\left\langle\boldsymbol{\mu},\widetilde{\boldsymbol{g}}_\mu(t)\right\rangle+\frac{1}{\eta_\mu}\mathcal{D}_{\text{KL}}(\boldsymbol{\mu}\|\boldsymbol{\mu}_t)\right\}$
end for
Return  $\overline{\boldsymbol{v}}_T = \frac{1}{T}\sum_{t=1}^T \boldsymbol{v}_t$, $\overline{\boldsymbol{\mu}}_T = \frac{1}{T}\sum_{t=1}^T \boldsymbol{\mu}_t$.
```

C. The proof of Theorem 4.1

We restate the result here for convenience of the reader.

Theorem C.1. *Let $\varrho_v=\eta_\mu$. Then, the output of* **COMIDA-MDP** *satisfies the following bound:*

$$\mathbb{E}\left[\left\langle\boldsymbol{\mu}^{\pi^*}-\boldsymbol{\mu}^{\overline{\pi}_T},\boldsymbol{r}\right\rangle\right]\leq\frac{\mathcal{D}_{\text{KL}}\left(\boldsymbol{\mu}^{\pi^*}\|\boldsymbol{\mu}_1\right)}{\eta_\mu T}+\eta_\mu+\left(\frac{1}{\eta_v T}+4\eta_\mu\right)\mathbb{E}\left[\left\|\boldsymbol{v}^{\overline{\pi}_T}\right\|_2^2\right]+4\eta_v$$

We start by stating a useful result (which we have learned from Cheng et al., 2020) that connects the duality gap with the suboptimality of the policy output by the algorithm.

Lemma C.2. *(cf. Proposition 4 of Cheng et al., 2020) The duality gap at $(\overline{\boldsymbol{\mu}}_T,\overline{\boldsymbol{v}}_T)$ satisfies*

$$G(\boldsymbol{\mu}^{\pi^*},\boldsymbol{v}^{\overline{\pi}_T})=\mathcal{L}(\boldsymbol{\mu}^{\pi^*};\overline{\boldsymbol{v}}_T)-\mathcal{L}(\overline{\boldsymbol{\mu}}_T;\boldsymbol{v}^{\overline{\pi}_T})=\rho^*-\rho^{\overline{\pi}_T}.$$

Proof. From Equation (13), recall that

$$G(\boldsymbol{\mu}^{\pi^*},\boldsymbol{v}^{\overline{\pi}_T})=\mathcal{L}(\boldsymbol{\mu}^{\pi^*};\overline{\boldsymbol{v}}_T)-\mathcal{L}(\overline{\boldsymbol{\mu}}_T;\boldsymbol{v}^{\overline{\pi}_T}). \tag{16}$$

By definition of the Lagrangian, we can write

$$\mathcal{L}(\boldsymbol{\mu}^{\pi^*};\overline{\boldsymbol{v}}_T) = \left\langle \boldsymbol{\mu}^{\pi^*}, \boldsymbol{r} \right\rangle + \left\langle \overline{\boldsymbol{v}}_T, \boldsymbol{P}^\intercal \boldsymbol{\mu}^{\pi^*} - \boldsymbol{E}^\intercal \boldsymbol{\mu}^{\pi^*} \right\rangle = \left\langle \boldsymbol{\mu}^{\pi^*}, \boldsymbol{r} \right\rangle,$$

since $\boldsymbol{\mu}^{\pi^*}$ is a valid stationary distribution that satisfies $\boldsymbol{P}^\intercal \boldsymbol{\mu}^{\pi^*} = \boldsymbol{E}^\intercal \boldsymbol{\mu}^{\pi^*}$. On the other hand, using that $\overline{\boldsymbol{\mu}}_T \in \Delta_{SA}$ and rearranging terms we have that:

$$\begin{aligned}
\mathcal{L}(\overline{\boldsymbol{\mu}}_T;\boldsymbol{v}^{\overline{\pi}_T}) &= \langle \overline{\boldsymbol{\mu}}_T, \boldsymbol{r} \rangle + \left\langle \boldsymbol{v}^{\overline{\pi}_T}, \boldsymbol{P}^\intercal \overline{\boldsymbol{\mu}}_T - \boldsymbol{E}^\intercal \overline{\boldsymbol{\mu}}_T \right\rangle + \rho^{\overline{\pi}_T}(1 - \langle \overline{\boldsymbol{\mu}}_T, \boldsymbol{1} \rangle) \\
&= \left\langle \overline{\boldsymbol{\mu}}_T, \boldsymbol{r} + \boldsymbol{P}\boldsymbol{v}^{\overline{\pi}_T} - \boldsymbol{E}\boldsymbol{v}^{\overline{\pi}_T} - \rho^{\overline{\pi}_T}\boldsymbol{1} \right\rangle + \rho^{\overline{\pi}_T} \\
&= \sum_{s,a}\sum_{a'} \overline{\boldsymbol{\mu}}_T(s,a')\overline{\pi}_T(a|s)\Big(r(s,a) + \left\langle p(\cdot|s,a), \boldsymbol{v}^{\overline{\pi}_T} \right\rangle - \boldsymbol{v}^{\overline{\pi}_T}(s) - \rho^{\overline{\pi}_T} \Big) + \rho^{\overline{\pi}_T} = \rho^{\overline{\pi}_T},
\end{aligned}$$

where the last equality holds by definition of $\overline{\pi}_T$ in the main text and the value functions in Equation (10). Combining both expressions in Equation (16) gives the desired result. □

C.1. Proof of Theorem 4.1

First, we prove that the gradient norms are bounded. By definition of the gradients,

$$\mathbb{E}_t\left[\|\widetilde{\boldsymbol{g}}_v(t)\|_2^2\right] = \mathbb{E}_t\left[\left\|\boldsymbol{e}_{s'_t} - \boldsymbol{e}_{s_t}\right\|_2^2\right] = \mathbb{E}_t\left[1 - 2\mathbb{I}_{\{s'_t = s_t\}} + 1\right] \leq 2. \tag{17}$$

Also, using that $r(s,a) \in [0,1]$ for any $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\left\|\widetilde{\boldsymbol{g}}_\mu(t)\right\|_\infty^2 \leq \max_{(s,a,s')} \left|r(s,a) + v_t(s') - v_t(s)\right|^2 \leq \left(1 + 2\left\|\boldsymbol{v}_t\right\|_\infty\right)^2 \leq 2 + 8\left\|\boldsymbol{v}_t\right\|_\infty^2, \tag{18}$$

where the last inequality is Cauchy–Schwarz.

In what follows, we let $\boldsymbol{v}^* = \boldsymbol{v}^{\overline{\pi}_T}$, and derive a bound on the duality gap evaluated at this comparator point. We start by appealing to Lemma C.2 and decomposing the duality gap as follows:

$$\begin{aligned}
\rho^* - \rho^{\overline{\pi}_T} = \mathbb{E}\left[G(\boldsymbol{v}^*;\boldsymbol{\mu}^*)\right] \leq \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\mathcal{L}\left(\boldsymbol{v}_t, \boldsymbol{\mu}^*\right) - \mathcal{L}\left(\boldsymbol{v}_t, \boldsymbol{\mu}_t\right)\right] \\
+ \frac{1}{T}\sum_{t=1}^T \mathbb{E}\left[\mathcal{L}^{(\mathrm{reg})}\left(\boldsymbol{v}_t, \boldsymbol{\mu}_t\right) - \mathcal{L}^{(\mathrm{reg})}\left(\boldsymbol{v}^*, \boldsymbol{\mu}_t\right)\right] \\
+ \frac{\varrho_v}{T}\sum_{t=1}^T \mathbb{E}\left[\|\boldsymbol{v}^*\|_\infty^2 - \|\boldsymbol{v}_t\|_\infty^2\right].
\end{aligned} \tag{19}$$

By the standard online mirror descent analysis, we obtain the following upper bound on the first term that corresponds to the regret of the μ-player:

$$\begin{aligned}
\sum_{t=1}^T \mathbb{E}\left[\mathcal{L}\left(\boldsymbol{v}_t;\boldsymbol{\mu}^*\right) - \mathcal{L}\left(\boldsymbol{v}_t;\boldsymbol{\mu}_t\right)\right] &\overset{(a)}{\leq} \sum_{t=1}^T \mathbb{E}\left[\left\langle \boldsymbol{\mu}^* - \boldsymbol{\mu}_t, \widetilde{\boldsymbol{g}}_\mu(t) \right\rangle\right] \\
&\overset{(b)}{\leq} \frac{\mathcal{D}_{\mathrm{KL}}\left(\boldsymbol{\mu}^*\|\boldsymbol{\mu}_1\right)}{\eta_\mu} + \frac{\eta_\mu}{2}\sum_{t=1}^T \mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_\mu(t)\right\|_\infty^2\right] \\
&\overset{(c)}{\leq} \frac{\mathcal{D}_{\mathrm{KL}}\left(\boldsymbol{\mu}^*\|\boldsymbol{\mu}_1\right)}{\eta_\mu} + \eta_\mu \sum_{t=1}^T \mathbb{E}\left[1 + 4\left\|\boldsymbol{v}_t\right\|_\infty^2\right]
\end{aligned} \tag{20}$$

Here, we have used *(a)* Definition 2.2, *(b)* Corollary D.2 with $\mathcal{U} = \Delta_{SA}$, $\ell_t(\cdot) = -\mathcal{L}(\boldsymbol{v}_t;\cdot)$, $\mathcal{D}_u(\boldsymbol{u}\|\boldsymbol{u}') = \mathcal{D}_{\mathrm{KL}}(\boldsymbol{u}\|\boldsymbol{u}')$, $\varrho_u = 0$ and $\boldsymbol{u}_1 = \boldsymbol{\mu}_1$, as well as (c) the bound on the gradient norm established in Equation (18).

As for the second term that corresponds to the regret of the v-player, the analysis is somewhat more involved. One challenge is that the comparator point $\boldsymbol{v}^* = \boldsymbol{v}^{\overline{\pi}_t}$ is dependent on the iterates. This will be addressed at the end of this analysis. For now, applying Lemma D.1 with the appropriate parameters including $\mathcal{D}_u(\boldsymbol{u}\|\boldsymbol{u}') = \frac{1}{2}\left\|\boldsymbol{u}-\boldsymbol{u}'\right\|_2^2$, $\mathcal{H}_u\left(\boldsymbol{u}\right) = \mathcal{H}_v\left(\boldsymbol{u}\right) = \left\|\boldsymbol{u}\right\|_\infty^2$, $\boldsymbol{u}_1 = \boldsymbol{v}_1 = \boldsymbol{0}$ so that $\mathcal{H}_u\left(\boldsymbol{u}_1\right) = 0$, as well as noting that the squared Euclidean norm is 1-strongly convex and is the dual norm of itself gives the following bound:

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}&\left[\mathcal{L}^{(\text{reg})}\left(\boldsymbol{v}_t;\boldsymbol{\mu}_t\right)-\mathcal{L}^{(\text{reg})}\left(\boldsymbol{v}^*;\boldsymbol{\mu}_t\right)\right]\\&\leq\frac{\mathbb{E}\left[\mathcal{D}_v(\boldsymbol{v}^*\|\boldsymbol{v}_1)\right]}{\eta_v}+\frac{\eta_v}{2}\sum_{t=1}^{T}\mathbb{E}\left[\|\widetilde{\boldsymbol{g}}_v(t)\|_2^2\right]+\sum_{t=1}^{T}\mathbb{E}\left[\langle\boldsymbol{g}_v(t)-\widetilde{\boldsymbol{g}}_v(t),\boldsymbol{v}^*-\boldsymbol{v}_t\rangle\right]\\&\leq\frac{1}{\eta_v}\mathbb{E}\left[\mathcal{D}_v(\boldsymbol{v}^*\|\boldsymbol{v}_1)\right]+2\eta_v T+\sum_{t=1}^{T}\mathbb{E}\left[\langle\boldsymbol{g}_v(t)-\widetilde{\boldsymbol{g}}_v(t),\boldsymbol{v}^*-\boldsymbol{v}_t\rangle\right].\end{aligned}$$

The last inequality follows from using that $\mathbb{E}\left[\|\widetilde{\boldsymbol{g}}_v(t)\|_2^2\right]\leq 2$.

We are left with the problem of bounding the last term. We first observe that $\mathbb{E}\left[\sum_{t=1}^{T}\langle\boldsymbol{g}_v(t)-\widetilde{\boldsymbol{g}}_v(t),\boldsymbol{v}_t\rangle\right]=0$ due to the $\mathcal{F}_{t-1}$-measurability of $\boldsymbol{v}_t$, so in fact all that remains is bounding $\mathbb{E}\left[\sum_{t=1}^{T}\langle\boldsymbol{g}_v(t)-\widetilde{\boldsymbol{g}}_v(t),\boldsymbol{v}^*\rangle\right]$. This requires some care because $\boldsymbol{v}^*=\boldsymbol{v}^{\overline{\pi}_T}$ depends on the entire sequence of iterates, and thus we cannot make direct use of the fact that $\mathbb{E}_t\left[\boldsymbol{g}_v(t)-\widetilde{\boldsymbol{g}}_v(t)\right]=0$. We will address this issue via a reduction to online learning inspired by the martingale tail bounds of Rakhlin & Sridharan (2017).

To this end, let us construct an auxiliary online learning game where in each round $t=1,2,\dots,T$, the following steps are repeated between an online learner and its environment:

- The online learner chooses a function $\widehat{\boldsymbol{v}}_t\in\mathbb{R}^S$,
- the environment chooses the cost function $\boldsymbol{c}_t=\widetilde{\boldsymbol{g}}_v(t)-\boldsymbol{g}_v(t)$,
- the online learner incurs cost $\langle\boldsymbol{c}_t,\widehat{\boldsymbol{v}}_t\rangle$ and observes $\boldsymbol{c}_t$.

By construction, we have that $\left\|\boldsymbol{c}_t\right\|_2\leq\left\|\boldsymbol{c}_t\right\|_1\leq 2$. We will study the case below where the online learner executes online gradient descent on the sequence $\{\boldsymbol{c}_t\}_{t=1}^T$, initialized at $\widehat{\boldsymbol{v}}_1=\boldsymbol{v}_1$ and using stepsize η_v. By standard techniques (e.g., Lemma D.1 with $\varrho_u=0$ and), the regret of this method can be bounded against any comparator $\boldsymbol{v}^*$ as follows:

$$\sum_{t=1}^{T}\langle\boldsymbol{c}_t,\widehat{\boldsymbol{v}}_t-\boldsymbol{v}^*\rangle\leq\frac{\left\|\boldsymbol{v}^*-\boldsymbol{v}_1\right\|^2}{2\eta_v}+\frac{\eta_v}{2}\sum_{t=1}^{T}\left\|\boldsymbol{c}_t\right\|_2^2\leq\frac{\left\|\boldsymbol{v}^*-\boldsymbol{v}_1\right\|^2}{2\eta_v}+2\eta_v T.$$

Then, the term we seek to bound can be controlled as follows:

$$\begin{aligned}\mathbb{E}\left[\sum_{t=1}^{T}\langle\boldsymbol{g}_v(t)-\widetilde{\boldsymbol{g}}_v(t),\boldsymbol{v}^*\rangle\right]&=-\mathbb{E}\left[\sum_{t=1}^{T}\langle\boldsymbol{c}_t,\boldsymbol{v}^*\rangle\right]=\mathbb{E}\left[\sum_{t=1}^{T}\langle\boldsymbol{c}_t,\widehat{\boldsymbol{v}}_t-\boldsymbol{v}^*\rangle\right]-\mathbb{E}\left[\sum_{t=1}^{T}\langle\boldsymbol{c}_t,\widehat{\boldsymbol{v}}_t\rangle\right]\\&\leq\frac{\mathbb{E}\left[\left\|\boldsymbol{v}^*-\boldsymbol{v}_1\right\|^2\right]}{2\eta_v}+2\eta_v T,\end{aligned}$$

where in the last step we used our regret bound stated just above, and also that $\mathbb{E}\left[\sum_{t=1}^{T}\langle\boldsymbol{c}_t,\widehat{\boldsymbol{v}}_t\rangle\right]=0$, which holds because the sum in question is a martingale. Indeed, note that for any t, we have

$$\mathbb{E}_t\left[\langle\boldsymbol{c}_t,\widehat{\boldsymbol{v}}_t\rangle\right]=\langle\mathbb{E}_t\left[\boldsymbol{c}_t\right],\widehat{\boldsymbol{v}}_t\rangle=\langle 0,\widehat{\boldsymbol{v}}_t\rangle=0,$$

which holds due to the fact that $\widehat{\boldsymbol{v}}_t$ was chosen before $\boldsymbol{c}_t$ was revealed to the online learner. Overall, this proves

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle\boldsymbol{g}_v(t)-\widetilde{\boldsymbol{g}}_v(t),\boldsymbol{v}^*-\boldsymbol{v}_t\rangle\right]\leq\frac{\mathbb{E}\left[\left\|\boldsymbol{v}^*-\boldsymbol{v}_1\right\|^2\right]}{2\eta_v}+2\eta_v T,$$

thus verifying the inequality

$$\sum_{t=1}^{T} \mathbb{E}\left[\mathcal{L}^{(\mathrm{reg})}\left(\boldsymbol{v}_t;\boldsymbol{\mu}_t\right) - \mathcal{L}^{(\mathrm{reg})}\left(\boldsymbol{v}^*;\boldsymbol{\mu}_t\right)\right] \leq \frac{\mathbb{E}\left[\left\|\boldsymbol{v}^* - \boldsymbol{v}_1\right\|_2^2\right]}{\eta_v} + 4\eta_v T.$$

Putting the above inequality together with Equations (19) and (20), we finally obtain the following bound:

$$\mathbb{E}\left[G(\boldsymbol{v}^*;\boldsymbol{\mu}^*)\right] \leq \frac{\mathcal{D}_{\mathrm{KL}}\left(\boldsymbol{\mu}^* \| \boldsymbol{\mu}_1\right)}{\eta_\mu T} + \frac{\eta_\mu}{T}\sum_{t=1}^{T}\mathbb{E}\left[1 + 4\left\|\boldsymbol{v}_t\right\|_\infty^2\right] + \frac{\mathbb{E}\left[\left\|\boldsymbol{v}^* - \boldsymbol{v}_1\right\|_2^2\right]}{\eta_v T} + 4\eta_v + \frac{\varrho_v}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left\|\boldsymbol{v}^*\right\|_\infty^2 - \left\|\boldsymbol{v}_t\right\|_\infty^2\right].$$

Recalling the choice $\boldsymbol{v}_1 = 0$, choosing $\varrho_v = 4\eta_\mu$, and bounding $\left\|\boldsymbol{v}^*\right\|_\infty \leq \left\|\boldsymbol{v}^*\right\|_2$ we obtain the result claimed in the theorem. □

D. Auxiliary Lemmas

Lemma D.1. *(cf. Theorem 8 of Duchi et al., 2010) Let $\ell_t : \mathcal{U} \to \mathbb{R}$ be convex, $\boldsymbol{g}_{\boldsymbol{u}}(t) \in \partial \ell_t(\boldsymbol{u}_t)$ and $\widetilde{\boldsymbol{g}}_u(t)$ be such that $\mathbb{E}_t\left[\widetilde{\boldsymbol{g}}_u(t)\right] = \boldsymbol{g}_{\boldsymbol{u}}(t)$. Given $\boldsymbol{u}_1 \in \mathcal{U}$, define $\tilde{\boldsymbol{g}}_u(1) \in \mathbb{R}^m$ and the sequence of vectors $\{(\boldsymbol{u}_t, \widetilde{\boldsymbol{g}}_u(t))\}_{t=2}^T$ via the following recursion for $t \in [T]$:*

$$\boldsymbol{u}_{t+1} = \underset{\boldsymbol{u} \in \mathcal{U}}{\textit{arg min}} \left\{ \langle \boldsymbol{u}, \widetilde{\boldsymbol{g}}_u(t) \rangle + \varrho_u \mathcal{H}_u\left(\boldsymbol{u}\right) + \frac{1}{\eta_u} \mathcal{D}_u(\boldsymbol{u} \| \boldsymbol{u}_t) \right\}. \tag{21}$$

Suppose the distance-generating function ω_u is γ_u-strongly convex with respect to $\|\cdot\|_u$. For any $\boldsymbol{u}^* \in \mathcal{U}$,

$$\begin{aligned}
&\sum_{t=1}^T \mathbb{E}\left[\ell_t^{(reg)}\left(\boldsymbol{u}_t\right) - \ell_t^{(reg)}\left(\boldsymbol{u}^*\right)\right] \\
&\qquad \leq \frac{\mathbb{E}\left[\mathcal{D}_u(\boldsymbol{u}^* \| \boldsymbol{u}_1)\right]}{\eta_u} + \frac{\eta_u}{2\gamma_u} \sum_{t=1}^T \mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_u(t)\right\|_{u,*}^2\right] + \varrho_u \mathbb{E}\left[\mathcal{H}_u(\boldsymbol{u}_1)\right] + \sum_{t=1}^T \mathbb{E}\left[\langle \boldsymbol{g}_{\boldsymbol{u}}(t) - \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}^* - \boldsymbol{u}_t \rangle\right].
\end{aligned}$$

where

$$\ell_t^{(reg)}\left(\boldsymbol{u}\right) = \ell_t\left(\boldsymbol{u}\right) + \varrho_u \mathcal{H}_u\left(\boldsymbol{u}\right). \tag{22}$$

Proof. Using the definition of $\ell_t^{(\text{reg})}$, consider the regret in terms of the regularized loss:

$$\begin{aligned}
\ell_t^{(\text{reg})}\left(\boldsymbol{u}_t\right) - \ell_t^{(\text{reg})}\left(\boldsymbol{u}^*\right) &= \ell_t\left(\boldsymbol{u}_t\right) - \ell_t\left(\boldsymbol{u}^*\right) + \varrho_u \mathcal{H}_u(\boldsymbol{u}_t) - \varrho_u \mathcal{H}_u(\boldsymbol{u}^*) \\
&= \Big(\ell_t\left(\boldsymbol{u}_t\right) - \ell_t\left(\boldsymbol{u}^*\right) + \varrho_u \mathcal{H}_u(\boldsymbol{u}_{t+1}) - \varrho_u \mathcal{H}_u(\boldsymbol{u}^*)\Big) + \varrho_u \Big(\mathcal{H}_u(\boldsymbol{u}_t) - \mathcal{H}_u(\boldsymbol{u}_{t+1})\Big).
\end{aligned}$$

To proceed, we let $\boldsymbol{h}_{\boldsymbol{u}}(t+1) \in \partial \mathcal{H}_u(\boldsymbol{u}_{t+1})$, so that we can use the convexity of ℓ_t and $\mathcal{H}_u$ to bound the first set of terms as

$$\begin{aligned}
&\ell_t\left(\boldsymbol{u}_t\right) - \ell_t\left(\boldsymbol{u}^*\right) + \varrho_u \mathcal{H}_u(\boldsymbol{u}_{t+1}) - \varrho_u \mathcal{H}_u(\boldsymbol{u}^*) \\
&\quad \leq \langle \boldsymbol{g}_{\boldsymbol{u}}(t), \boldsymbol{u}_t - \boldsymbol{u}^* \rangle + \varrho_u \langle \boldsymbol{h}_{\boldsymbol{u}}(t+1), \boldsymbol{u}_{t+1} - \boldsymbol{u}^* \rangle \\
&\quad = \langle \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}_t - \boldsymbol{u}^* \rangle + \varrho_u \langle \boldsymbol{h}_{\boldsymbol{u}}(t+1), \boldsymbol{u}_{t+1} - \boldsymbol{u}^* \rangle + \langle \boldsymbol{g}_{\boldsymbol{u}}(t) - \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}_t - \boldsymbol{u}^* \rangle.
\end{aligned} \tag{23}$$

Before we proceed to bound the first two terms, note that $\boldsymbol{u}_{t+1}$ in Equation (21) is a solution to a constrained convex optimization problem, and as a result it satisfies the following optimality condition for any $\boldsymbol{u} \in \mathcal{U}$:

$$\left\langle \boldsymbol{u} - \boldsymbol{u}_{t+1}, \widetilde{\boldsymbol{g}}_u(t) + \varrho_u \boldsymbol{h}_{\boldsymbol{u}}(t+1) + \frac{1}{\eta_u}\left(\nabla \omega_u(\boldsymbol{u}_{t+1}) - \nabla \omega_u(\boldsymbol{u}_t)\right) \right\rangle \geq 0. \tag{24}$$

Thus, we bound the first two terms on the right-hand side of the inequality (23) as follows:

$$\begin{aligned}
&\langle \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}_t - \boldsymbol{u}^* \rangle + \varrho_u \langle \boldsymbol{h}_{\boldsymbol{u}}(t+1), \boldsymbol{u}_{t+1} - \boldsymbol{u}^* \rangle \\
&\quad = \langle \widetilde{\boldsymbol{g}}_u(t) + \varrho_u \boldsymbol{h}_{\boldsymbol{u}}(t+1), \boldsymbol{u}_{t+1} - \boldsymbol{u}^* \rangle + \langle \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}_t - \boldsymbol{u}_{t+1} \rangle \\
&\quad = \left\langle \widetilde{\boldsymbol{g}}_u(t) + \varrho_u \boldsymbol{h}_{\boldsymbol{u}}(t+1) + \frac{1}{\eta_u}\left(\nabla \omega_u(\boldsymbol{u}_{t+1}) - \nabla \omega_u(\boldsymbol{u}_t)\right), \boldsymbol{u}_{t+1} - \boldsymbol{u}^* \right\rangle \\
&\qquad + \frac{1}{\eta_u} \langle \nabla \omega_u(\boldsymbol{u}_{t+1}) - \nabla \omega_u(\boldsymbol{u}_t), \boldsymbol{u}^* - \boldsymbol{u}_{t+1} \rangle + \langle \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}_t - \boldsymbol{u}_{t+1} \rangle \\
&\quad \overset{(a)}{\leq} \frac{1}{\eta_u} \langle \nabla \omega_u(\boldsymbol{u}_{t+1}) - \nabla \omega_u(\boldsymbol{u}_t), \boldsymbol{u}^* - \boldsymbol{u}_{t+1} \rangle + \langle \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}_t - \boldsymbol{u}_{t+1} \rangle \\
&\quad \overset{(b)}{=} \frac{1}{\eta_u}\Big(\mathcal{D}_u(\boldsymbol{u}^* \| \boldsymbol{u}_t) - \mathcal{D}_u(\boldsymbol{u}^* \| \boldsymbol{u}_{t+1})\Big) - \frac{1}{\eta_u} \mathcal{D}_u(\boldsymbol{u}_{t+1} \| \boldsymbol{u}_t) + \langle \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}_t - \boldsymbol{u}_{t+1} \rangle \\
&\quad \overset{(c)}{\leq} \frac{1}{\eta_u}\Big(\mathcal{D}_u(\boldsymbol{u}^* \| \boldsymbol{u}_t) - \mathcal{D}_u(\boldsymbol{u}^* \| \boldsymbol{u}_{t+1})\Big) - \frac{\gamma_u}{2\eta_u} \left\|\boldsymbol{u}_{t+1} - \boldsymbol{u}_t\right\|_u^2 + \langle \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}_t - \boldsymbol{u}_{t+1} \rangle \\
&\quad \leq \frac{1}{\eta_u}\Big(\mathcal{D}_u(\boldsymbol{u}^* \| \boldsymbol{u}_t) - \mathcal{D}_u(\boldsymbol{u}^* \| \boldsymbol{u}_{t+1})\Big) + \frac{\gamma_u}{\eta_u} \sup_{\boldsymbol{u}} \left(\left\langle \frac{\eta_u}{\gamma_u} \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u} \right\rangle - \frac{1}{2} \left\|\boldsymbol{u}\right\|_u^2 \right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(d)}{=} \frac{1}{\eta_u}\Big(\mathcal{D}_u(\boldsymbol{u}^*\|\boldsymbol{u}_t) - \mathcal{D}_u(\boldsymbol{u}^*\|\boldsymbol{u}_{t+1})\Big) + \frac{\gamma_u}{2\eta_u}\left\|\frac{\eta_u}{\gamma_u}\widetilde{\boldsymbol{g}}_u(t)\right\|_{u,*}^2 \\
&= \frac{1}{\eta_u}\Big(\mathcal{D}_u(\boldsymbol{u}^*\|\boldsymbol{u}_t) - \mathcal{D}_u(\boldsymbol{u}^*\|\boldsymbol{u}_{t+1})\Big) + \frac{\eta_u}{2\gamma_u}\left\|\widetilde{\boldsymbol{g}}_u(t)\right\|_{u,*}^2 .
\end{aligned}
$$

We have used (a) the optimality condition stated in Equation (24), (b) the so-called *three-points identity* of Bregman divergences (cf. Lemma 4.1 in (Beck & Teboulle, 2003)), (c) the strong convexity of $\mathcal{D}_u(\cdot\|\boldsymbol{u}_t)$ and (d) the fact that for any norm $\|\cdot\|$, we have $\sup_{\boldsymbol{u}}\left\{\langle \boldsymbol{u}, \boldsymbol{g}\rangle - \frac{1}{2}\|\boldsymbol{u}\|^2\right\} = \frac{1}{2}\|\boldsymbol{g}\|_*^2$.

Thus, putting together all the above calculations, we arrive at the following bound:

$$
\begin{aligned}
&\ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u}^*) + \varrho_u\mathcal{H}_u(\boldsymbol{u}_{t+1}) - \varrho_u\mathcal{H}_u(\boldsymbol{u}^*) \\
&\qquad \le \langle \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}_t - \boldsymbol{u}^*\rangle + \varrho_u\langle \boldsymbol{h}_{\boldsymbol{u}}(t+1), \boldsymbol{u}_{t+1} - \boldsymbol{u}^*\rangle + \langle \boldsymbol{g}_{\boldsymbol{u}}(t) - \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}_t - \boldsymbol{u}^*\rangle \\
&\qquad \le \frac{1}{\eta_u}\Big(\mathcal{D}_u(\boldsymbol{u}^*\|\boldsymbol{u}_t) - \mathcal{D}_u(\boldsymbol{u}^*\|\boldsymbol{u}_{t+1})\Big) + \frac{\eta_u}{2\gamma_u}\left\|\widetilde{\boldsymbol{g}}_u(t)\right\|_{u,*}^2 + \langle \boldsymbol{g}_{\boldsymbol{u}}(t) - \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}_t - \boldsymbol{u}^*\rangle .
\end{aligned}
$$

Furthermore, plugging in the definition of $\ell_t^{(\mathrm{reg})}$ we get

$$
\begin{aligned}
\ell_t^{(\mathrm{reg})}(\boldsymbol{u}_t) - \ell_t^{(\mathrm{reg})}(\boldsymbol{u}^*) &= \left(\ell_t(\boldsymbol{u}_t) - \ell_t(\boldsymbol{u}^*) + \varrho_u\mathcal{H}_u(\boldsymbol{u}_{t+1}) - \varrho_u\mathcal{H}_u(\boldsymbol{u}^*)\right) + \varrho_u\Big(\mathcal{H}_u(\boldsymbol{u}_t) - \mathcal{H}_u(\boldsymbol{u}_{t+1})\Big) \\
&\le \frac{1}{\eta_u}\Big(\mathcal{D}_u(\boldsymbol{u}^*\|\boldsymbol{u}_t) - \mathcal{D}_u(\boldsymbol{u}^*\|\boldsymbol{u}_{t+1})\Big) + \frac{\eta_u}{2\gamma_u}\left\|\widetilde{\boldsymbol{g}}_u(t)\right\|_{u,*}^2 + \langle \boldsymbol{g}_{\boldsymbol{u}}(t) - \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}_t - \boldsymbol{u}^*\rangle \\
&\quad + \varrho_u\Big(\mathcal{H}_u(\boldsymbol{u}_t) - \mathcal{H}_u(\boldsymbol{u}_{t+1})\Big).
\end{aligned}
$$

Hence, taking marginal expectations on both sides, summing over $t = 1, \cdots, T$ steps, evaluating the telescoping terms and upper bounding some negative terms by zero, we finally obtain the following bound on the total regret of **COMID** on the regularized objective:

$$
\begin{aligned}
&\sum_{t=1}^T \mathbb{E}\left[\ell_t^{(\mathrm{reg})}(\boldsymbol{u}_t) - \ell_t^{(\mathrm{reg})}(\boldsymbol{u}^*)\right] \\
&\quad \le \frac{\mathbb{E}\left[\mathcal{D}_u(\boldsymbol{u}^*\|\boldsymbol{u}_1)\right]}{\eta_u} + \frac{\eta_u}{2\gamma_u}\sum_{t=1}^T \mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_u(t)\right\|_{u,*}^2\right] + \sum_{t=1}^T \mathbb{E}\left[\langle \boldsymbol{g}_{\boldsymbol{u}}(t) - \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}_t - \boldsymbol{u}^*\rangle\right] + \varrho_u\mathbb{E}\left[\mathcal{H}_u(\boldsymbol{u}_1)\right].
\end{aligned}
$$

This completes the proof. □

Corollary D.2. *Suppose the sequence of vectors $\{(\boldsymbol{x}_t, \widetilde{\boldsymbol{g}}_x(t))\}_{t=1}^T$ is as described above. If the comparator $\boldsymbol{x}^*$ is fixed and independent of the iterates, the following inequality holds:*

$$
\sum_{t=1}^T \mathbb{E}\left[\ell_t^{(\mathit{reg})}(\boldsymbol{u}_t) - \ell_t^{(\mathit{reg})}(\boldsymbol{u}^*)\right] \le \frac{\mathcal{D}_u(\boldsymbol{u}^*\|\boldsymbol{u}_1)}{\eta_u} + \frac{\eta_u}{2\gamma_u}\sum_{t=1}^T \mathbb{E}\left[\left\|\widetilde{\boldsymbol{g}}_u(t)\right\|_{u,*}^2\right] + \varrho_u\mathbb{E}\left[\mathcal{H}_u(\boldsymbol{u}_1)\right].
$$

Proof. Since $(\boldsymbol{u}_t, \boldsymbol{y}_t)$ is $\mathcal{F}_{t-1}$-measurable, $\mathbb{E}_t\left[\widetilde{\boldsymbol{g}}_u(t)\right] = \boldsymbol{g}_{\boldsymbol{u}}(t)$ and $\boldsymbol{u}^*$ does not depend on the iterates, the term $\langle \boldsymbol{g}_{\boldsymbol{u}}(t) - \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}^* - \boldsymbol{u}_t\rangle$ is zero in expectation. Precisely,

$$
\begin{aligned}
\mathbb{E}\left[\langle \boldsymbol{g}_{\boldsymbol{u}}(t) - \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}^* - \boldsymbol{u}_t\rangle\right] &= \mathbb{E}\left[\mathbb{E}_t\left[\langle \boldsymbol{g}_{\boldsymbol{u}}(t) - \widetilde{\boldsymbol{g}}_u(t), \boldsymbol{u}^* - \boldsymbol{u}_t\rangle\right]\right] \\
&= \mathbb{E}\left[\langle \mathbb{E}_t\left[\boldsymbol{g}_{\boldsymbol{u}}(t) - \widetilde{\boldsymbol{g}}_u(t)\right], \boldsymbol{u}^* - \boldsymbol{u}_t\rangle\right] = 0
\end{aligned}
$$

The stated result follows from this observation. □