Let's be Honest: An Optimal No-Regret Framework for Zero-Sum Games

Ehsan Asadi Kangarshahi [*1] **Ya-Ping Hsieh** [*1] **Mehmet Fatih Sahin** [1] **Volkan Cevher** [1]

Abstract

We revisit the problem of solving two-player zero-sum games in the decentralized setting. We propose a simple algorithmic framework that simultaneously achieves the best rates for honest regret as well as adversarial regret, and in addition resolves the open problem of removing the logarithmic terms in convergence to the value of the game. We achieve this goal in three steps. First, we provide a novel analysis of the optimistic mirror descent (OMD), showing that it can be modified to guarantee fast convergence for both honest regret and value of the game, when the players are playing collaboratively. Second, we propose a new algorithm, dubbed as robust optimistic mirror descent (ROMD), which attains optimal adversarial regret without knowing the time horizon beforehand. Finally, we propose a simple signaling scheme, which enables us to bridge OMD and ROMD to achieve the best of both worlds. Numerical examples are presented to support our theoretical claims and show that our non-adaptive ROMD algorithm can be competitive to OMD with adaptive step-size selection.

1. Introduction

The simple zero-sum games have been studied extensively, often from the standpoint of analyzing the convergence to the Nash equilibrium. At the equilibrium, the players employ a min-max pair of strategies where no player can improve their pay-off by a unilateral deviation (Von Neumann, 1928).

In this setting, one can expect that the players arrive at the equilibrium via decentralized, no-regret learning algorithms, which hold even in the presence of potential adversarial behavior, and which also better model selfish play. The resulting dynamics is of great interest in optimization and behavioral economics (Myerson, 1999), especially under communication constraints.

When the behavior of each player is explained by a no-regret algorithm, it is possible to significantly improve convergence rates beyond the so-called black-box, adversarial dynamics. This observation was first made by (Daskalakis et al., 2011), which tailored a decentralized version of Nesterov's primal-dual method based on the excessive gap condition.

Intriguingly, (Daskalakis et al., 2011) left it as an open question on the existence of a simple algorithm that converges at optimal rates for both regret and the value of the game in an uncoupled manner, both against honest (i.e., cooperative) and dishonest (i.e., arbitrarily adversarial) behavior.

The challenge was partially settled by the modified optimistic mirror descent (OMD) framework in (Rakhlin & Sridharan, 2013b). While the framework of (Daskalakis et al., 2011) is considered unnatural and involves additional logarithmic factors, similar arguments apply to Rakhlin & Sridharan (2013b)'s framework: The modified OMD needs to know the game horizon a priori to determine the step-sizes. Their analysis also results in non-optimal regret and logarithmic factors in convergence to the value of the game.

Besides the aforementioned drawbacks, neither approaches can accommodate natural switches between honest and dishonest behavior.

In this work, we propose a simple algorithmic framework that closes the gap between upper and lower bounds for adversarial regret as well as convergence to the value of the game, while maintaining the best known rate for honest regret, thereby resolving the open problem posed by (Daskalakis et al., 2011).

We achieve the desiderata as follows: First, we provide a novel analysis of OMD and show that it can obtain fast convergence for both honest regret and value of the game, when both players are honest. Second, we introduce robust optimistic mirror descent (ROMD), which attains optimal adversarial regret without knowing the time horizon. Finally, we propose a simple signaling scheme, which enables us to bridge OMD and ROMD to achieve the best of both worlds, and seamlessly handle honest and dishonest behavior.

[*]Equal contribution [1]LIONS, EPFL, Switzerland. Correspondence to: Ya-Ping Hsieh <ya-ping.hsieh@epfl.ch>, Volkan Cevher <volkan.cevher@epfl.ch>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018. Copyright 2018 by the author(s).

	Honest R_T	Adversarial R_T	Game Value	Oracle	Algorithm
Daskalakis et al. (2011)	$O(\log T)$	$O(\sqrt{T})$	$O(T^{-1}\log^{\frac{3}{2}} T)$	$\|A\|_{\max}$	Complicated
Rakhlin & Sridharan (2013b)	?	$O(\sqrt{T}\log T)$	$O\left(T^{-1}\log T\right)$	$T, \|A\|_{\max}$	Simple
This paper	$O(\log T)$	$O(\sqrt{T})$	$O\left(T^{-1}\right)$	$\|A\|_{\max}$	Simple

Table 1. A convergence rate comparison in the context of assumptions.

1.1. Related Work

Algorithms for Decentralized Games: To our knowledge, the only two explicit algorithms capable of solving zero-sum games in the decentralized setting are given by (Daskalakis et al., 2011) and (Rakhlin & Sridharan, 2013b), respectively. A comparison of their convergence rates versus ours is presented in Table 1.

The algorithm of (Daskalakis et al., 2011) is a decentralized primal-dual method based on Nesterov's excessive gap technique (Nesterov, 2005). Its convergence guarantees are only slightly worse than ours (*cf.,* Table 1). However, due to the presence of complicated and unnatural scheduling steps, the authors in (Daskalakis et al., 2011) themselves were not convinced by the practicality of their algorithm and stated the result as merely an "existence proof."

Later on, Rakhlin & Sridharan (2013b) proposed an algorithm based on the Optimistic Mirror Descent (OMD), initially introduced in a special case by (Chiang et al., 2012) and also studied in detail by (Rakhlin & Sridharan, 2013a). While the algorithm is simple, it features several drawbacks. Foremost, it requires the time horizon beforehand, which is unsatisfactory. Second, when both players are playing collaboratively, their regret is sub-optimal. Third, its adversarial regret and convergence to the game value has extra $\log T$ factors, which require additional cautions to remove. Finally, the algorithm uses *adaptive* step-sizes, requiring additional work per-iteration.

Meta-Algorithms: There exist some work on "meta-algorithms" for games (Syrgkanis et al., 2015; Foster et al., 2016), which can turn certain learning algorithms into solving zero-sum games. For instance, leveraging the framework in (Syrgkanis et al., 2015), one can modify OMD to achieve $O(T^{\frac{1}{4}})$ for honest regret + $\tilde{O}\left(\sqrt{T}\right)$ for adversarial regret. Our algorithm uniformly outperforms these rates.

2. Preliminaries and Notation

Let ψ be a mirror map over the convex domain $\mathcal{D}$, and let $D(\cdot,\cdot)$ be the Bregman divergence associated with ψ. We assume the knowledge of the three-point identity for Bregman divergence in the sequel:

$$D(\mathbf{x},\mathbf{y})+D(\mathbf{y},\mathbf{z}) = D(\mathbf{x},\mathbf{z})+\langle \mathbf{x}-\mathbf{y}, \nabla\psi(\mathbf{z})-\nabla\psi(\mathbf{y})\rangle\,.$$

We use the notation $\mathbf{z} = MD_\eta(\mathbf{x},\mathbf{g})$ to denote:

$$\mathbf{z} = \nabla\psi^\star\Big(\nabla\psi(\mathbf{x})-\eta\mathbf{g}\Big)$$

where $\psi^\star$ is the Fenchel dual of ψ.

Let ψ be 1-strongly convex with respect to the norm $\|\cdot\|$. We define

$$D^2 \coloneqq \max\left\{\sup_{\mathbf{x},\mathbf{x}'\in\mathcal{D}} \frac{1}{2}\|\mathbf{x}-\mathbf{x}'\|^2, \sup_{\mathbf{x}\in\mathcal{D}} D(\mathbf{x},\mathbf{x}_c)\right\}$$

where $\mathbf{x}_c \coloneqq \arg\min_{\mathbf{x}\in\mathcal{D}}\psi(\mathbf{x})$ is the prox center. Hence D controls both the diameter (in $\|\cdot\|$) and the Bregman divergence to the prox center.

We frequently use the fact that

$$\langle \mathbf{x}, A\mathbf{y}\rangle \le |A|_{\max} \quad \forall \mathbf{x}\in\Delta_m,\ \mathbf{y}\in\Delta_n$$

where $|A|_{\max}$ is the maximum entry of A in absolute value, and $\Delta_m \coloneqq \{\mathbf{x}\in\mathbb{R}^m \mid \sum_{i=1}^m x_i = 1, x_i \ge 0\}$ is the standard simplex. On a simplex, we will only consider the entropic mirror map:

$$\psi(\mathbf{x}) = \sum_{i=1}^k x_i\log x_i, \quad k = m \text{ or } n$$

which is well-known to be 1-strongly convex in $\|\cdot\|_1$.

We use $\frac{1}{m}1_m$ to denote the uniform distribution on Δ_m.

3. Problem Formulation and Main Result

An (offline) two-player zero-sum game with payoff matrix A refers to the solving the minimax problem:

$$V \coloneqq \min_{\mathbf{y}\in\Delta_n}\max_{\mathbf{x}\in\Delta_m}\langle \mathbf{x}, A\mathbf{y}\rangle\,. \tag{1}$$

The quantity V in (1) is called the **value** of the game, or the Nash Equilibrium Value. Any pair $(\bar{\mathbf{x}},\bar{\mathbf{y}})$ attaining the game value is called an equilibrium strategy.

In the decentralized setting (aka., the "strongly uncoupled" setting), the payoff matrix and the number of opponent's strategies are unknown to both players, and their goal is to learn a pair of equilibrium strategy through repeated game plays. Moreover, each player aims to suffer a low individual regret, even in the presence of an adversary or a corrupted channel that distorts the feedback.

Specifically, at each round t, the players take actions $\mathbf{x}_t$ and $\mathbf{y}_t$, and then receive the loss vectors $-A\mathbf{y}_t$ (for $\mathbf{x}$-player) and $A^\top \mathbf{x}_t$ (for $\mathbf{y}$-player). In the honest setting, we assume that the two players take actions according to a prescribed algorithm, and we say the setting is adversarial if only one player (the $\mathbf{x}$-player in this paper) adheres to the prescribed algorithm and the other player arbitrary.

As in previous work, we assume that an upper bound $|A|_{\max}$ on the maximum absolute entry of A is available to both players. The goal is to achieve

$$|V - \langle \mathbf{x}_T, A\mathbf{y}_T \rangle| \le r_1(T),$$
$$R_T := \max_{\mathbf{x} \in \Delta_m} \sum_{t=1}^{T} \langle \mathbf{x}_t - \mathbf{x}, -A\mathbf{y}_t \rangle \le r_2(T)$$

for fast-decaying r_1 and sublinear r_2 in T. The first requirement is to approximate the game value in (1), and the second one asks to minimize the regret R_T.

Our main result can be stated as follows:

Theorem 1 (Main result, informal)**.** *For (1), there is a simple decentralized algorithm with non-adaptive step-size such that*

$$r_1(T) = O\left(\frac{1}{T}\right), \qquad r_2(T) = O\left(\log T\right),$$

if the opponent is honest (i.e., playing collaboratively to solve the game). Moreover, against any adversary, we have

$$r_2(T) = O\left(\sqrt{T}\right).$$

Except for the $O(\log T)$ honest regret, these rates are known to be optimal (Cesa-Bianchi & Lugosi, 2006; Daskalakis et al., 2015). We are also the first to remove $\log T$ factors in convergence to the value of the game, an open question posed by the very first work in learning decentralized games (Daskalakis et al., 2011).

4. A family of optimistic mirror descents: Classical, Robust, and Let's be honest

We first illustrate the high-level ideas to prove **Theorem 1** in Section 4.1. A novel analysis for OMD in the honest setting is given in Section 4.2, and we propose a new algorithm for the adversarial setting in Section 4.3. Finally, the full algorithm is presented in Section 4.4, along with the rigorous version of the main result (*cf.,* **Theorem 4**).

4.1. High-Level Ideas

Our algorithms are inspired by the iterates of the form:

$$\begin{cases} \mathbf{x}_{t+1} = MD_\eta(\mathbf{x}_t, -2A\mathbf{y}_t + A\mathbf{y}_{t-1}) \\ \mathbf{y}_{t+1} = MD_\eta(\mathbf{y}_t, 2A^\top \mathbf{x}_t - A^\top \mathbf{x}_{t-1}) \end{cases}, \tag{2}$$

which are equivalent to the OMD in (Rakhlin & Sridharan, 2013b) (see Appendix A). It is known that directly applying (2) to (1) yields $O\left(\frac{1}{T}\right)$ convergence in the game value, however without any guarantee on the regret.

To make OMD optimal for zero-sum games, we improve (2) on two fronts. First, in the honest setting, we make the following simple observation: Although the iterates $\mathbf{x}_t$ are not guaranteed to possess sublinear regret, the averaged iterates $\frac{1}{t}\sum_{i=1}^{t} \mathbf{x}_i$ do enjoy logarithmic regret, and hence, it suffices to play the averaged iterates in the honest setting.

Second, in order to make OMD robust against any adversary, we utilize the "mixing steps" of (Rakhlin & Sridharan, 2013b) with an important improvement: Our step-sizes do not depend on the time horizon. This new feature is crucial in removing $\log T$ factors in both the convergence to game value and adversarial regret. In fact, our analysis is arguably simpler than (Rakhlin & Sridharan, 2013b).

4.2. Optimistic Mirror Descent

Algorithm 1 Optimistic Mirror Descent: $\mathbf{x}$-Player

Set $\eta = \frac{1}{2|A|_{\max}}$
Play $\mathbf{z}_1 = \mathbf{z}_2 = \mathbf{z}_3 = \frac{1}{m}1_m$
For $t \ge 3$:

1: Compute

$$\mathbf{x}_{t+1} = MD_\eta(\mathbf{x}_t, -2(t-2)A\mathbf{w}_t + 3(t-3)A\mathbf{w}_{t-1} - (t-4)A\mathbf{w}_{t-2})$$

2: Play $\mathbf{z}_{t+1} = \frac{1}{t-1}\sum_{i=3}^{t+1} \mathbf{x}_i$
3: Observe $-A\mathbf{w}_{t+1}$

Algorithm 2 Optimistic Mirror Descent: $\mathbf{y}$-Player

Set $\eta = \frac{1}{2|A|_{\max}}$
Play $\mathbf{w}_1 = \mathbf{w}_2 = \mathbf{w}_3 = \frac{1}{n}1_n$
For $t \ge 3$:

1: Compute

$$\mathbf{y}_{t+1} = MD_\eta(\mathbf{y}_t, 2(t-2)A^\top \mathbf{z}_t - 3(t-3)A^\top \mathbf{z}_{t-1} + (t-4)A^\top \mathbf{z}_{t-2})$$

2: Play $\mathbf{w}_{t+1} = \frac{1}{t-1}\sum_{i=3}^{t+1} \mathbf{y}_i$
3: Observe $A^\top \mathbf{z}_{t+1}$

As alluded to in Section 4.1, we will play OMD with the averaged iterates. The algorithms are given explicitly in **Algorithm 1** and **2**.

Remark 1. *Note that there is no need to play $\frac{1}{m}1_m$ and $\frac{1}{n}1_n$ three times in **Algorithm 1** and **2**. The players could just play once $\left(\frac{1}{m}1_m\right)^\top A \left(\frac{1}{n}1_n\right)$ and would have enough*

information to run OMD from $\mathbf{x}_4$ and $\mathbf{y}_4$. Our choices are motivated by the resulting ease of the notation.

We analyze our version of OMD below. The crux of our analysis is to first look at the regrets of auxiliary sequences $\mathbf{x}_t$ and $\mathbf{y}_t$, and we show that the *sum* of the auxiliary regrets, not any individual of them, controls both the convergence to the value of the game and the honest regret for the averaged sequences $\mathbf{z}_t$ and $\mathbf{w}_t$.

Theorem 2. *Suppose two players of a zero-sum game have played T rounds according to the OMD algorithm with $\eta = \frac{1}{2|A|_{\max}}$. Then*

1. *The $\mathbf{x}$-player suffers an $O(\log T)$ regret:*

$$\max_{\mathbf{z}\in\Delta_m} \sum_{t=3}^{T} \langle \mathbf{z}_t - \mathbf{z}, -A\mathbf{w}_t \rangle \le \log 2(T-2)|A|_{\max} \times \left(20 + \log m + \log n\right) = O(\log T) \tag{3}$$

   and similarly for the $\mathbf{y}$-player.

2. *The strategies $(\mathbf{z}_T, \mathbf{w}_T)$ constitutes an $O\left(\frac{1}{T}\right)$-approximate equilibrium to the value of the game:*

$$|V - \langle \mathbf{z}_T, A\mathbf{w}_T \rangle| \le \frac{\left(20 + \log m + \log n\right)|A|_{\max}}{T-2} = O\left(\frac{1}{T}\right). \tag{4}$$

Proof. See **Appendix B**. □

4.3. Robust Optimistic Mirror Descent

In this section, we introduce **robust optimistic mirror descent** (ROMD), which is a novel algorithm even for online convex optimization.

Let ψ be 1-strongly convex with respect to $\|\cdot\|$, and suppose we are minimizing the regret against an arbitrary sequence of convex functions $f_1, f_2, \ldots$ in a constraint set $\mathcal{D}$. Assume that each function is G-Lipschitz in $\|\cdot\|$. Assume also that no Bregman projection is needed (i.e., $MD_\eta(\mathbf{x}, \mathbf{g}) \in \mathcal{D}$ for any $\mathbf{x}$ and $\mathbf{g}$); this is, for instance, the case for the entropic mirror map.

We state ROMD in the general form in **Algorithm 3**.

Theorem 3 ($O(\sqrt{T})$-Adversarial Regret)**.** *Suppose that $\|\nabla f_t\|_* \le G$ for all t. Then playing T rounds of **Algorithm 3** with $\eta_t = \frac{1}{G\sqrt{t}}$ against an arbitrary sequence of convex functions has the following guarantee on the regret:*

$$\max_{\mathbf{x}\in\Delta_m} \sum_{t=1}^{T} \langle \mathbf{x}_t - \mathbf{x}, \nabla f_t(\mathbf{x}_t) \rangle \le G\sqrt{T}\left(18 + 2D^2\right) + GD\left(3\sqrt{2} + 4D\right) = O\left(\sqrt{T}\right).$$

Proof. See **Appendix C**. □

Algorithm 3 Robust Optimistic Mirror Descent

Initialize $\mathbf{x}_1 = \mathbf{x}_c$, $\nabla f_0 = 0$, $\eta_t = \frac{1}{G\sqrt{t}}$
for $t = 1, 2, \ldots,$ **do**
  $\tilde{\mathbf{x}}_t = (\frac{t-1}{t})\mathbf{x}_t + \frac{1}{t}\mathbf{x}_c$
  Set $\tilde{\nabla}_t = 2\nabla f_t(\mathbf{x}_t) - \nabla f_{t-1}(\mathbf{x}_{t-1})$, play $\mathbf{x}_{t+1} = MD_{\eta_t}(\tilde{\mathbf{x}}_t, \tilde{\nabla}_t)$
  Observe f_{t+1}
end for

When specialized to zero-sum games, it suffices to take $\mathbf{x}_c = \frac{1}{m}1_m$, $G = |A|_{\max}$, $D = \log m$, and ψ being the entropic mirror map.

Remark 2. *Our analysis of ROMD crucially relies on the assumption that no Bregman projection is needed. We have not been able to generalize our analysis to the case with Bregman projections.*

4.4. Let's be honest: The full framework

We now present our approach for solving (1).

To ease the notation, define

$$\mathbf{z}_t^* := \arg\min_{\mathbf{x}\in\Delta_m} \langle \mathbf{x}, -A\mathbf{w}_t \rangle$$

and

$$\mathbf{w}_t^* = \arg\min_{\mathbf{y}\in\Delta_n} \langle \mathbf{z}_t, A\mathbf{y} \rangle.$$

Let constants C_1, C_2, and C_3 be such that (see **Thereom 2**, **Theorem 3**, and (B.10))

$$\langle \mathbf{z}_t - \mathbf{z}_t^*, -A\mathbf{w}_t \rangle \le \frac{C_1}{t}, \quad \mathbf{z}_t, \mathbf{w}_t \text{ from OMD}, \tag{5}$$

$$\langle \mathbf{w}_t - \mathbf{w}_t^*, A^\top \mathbf{z}_t \rangle \le \frac{C_1}{t}, \quad \mathbf{z}_t, \mathbf{w}_t \text{ from OMD}, \tag{6}$$

$$\sum_{t=1}^{T} \langle \mathbf{z}_t - \mathbf{z}^*, -A\mathbf{y}_t \rangle \le C_2\sqrt{T}, \quad \mathbf{z}_t \text{ from ROMD and } \mathbf{y}_t \text{ arbitrary}, \tag{7}$$

$$|V - \mathbf{z}_T A \mathbf{w}_T| \le \frac{C_3}{T}, \quad \mathbf{z}_T, \mathbf{w}_T \text{ from OMD}. \tag{8}$$

From a high-level, our approach exploits the following simple observation: Suppose that we know C_1 above. If the

Algorithm 4 Let's Be Honest Optimistic Mirror Descent: $\mathbf{x}$-Player

1: Initialize $b = 1, t = 1, \mathbf{w}_0 = \frac{1}{n}1_n$ and $\mathbf{z}_0 = \frac{1}{m}1_m$
2: Play t-th round of OMD-$\mathbf{x}$, observe $-A\mathbf{p}_t$
3:
4: **if** $G_t^{\mathbf{w}} := \langle \mathbf{w}_{t-1}, A^\top \mathbf{z}_{t-1} \rangle - \langle \mathbf{p}_t, A^\top \mathbf{z}_{t-1} \rangle > \frac{b}{t-1}$ **then**
5: Play $b^4 - 1$ rounds of ROMD
6: $t \leftarrow t + 1$
7: $b \leftarrow 2b$
8: Go to line 2.
9: **end if**
10: $-A\mathbf{w}_t \leftarrow -A\mathbf{p}_t$
11:
12: **if** $G_t^{\mathbf{z}} := \langle \mathbf{z}_t, -A\mathbf{w}_t \rangle - \langle \mathbf{z}_t^*, -A\mathbf{w}_t \rangle > \frac{b}{t}$ **then**
13: Play $\check{\mathbf{x}}_{t+1} := \mathbf{z}_t^*$
14: Play $b^4 - 1$ rounds of ROMD
15: $t \leftarrow t + 2$
16: $b \leftarrow 2b$
17: Go to line 2.
18: **end if**
19: $t \leftarrow t + 1$
20: Go to line 2.

Algorithm 5 Let's Be Honest Optimistic Mirror Descent: $\mathbf{y}$-Player

1: Initialize $b = 1, t = 1, \mathbf{w}_0 = \frac{1}{n}1_n$ and $\mathbf{z}_0 = \frac{1}{m}1_m$
2: Play t-th round of OMD-$\mathbf{y}$, observe $A^\top \mathbf{o}_t$
3:
4: **if** $G_t^{\mathbf{z}} := \langle \mathbf{z}_{t-1}, -A\mathbf{w}_{t-1} \rangle - \langle \mathbf{o}_t, -A\mathbf{w}_{t-1} \rangle > \frac{b}{t-1}$ **then**
5: Play $b^4 - 1$ rounds of ROMD
6: $t \leftarrow t + 1$
7: $b \leftarrow 2b$
8: Go to line 2.
9: **end if**
10: $A\mathbf{z}_t \leftarrow A^\top \mathbf{o}_t$
11:
12: **if** $G_t^{\mathbf{w}} := \langle \mathbf{w}_t, A^\top \mathbf{z}_t \rangle - \langle \mathbf{w}_t^*, A^\top \mathbf{z}_t \rangle > \frac{b}{t}$ **then**
13: Play $\check{\mathbf{y}}_{t+1} := \mathbf{w}_t^*$
14: Play $b^4 - 1$ rounds of ROMD
15: $t \leftarrow t + 2$
16: $b \leftarrow 2b$
17: Go to line 2.
18: **end if**
19: $t \leftarrow t + 1$
20: Go to line 2.

instantaneous regret bound (5) and (6) hold true for all t, then we would trivially have the desired convergence.

In contrast, if at any round the bound (5) is violated for the $\mathbf{x}$-player, then it must be due to an adversarial play, and we can simply switch to ROMD to get $O(\sqrt{T})$ regret. However, since C_1 (*cf.*, (B.10)) involves n, the number of opponent's strategies, the $\mathbf{x}$-player cannot compute it exactly. The situation is similar for the $\mathbf{y}$-player. We hence need to come up with a way to estimate C_1 for both players.

It is important to note that one can not naïvely estimate C_1 by binary search separately on both players. The reason, and the major difficultly to the above approach, is as follows: Since in general $\langle \mathbf{z}_t - \mathbf{z}_t^*, -A\mathbf{w}_t \rangle \neq \langle \mathbf{w}_t - \mathbf{w}_t^*, A^\top \mathbf{z}_t \rangle$, it could be the case that, at the same round, the $\mathbf{x}$-player detects a bad instantaneous regret and switch to ROMD, while the $\mathbf{y}$-player remains in OMD, even though two players are both honest. However, our entire analysis of OMD would breakdown if the OMD is not played cohesively.

Furthermore, recall that we also want robustness against any adversary. Therefore, a bad instantaneous regret indicates the possibility of receiving an adversarial play, and we need to switch to ROMD whenever this occurs.

To resolve such issues, we devise a simple **signaling** scheme ($\check{\mathbf{x}}_t$ and $\check{\mathbf{y}}_t$ below), which synchronizes both players' C_1 estimate and also the OMD plays while guaranteeing robustness.

In words, our signaling scheme is a "Let's be honest" message to the opponent: "I am having a bad instantaneous regret. Please update your C_1 with me, and please pretend that I am adversarial for a small number of rounds, so that we can play honest OMD cohesively." It turns out that doing these extra signaling rounds do not hurt the convergence rates in OMD and ROMD at all.

Our full algorithm, termed **Let's Be Honest** (LbH) **Optimistic Mirror Descent**, is presented in **Algorithm 4** and **5**.

Remark 3. *In **Algorithm 4** and **5**, the role of b is to estimate the constant C_1 in (5). Since our analysis requires b to be the same for both players throughout the algorithm run, a simple way is to assume that, say, $m = n = 5$, compute the corresponding $\tilde{C}_1$, and set the initial $b \leftarrow \tilde{C}_1$. Doing so indeed improves upon constants in our convergence; we chose $b = 1$ only for simplicity.*

Remark 4. *There are some degree of freedom in **Algorithm 4** and **5**. For instance, instead of doubling b in Line 16, one can do $b \leftarrow (1 + \epsilon)b$ for some $\epsilon > 0$. In Line 5, one can also play $b^2 - 1$ rounds, rather than $b^4 - 1$. As will become apparent in **Theorem 4**, these variants only effect the constants but not the convergence rates. However, they do have impact on empirical performance; cf., Section 5.*

The following key lemma ensures the two players to enter the ROMD plays coherently.

Lemma 1. *If the $\mathbf{y}$-player enters Line 12 of **Algorithm 5** at*

*the t-th round, then the $\mathbf{x}$-player enters Line 4 of **Algorithm 4** at the $(t+1)$-th round. Conversely, if, at the t-th round, the $\mathbf{y}$-player does not enter Line 12 of **Algorithm 5**, then the $\mathbf{x}$-player does not enter Line 4 of **Algorithm 4** at the $(t+1)$-th round.*

Exactly the same statements hold when the $\mathbf{x}$- and $\mathbf{y}$-player are reversed above.

Proof. If the $\mathbf{y}$-player enters Line 12 of **Algorithm 5** at the t-th round, then $\check{\mathbf{y}}_{t+1}$ is signalled at the $(t+1)$-th round, and it must be the case that $\langle \mathbf{w}_t - \mathbf{w}_t^*, A^\top \mathbf{z}_t \rangle > \frac{b}{t}$ (*cf.*, Line 12 of **Algorithm 5**). Therefore, at the $(t+1)$-th round, the $\mathbf{x}$-player would receive $-A\check{\mathbf{y}}_{t+1} = -A\mathbf{w}_t^*$ and compute

$$\begin{aligned} G_{t+1}^{\mathbf{w}} &= \langle \mathbf{w}_t, A^\top \mathbf{z}_t \rangle - \langle \check{\mathbf{y}}_{t+1}, A^\top \mathbf{z}_t \rangle \\ &= \langle \mathbf{w}_t - \mathbf{w}_t^*, A^\top \mathbf{z}_t \rangle > \frac{b}{t} \end{aligned}$$

which then enters the Line 4 of **Algorithm 4**.

Conversely, suppose that the $\mathbf{y}$-player does not enter Line 12 of **Algorithm 5** at the t-th round (or, equivalently, plays OMD at the $(t+1)$-th round). Then $\langle \mathbf{w}_t - \mathbf{w}_t^*, A^\top \mathbf{z}_t \rangle \le \frac{b}{t}$, implying that

$$\begin{aligned} G_{t+1}^{\mathbf{w}} &= \langle \mathbf{w}_t - \mathbf{w}_{t+1}, A^\top \mathbf{z}_t \rangle \\ &\le \langle \mathbf{w}_t - \mathbf{w}_t^*, A^\top \mathbf{z}_t \rangle \le \frac{b}{t} \end{aligned}$$

hence preventing the $\mathbf{x}$-player from entering Line 4 of **Algorithm 4**.

Exactly the same computation holds when we reverse the role of $\mathbf{x}$- and $\mathbf{y}$-player. □

Given **Lemma 1**, we now know that the $\mathbf{x}$-player switches to ROMD **if and only if** the $\mathbf{y}$-player does. The rest of the proof then readily follows from **Theorems 2** and **3**.

Theorem 4. *Suppose the $\mathbf{x}$-player plays according to **Algorithm 4** for T rounds, and let R_T be the regret up to time T. Then*

1. *Let $T = T_1 + T_2 + T_3$ where T_1 is the number of OMD plays, T_2 is the number of ROMD plays, and T_3 is the number of signaling rounds (playing $\check{\mathbf{x}}_t$ or $\check{\mathbf{y}}_t$). Then there are constants C and C', depending only on m, n and $|A|_{\max}$, such that*

$$\frac{1}{T}\mathrm{R}_T \le \frac{C \log T_1 + C'\sqrt{T_2}}{T_1 + T_2}. \tag{9}$$

   In particular, if the opponent plays honestly, then $\mathrm{R}_T = O(\log T_1) = O(\log T)$. If the opponent is adversarial, we have $\mathrm{R}_T = O(\sqrt{T_2}) = O(\sqrt{T})$.

2. *Suppose that the honest $\mathbf{y}$-player plays **Algorithm 5**. Then the pair $(\mathbf{z}_T, \mathbf{w}_T)$ constitutes an $O\left(\frac{1}{T}\right)$-approximate equilibrium:*

$$|V - \langle \mathbf{z}_T, A\mathbf{w}_T \rangle| \le \frac{C''}{T} \tag{10}$$

   for some constant C''.

Proof. Suppose first that both players are honest.

We first prove the individual regret for the $\mathbf{x}$-player. We split the terms as follows:

$$\begin{aligned} \mathrm{R}_T = \mathrm{R}_{T_1}(\text{playing OMD}) + \mathrm{R}_{T_2}(\text{playing ROMD}) \\ + \mathrm{R}_{T_3}(\text{signaling}). \end{aligned} \tag{11}$$

Recall (5)-(8). We claim that

(a) $T_3 \le \lceil \log C_1 \rceil$.

(b) $T_2 \le 16 \cdot \frac{16^{T_3-1}-1}{15} \coloneqq C_1'$.

Indeed, after $\lceil \log C_1 \rceil$-times signaling, we would have $b = 2^{T_3} > C_1$. Then (5) and (6) imply that we will never enter Line 12 again. On the other hand, we have

$$T_2 \le \sum_{r=1}^{T_3} 2^{4r} = \frac{16^{T_3-1}-1}{15}.$$

Combining (a), (b) and using (5), (7) in (11), we conclude that

$$\begin{aligned} \mathrm{R}_T &\le C_1 \log T_1 + C_2\sqrt{T_2} + 2|A|_{\max} T_3 \\ &\le C_1 \log T_1 + C_2\sqrt{C_1'} + 2|A|_{\max}\lceil \log C_1 \rceil \\ &= O(\log T_1) = O(\log T) \end{aligned}$$

which establishes (9) in the honest case.

For convergence to the value of the game, we have, by (8),

$$|V - \langle \mathbf{z}_T, A\mathbf{w}_T \rangle| \le \frac{C_3}{T - T_2 - T_3} \le \frac{C_3}{T - C^*}$$

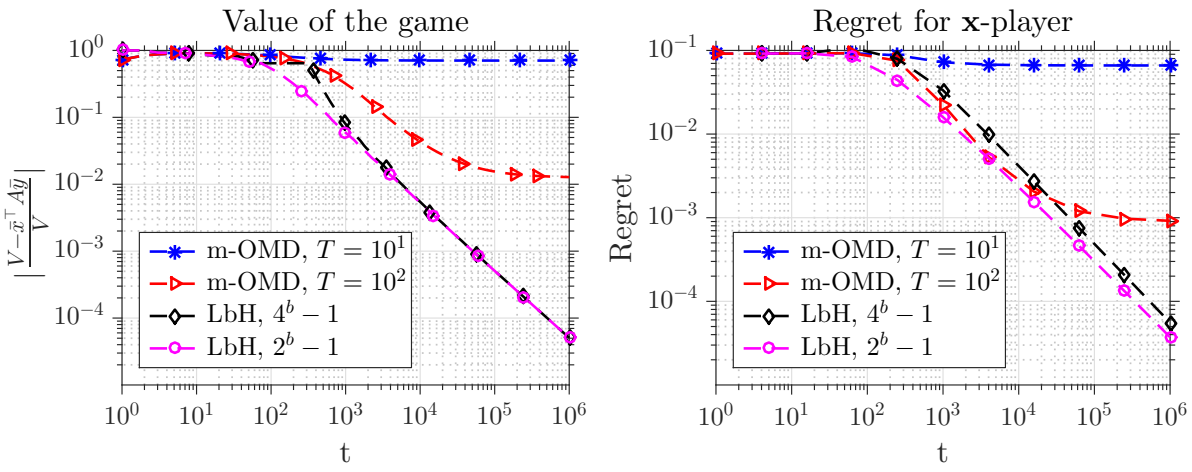
where $C^* = \lceil \log C_1 \rceil + C_1'$. The proof of (10) is completed by using the fact that $\frac{1}{T-C^*} \le \frac{C^*}{T}$ when $T \ge \frac{C^{*2}}{C^*-1}$.

Finally, we show (9) in the adversarial case.

Let T_1, T_2, and T_3 be as before, and we again split the regret into:

$$\begin{aligned} \mathrm{R}_T = \mathrm{R}_{T_1}(\text{playing OMD}) + \mathrm{R}_{T_2}(\text{playing ROMD}) \\ + \mathrm{R}_{T_3}(\text{signaling}). \end{aligned}$$

Notice that this time the inequalities (5) and (6) do not apply since the opponent no longer plays OMD collaboratively.



(a) Value of the game.

(b) Regret.

Figure 1. Honest setting.

However, by Line 12 of **Algorithm 4**, for every OMD play we must have

$$\langle \mathbf{z}_t, -A\mathbf{w}_t\rangle - \langle \mathbf{z}_t^*, -A\mathbf{w}_t\rangle \le \frac{b}{t} \le \frac{2^{T_3}}{t}.$$

Following the analysis as in the honest setting, we may further write

$$\mathrm{R}_T \le 2^{T_3}\log T_1 + C_2\sqrt{T_2} + 2|A|_{\max}T_3.$$

It hence suffices to show that

$$2^{T_3}\log T_1 \le C^{**}\sqrt{T_1+T_2}. \tag{12}$$

for some constant C^{**}. To see (12), recall that

$$T_2 = \frac{16(16^{T_3}-1)}{15} \ge 16^{T_3-1}.$$

But then

$$\frac{2^{T_3}\log T_1}{\sqrt{T_1+T_2}} \le \frac{2^{T_3}\log T_1}{\sqrt{2\sqrt{T_1T_2}}} \le \frac{2^{T_3}\log T_1}{2^{T_3-1}\cdot\sqrt{2}\cdot\sqrt[4]{T_1}} \le C^{**}.$$

for some universal constant C^{**}.

□

Remark 5. *As is evident from the proof, we have made no attempt to sharpening the constants, and hence our bounds can be numerically loose.*

5. Experiments

The purpose of this section is to provide numerical evidence to the following claims of our theory:

1. The LbH algorithm does not require knowing the time horizon beforehand, and our step-sizes are non-adaptive. Therefore, all quantities of interest, such as regrets or game value, should steadily decrease along the algorithm run.

2. The LbH algorithm automatically adjusts to honest and adversarial opponents.

For comparison, we include the modified OMD (henceforth abbreviated as m-OMD) of (Rakhlin & Sridharan, 2013b) in our experiment, for different choices of time horizon.

We generate the entries of A uniformly at random in the interval $[-1, 1]$, and we set $m = 200$ and $n = 300$.

We consider two scenarios:

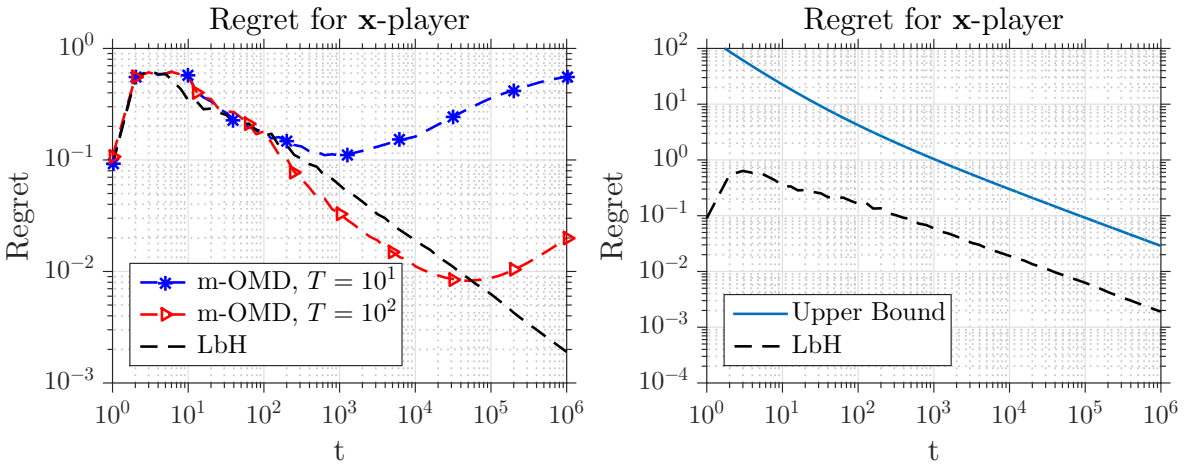
1. *Honest setting*: Both players adhere to the prescribed algorithms and try to reach the Nash equilibrium collaboratively.

2. *Adversarial setting*: The $\mathbf{y}$-player greedily maximizes the instantaneous regret of the $\mathbf{x}$-player.

5.1. Honest Setting

The convergence for the honest setting is reported in **Figure 1**, for two different parameter choices of LbH and m-OMD.

For both convergence to the game value and individual regret, after a short burn-in period (due to not knowing the C_1 in (5) and (6)), the LbH algorithm enters a steady $O\left(\frac{1}{T}\right)$-decreasing phase, as expected from our theory. On the other hand, as the m-OMD chooses step-sizes according to the time horizon, it eventually saturates in both plots.



(a) Regret comparison.

(b) Upper bound.

Figure 2. Adversarial setting.

As noted by (Rakhlin & Sridharan, 2013b), it is possible to prevent the saturation of m-OMD by employing the doubling trick or the techniques in (Auer et al., 2002). However, doing so not only complicates the algorithm, but also introduces extra $\log T$ factors in the convergence of honest regret, since the doubling trick loses a $\log T$ factor for logarithmic regrets. Such rates are sub-optimal given our results.

5.2. Adversarial Setting

We report the regret comparison in **Figure 2**.

In the adversarial setting, the LbH algorithm is essentially running the ROMD, and hence we see a straight $O(T^{-\frac{1}{2}})$ decrease in the regret, as dictated by our upper bound in **Theorem 3**; see **Figure 2**-(b). The parameter choice does not effect the performance.

The m-OMD slightly outperforms LbH for a short period, but eventually blows up in regret. We remark that the short-term good empirical performance is due to the adaptive step-sizes of m-OMD, which require additional work per-iteration. Our LbH algorithm is non-adaptive, but is already competitive in terms of empirical performance.

6. Conclusion and Future Work

We studied the problem of zero-sum games in the decentralized setting, and we resolved an open problem of achieving optimal convergence to the game value while maintaining low regrets. Our techniques were based on several simple but novel observations in the game dynamics. Namely, we noticed that the averaged iterates of OMD enjoy logarithmic regret in the honest setting, we provided horizon-independent mixing steps for the OMD to achieve optimal adversarial regret, and we designed a singaling scheme to losslessly bridge OMD and ROMD. In essence, we showed that it is not necessary, as done in the work of (Rakhlin & Sridharan, 2013b), to fix the time horizon beforehand and modify OMD accordingly. Our observations were instrumental in removing $\log T$ terms in all convergence rates.

Our framework suggests several research directions. First, instead of assuming that we observe the full loss vector, we may pose our problem in the *bandit* setting, where only the payoff value of the current strategy is observed. Second, for practical purposes, it is interesting to see whether there exists an *adaptive* step-size version of our algorithm. Finally, generalizing our framework to *multiplayer* games is a challenging future work.

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 n^o 725594 - time-data), and was supported by the Swiss National Science Foundation (SNSF) under grant number 200021_178865 / 1.

References

Auer, Peter, Cesa-Bianchi, Nicolo, and Gentile, Claudio. Adaptive and self-confident on-line learning algorithms. *Journal of Computer and System Sciences*, 64(1):48–75, 2002.

Cesa-Bianchi, Nicolo and Lugosi, Gábor. *Prediction, learning, and games*. Cambridge university press, 2006.

Chiang, Chao-Kai, Yang, Tianbao, Lee, Chia-Jung, Mah-

davi, Mehrdad, Lu, Chi-Jen, Jin, Rong, and Zhu, Shenghuo. Online optimization with gradual variations. In *Conference on Learning Theory*, pp. 6–1, 2012.

Daskalakis, Constantinos, Deckelbaum, Alan, and Kim, Anthony. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete algorithms*, pp. 235–254. Society for Industrial and Applied Mathematics, 2011.

Daskalakis, Constantinos, Deckelbaum, Alan, and Kim, Anthony. Near-optimal no-regret algorithms for zero-sum games. *Games and Economic Behavior*, 92:327–348, 2015.

Foster, Dylan J, Li, Zhiyuan, Lykouris, Thodoris, Sridharan, Karthik, and Tardos, Eva. Learning in games: Robustness of fast convergence. In *Advances in Neural Information Processing Systems*, pp. 4734–4742, 2016.

Myerson, Roger B. Nash equilibrium and the history of economic theory. *Journal of Economic Literature*, 37(3): 1067–1082, 1999.

Nesterov, Yu. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, 16(1):235–249, 2005.

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

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

Syrgkanis, Vasilis, Agarwal, Alekh, Luo, Haipeng, and Schapire, Robert E. Fast convergence of regularized learning in games. In *Advances in Neural Information Processing Systems*, pp. 2989–2997, 2015.

Von Neumann, John. On the theory of parlor games. *Mathematische Annalen*, 100:295–320, 1928.