A Universal Algorithm for Variational Inequalities Adaptive to Smoothness and Noise

Francis Bach   FRANCIS.BACH@INRIA.FR
INRIA, ENS, PSL Research University Paris, France.

Kfir Y. Levy   YEHUDA.LEVY@INF.ETHZ.CH
Department of Computer Science, ETH Zurich.

Abstract

We consider variational inequalities coming from monotone operators, a setting that includes convex minimization and convex-concave saddle-point problems. We assume an access to potentially noisy unbiased values of the monotone operators and assess convergence through a compatible gap function which corresponds to the standard optimality criteria in the aforementioned subcases. We present a universal algorithm for these inequalities based on the Mirror-Prox algorithm. Concretely, our algorithm *simultaneously* achieves the optimal rates for the smooth/non-smooth, and noisy/noiseless settings. This is done without any prior knowledge of these properties, and in the general set-up of arbitrary norms and compatible Bregman divergences. For convex minimization and convex-concave saddle-point problems, this leads to new adaptive algorithms. Our method relies on a novel yet simple adaptive choice of the step-size, which can be seen as the appropriate extension of AdaGrad to handle constrained problems.

Keywords: Online learning, convex optimization, first-order methods, universal methods, minimax games.

1. Introduction

Variational inequalities are a classical and general framework to encompass a wide variety of optimization problems such as convex minimization and convex-concave saddle-point problems, which are ubiquitous in machine learning and optimization (Nemirovski, 2004; Juditsky et al., 2011; Juditsky and Nemirovski, 2016). Given a convex subset $\mathcal{K}$ of $\mathbb{R}^d$, these inequalities are often defined from a *monotone operator* $F : \mathcal{K} \mapsto \mathbb{R}^d$ (which we will assume single-valued for simplicity), such that for any $(x, y) \in \mathcal{K} \times \mathcal{K}, (x - y) \cdot (F(x) - F(y)) \geqslant 0$. The goal is then to find a strong solution $x^* \in \mathcal{K}$ to the variational inequality, that is, such that

$$\forall x \in \mathcal{K}, \ (x^* - x) \cdot F(x^*) \leq 0. \tag{1}$$

For convex minimization problems, the operator F is simply the subgradient operator, while for convex-concave saddle-point problems, the operator F is composed of the subgradient with respect to the primal variable, and the negative supergradient with respect to the dual variables (see a detailed description in Section 5.2). In these two classical cases, solving the variational inequality corresponds to the usual notion of solution for these two problems. While our main motivation is to have a unique framework for these two subcases, the variational inequality framework is more general (see e.g. Nemirovski (2004) and references therein).

In this paper we are interested in algorithms to solve the inequality in Eq. (1), while only accessing an oracle for $F(x)$ for any given $x \in \mathcal{K}$, or only an unbiased estimate of $F(x)$. We also assume that we may efficiently project onto the set $\mathcal{K}$ (which we assume compact throughout this paper) using Bregman divergences. In terms of complexity bounds, this problem is by now well-understood with matching upper and lower bounds in a variety of situations. In particular the notion of smoothness (i.e., Lipschitz-continuity of F vs. simply assuming that F is bounded) and the presence of noise are the two important factors influencing the convergence rates. For example, the "Mirror-Prox" algorithm of Nemirovski (2004) and Juditsky et al. (2011), given the correct step-size (that depends heavily on the properties of the problem, see Section 2), attains the following bounds:

- For non-smooth problems where the operator (and its unbiased estimates) is bounded by G, the rate $O(GD/\sqrt{T})$ is attained after T iterations, where D is the proper notion of diameter for the set $\mathcal{K}$.

- For smooth problems with L-Lipschitz operators, and a noise variance of σ^2, the convergence rate is $O(LD^2/T + \sigma D/\sqrt{T})$.

These rates are actually optimal for this class of problems[1]. However, practitioners may not know in which class their problem lies or know all the required constants needed for running the algorithms. Thus universal (sometimes called adaptive) algorithms are needed to leverage the potentially unknown properties of an optimization problem. Moreover, locally, the problem could be smoother or less noisy than globally, and thus classical algorithms would not benefit from extra local speed-ups.

In this paper we make the following contributions:

- We present a universal algorithm for variational inequalities based on the Mirror-Prox algorithm, for both deterministic and stochastic settings. Our method employs a simple adaptive choice of the step-size that leads to optimal rates for smooth and non-smooth variational inequalities. Our algorithm does not require prior knowledge regarding the smoothness or noise properties of the problem.

- This is done in the general set-up of arbitrary norms and compatible Bregman divergences.

- For convex minimization and convex-concave saddle-point problems, this leads to new adaptive algorithms. In particular, our new adaptive method can be seen as extension of Ada-Grad (McMahan and Streeter, 2010; Duchi et al., 2011), that is more appropriate to handling constrained problems.

On the technical side, our work combines the Mirror-Prox method with a novel adaptive learning rate rule inspired by online learning techniques such as AdaGrad (McMahan and Streeter, 2010; Duchi et al., 2011), and optimistic OGD (Chiang et al., 2012; Rakhlin and Sridharan, 2013).

Related work. Algorithms for solving variational inequalities date back to Korpelevich (1976) who was the first to suggest the *extragradient* method. The key idea behind this method is the following: in each round t we make two updates. First, we take a gradient step from the current iterate y_t, which leads to a point $y_{t+1/2}$. Then, instead of applying another gradient step starting in $y_{t+1/2}$, we go back to y_t and take a step using the gradient of $y_{t+1/2}$, which leads to y_{t+1}.

---

1. The class of problems indeed includes convex optimization with lower bounds in $O(1/\sqrt{T})$ (Nemirovskii and Yudin, 1983) and bilinear saddle-point problems with lower bound in $O(1/T)$ (Nemirovsky, 1992).

The work of Korpelevich (1976) was followed by Korpelevich (1983); Noor (2003), who further explored the asymptotic behaviour of such extragradient-like algorithms. The seminal work of Nemirovski (2004) was the first to establish non-asymptotic convergence guarantees of such a method, establishing a rate of $O(LD^2/T)$ for smooth problems. Nemirovski's method named *Mirror-Prox* was further explored by Juditsky et al. (2011), who analyze the stochastic setting, and present a Mirror-Prox version that obtains a rate of $O(LD^2/T + \sigma D/\sqrt{T})$, where σ^2 is the variance of the noise terms. It is also known that in the non-smooth case, Mirror-Prox obtains a rate of $O(GD/\sqrt{T})$ (Juditsky and Nemirovski, 2011). Note that the Mirror-Prox versions that we have mentioned so far require prior knowledge about the smoothness/non-smoothness and on the noise properties of the problem (i.e., σ), in order to obtain the optimal bounds for each case[2]. Conversely, our method obtains these optimal rates without any such prior knowledge. Note that Yurtsever et al. (2015); Dvurechensky et al. (2018) devise universal methods to solve variational inequalities that adapt to the smoothness of the problem. Nevertheless, these methods build on a line search technique that is inappropriate for handling noisy problems. Moreover, these methods require a predefined accuracy parameter as an input, which requires careful hyperparameter tuning.

In the past years there have been several works on universal methods for convex optimization (which is a particular case of the variational inequalities framework). Nesterov (2015) designed a universal method that obtains the optimal convergence rates of $O(LD^2/T^2)$ and $O(GD/\sqrt{T})$ for smooth/non-smooth optimization, without any prior knowledge of the smoothness. Yet, this method builds on a line search technique that is inappropriate to handling noisy problems. Moreover, it also requires a predefined accuracy parameter as an input, which requires careful tuning.

Levy (2017) designed alternative universal methods for convex minimization that do not require line search, yet these methods obtain a rate of $O(1/T)$ rather than the accelerated $O(1/T^2)$ rate for smooth objectives. Moreover, their results for the smooth case only holds for *unconstrained problems*. The same also applies to the well known AdaGrad method (McMahan and Streeter, 2010; Duchi et al., 2011). Recently, Levy et al. (2018) have presented a universal method that obtains the optimal rates for smooth/non-smooth and noisy/noiseless settings, without any prior knowledge of these properties. Nevertheless, their results for the smooth case are only valid in the *unconstrained setting*. Finally, note that these convex optimization methods are usually not directly applicable to the more general variational inequality framework.

Methods for solving convex-concave zero-sum games or saddle-point problems (another particular case of the variational inequality framework) were explored by the online learning community. The seminal work of Freund et al. (1999) has shown how to employ regret minimization algorithms to solve such games at a rate of $O(1/\sqrt{T})$. While the Mirror-Prox method solves such games at a faster rate of $O(1/T)$, it requires communication between the players. Interestingly, Daskalakis et al. (2011) have shown how to achieve a rate of $O(1/T)$ without communication. Finally, Rakhlin and Sridharan (2013) have provided a much simpler algorithm that obtains the same guarantees.

2. Variational Inequalities and Gap Functions

Here we present our general framework of variational inequalities with monotone operators, and introduce the notion of associated convex gap function. In Section 5.1 and 5.2, we show how this framework captures the settings of convex optimization, as well as convex-concave minimax games.

---

2. For the special case of bi-linear saddle-point problems, Juditsky et al. (2013) designed an algorithm that is adaptive to noise, but not to non-smoothness (which is irrelevant for bi-linear problems).

Preliminaries. Let $\| \cdot \|$ be a general norm and $\| \cdot \|_*$ be its dual norm. A function $f : \mathcal{K} \mapsto \mathbb{R}$ is *μ-strongly convex* over a convex set $\mathcal{K}$, if for any $x \in \mathcal{K}$ and any $\nabla f(x)$, a subgradient of f at x,

$$f(y) \geq f(x) + \nabla f(x) \cdot (y - x) + \frac{\mu}{2}\|x - y\|^2 ; \quad \forall x, y \in \mathcal{K}.$$

A function $f : \mathcal{K} \mapsto \mathbb{R}$ is *L-smooth* over $\mathcal{K}$ if, $\|\nabla f(x) - \nabla f(y)\|_* \leq L\|x - y\| ; \quad \forall x, y \in \mathcal{K}$. Also, for a convex differentiable function $f(\cdot)$, we define its *Bregman divergence* as follows,

$$\mathcal{D}_f(x, y) = f(x) - f(y) - \nabla f(y) \cdot (x - y) .$$

Note that $\mathcal{D}_f(\cdot, \cdot)$ is always non-negative. For more properties, see, e.g., Nemirovskii and Yudin (1983) and references therein.

2.1. Gap functions

We are considering a monotone operator F from $\mathcal{K}$ to $\mathbb{R}^d$, which is single-valued for simplicity[3]. Formally, a *monotone operator* satisfies,

$$(x - y) \cdot (F(x) - F(y)) \geq 0; \qquad \forall(x, y) \in \mathcal{K} \times \mathcal{K} .$$

And we are usually looking for a *strong* solution $x^* \in \mathcal{K}$ of the variational inequality, that satisfies

$$\sup_{x \in \mathcal{K}} (x^* - x) \cdot F(x^*) \leq 0.$$

When F is monotone, as discussed by Juditsky and Nemirovski (2016), a strong solution is also a *weak* solution, that is, $\sup_{x \in \mathcal{K}} (x^* - x) \cdot F(x) \leq 0$. Note that we do not use directly the monotonicity property of F; we only use the existence of a compatible gap function with respect to F, which is an adapted notion of merit function to characterize convergence, that we define in Def. 1. We show below that this definition captures the settings of convex optimization and convex-concave games.

We thus assume that we are given a convex set $\mathcal{K}$, as well as a *gap function* $\Delta : \mathcal{K} \times \mathcal{K} \mapsto \mathbb{R}$. For a given solution $x \in \mathcal{K}$, we define its *duality gap* as follows,

$$\mathrm{DualGap}(x) := \max_{y \in \mathcal{K}} \Delta(x, y) . \tag{2}$$

We assume to have an access to an oracle for F, i.e., upon querying this oracle with $x \in \mathcal{K}$, we receive $F(x)$. Our goal is to find a solution such that its duality gap is (approximately) zero. We also consider a stochastic setting (similarly to Juditsky et al. (2011)), where our goal is to provide guarantees on the *expected* duality gap. Next we present the *central definition* of this paper:

Definition 1 (Compatible gap function) *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex set, and let $\Delta : \mathcal{K} \times \mathcal{K} :\mapsto \mathbb{R}$, such that Δ is convex with respect to its first argument. We say that the function Δ is a gap function compatible with the monotone operator $F : \mathcal{K} \mapsto \mathbb{R}^d$ if,*

$$\Delta(x, y) \leq F(x) \cdot (x - y), \qquad \forall x, y \in \mathcal{K} ,$$

and $x^* \in \mathcal{K}$ is a solution of Eq. (1) if and only if $\mathrm{DualGap}(x^*) := \max_{y \in \mathcal{K}} \Delta(x^*, y) = 0$.

Note that given the notion of solution to the variational inequality in Eq. (1), the function $(x, y) \mapsto F(x) \cdot (x - y)$ is a good candidate for Δ, but it is not convex in x in general and thus Jensen's inequality cannot be applied.

---

3. That is, each $x \in \mathcal{K}$ is mapped to a *single* $F(x) \in \mathbb{R}^d$; we could easily extend to the multi-valued setting (Bauschke and Combettes, 2011), at the expense of more cumbersome notations.

Assumptions on F.   Throughout this paper we will assume there exists a bound G on the magnitude of F (and all of its unbiased estimates), i.e.,

$$\|F(x)\|_* \leq G, \qquad \forall x \in \mathcal{K} \,.$$

We will sometimes consider the extra assumption that F is L-smooth w.r.t. a given norm $\|\cdot\|$, i.e.,

$$\|F(x) - F(y)\|_* \leq L\|x - y\|, \qquad \forall x, y \in \mathcal{K} \,,$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. Note that we define the notion of smoothness for functions $f : \mathcal{K} \mapsto \mathbb{R}$, as well as to monotone operators $F : \mathcal{K} \mapsto \mathbb{R}^d$. These two different notions coincide when F is the gradient of f (see Sec. 5.1).

3. Universal Mirror-Prox

This section presents our variational inequality algorithm. We first introduce the optimistic-OGD algorithm of Rakhlin and Sridharan (2013), and present its guarantees. Then we show how to adapt this algorithm together with a novel learning rate rule in order to solve variational inequalities in a universal manner. Concretely, we present an algorithm that, without any prior knowledge regarding the problem's smoothness, obtains a rate of $O(1/T)$ for smooth problems (Thm. 4), and an $O(\sqrt{\log T/T})$ rate for non-smooth problems (Thm. 5). Our algorithm can be seen as an adaptive version of the Mirror-Prox method (Nemirovski, 2004).

We provide a proof sketch of Thm. 4 in Section 3.3. The full proofs are deferred to the Appendix.

3.1. Optimistic OGD

Here we introduce the optimistic online gradient descent (OGD) algorithm of Rakhlin and Sridharan (2013). This algorithm applies to the online linear optimization setting that can be described as a sequential game over T rounds between a learner and an adversary. In each round $t \in [T]$,

- the learner picks a decision point $x_t \in \mathcal{K}$,

- the adversary picks a loss vector $g_t \in \mathbb{R}^d$,

- the learner incurs a loss of $g_t \cdot x_t$, and gets to view g_t as a feedback.

The performance measure for the learner is the regret which is defined as follows,

$$\text{Regret} := \sum_{t=1}^{T} g_t \cdot x_t - \min_{x \in \mathcal{K}} \sum_{t=1}^{T} g_t \cdot x \,,$$

and we are usually interested in learning algorithms that ensure a regret which is sublinear in T.

Hint Vectors.   Rakhlin and Sridharan (2013) assume that in addition to viewing the loss sequence $\{g_t :\in \mathbb{R}^d\}_t$, the learner may access a sequence of "hint vectors" $\{M_t \in \mathbb{R}^d\}_t$. Meaning that in each round t, prior to choosing x_t, the player gets to view a "hint vector" $M_t \in \mathbb{R}^d$. In the case where the hints are good predictions for the loss vectors, i.e., $M_t \approx g_t$, Rakhlin and Sridharan (2013) show

5

that this could be exploited to provide improved regret guarantees. Concretely, they suggest to use the following *optimistic OGD* method: Choose $y_0 = \arg\min_{x \in \mathcal{K}} \mathcal{R}(x)$, and $\forall t \geq 1$,

$$x_t \leftarrow \arg\min_{x \in \mathcal{K}} M_t \cdot x + \frac{1}{\eta_t} \mathcal{D}_{\mathcal{R}}(x, y_{t-1}), \quad \text{and} \quad y_t \leftarrow \arg\min_{x \in \mathcal{K}} g_t \cdot x + \frac{1}{\eta_t} \mathcal{D}_{\mathcal{R}}(x, y_{t-1}), \quad (3)$$

where $\mathcal{R}(\cdot)$ is a 1-strongly-convex function over $\mathcal{K}$ w.r.t. a given norm $\| \cdot \|$, and $\mathcal{D}_{\mathcal{R}}$ is the Bregman divergence of $\mathcal{R}$. The following guarantees for *optimistic OGD* hold, assuming that the learning rate sequence is non-increasing (see proof in Appendix C.1) ,

Lemma 2 (Rakhlin and Sridharan (2013))

$$Regret \leq \frac{D^2}{\eta_1} + \frac{D^2}{\eta_T} + \sum_{t=1}^{T} \|g_t - M_t\|_* \cdot \|x_t - y_t\| - \frac{1}{2} \sum_{t=1}^{T} \eta_t^{-1} \left(\|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2 \right), \quad (4)$$

where $D^2 = \max_{x \in \mathcal{K}} \mathcal{R}(x) - \min_{x \in \mathcal{K}} \mathcal{R}(x)$, and $\| \cdot \|_*$ is the dual norm of $\| \cdot \|$.

3.2. Universal Mirror-Prox

Here we describe a new adaptive scheme for the learning rate of the above mentioned optimistic OGD. Then we show that applying this adaptive scheme to solving variational inequalities yields an algorithm that adapts to smoothness and noise.

A new Adaptive Scheme.   Rakhlin and Sridharan (2013) suggest to apply the following learning rate scheme inside optimistic OGD (Equation (3)),

$$\eta_t = D / \max \left\{ \sqrt{\sum_{t=1}^{t-1} \|g_t - M_t\|^2} + \sqrt{\sum_{t=1}^{t-2} \|g_t - M_t\|^2}, 1 \right\}.$$

They show by employing this rule with a version of optimistic OGD yields an algorithm that solves zero-sum matrix games at a fast $O(1/T)$ rate, without any communication between the players. While the Mirror-Prox algorithm (Nemirovski, 2004) achieves such a fast rate, it requires both players to communicate their iterates to each other in every round.

Our goal here is different. We would like to adapt to the smoothness and noise of the objective, while allowing players to communicate. To do so, we suggest to use the following adaptive scheme,

$$\eta_t = D / \sqrt{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2}, \quad \text{where} \quad Z_\tau^2 := \frac{\|x_\tau - y_\tau\|^2 + \|x_\tau - y_{\tau-1}\|^2}{5\eta_\tau^2}, \quad (5)$$

with the same definition of the diameter D as in Lemma 2, and $G_0 > 0$ is an arbitrary constant. Note that the best choice for G_0 is a tight upper bound on the dual norms of the g_t's and M_t's, which we denote here by G, i.e., $G := \max_{t \in [T]} \max\{\|g_t\|_*, \|M_t\|_*\}$. Nevertheless, even if $G_0 \neq G$ we still achieve convergence guarantees that scales with

$$\alpha := \max\{G/G_0, G_0/G\} . \quad (6)$$

In this work we assume to know D, yet we do not assume any prior knowledge of G.

Finally, note that $Z_\tau \in [0, G]$ for any $\tau \geq 1$; this immediately follows by the next lemma.

Lemma 3 *Let G be a bound on the dual norms of $\{g_t\}_t, \{M_t\}_t$. Then the above holds for y_{t-1}, x_t, y_t, that are used in Optimistic OGD (Eq. (3)),*

$$\|x_t - y_{t-1}\|/\eta_t \leq G, \quad \text{and} \quad \|y_t - y_{t-1}\|/\eta_t \leq G.$$

Algorithm 1 Universal Mirror-Prox

Input: #Iterations T, $y_0 = \arg\min_{x \in \mathcal{K}} \mathcal{R}(x)$, learning rate $\{\eta_t\}_t$ as in Eq. (5)
for $t = 1 \dots T$ **do**
    Set $M_t = F(y_{t-1})$
    Update:
$$x_t \leftarrow \arg\min_{x \in \mathcal{K}} M_t \cdot x + \frac{1}{\eta_t} \mathcal{D}_{\mathcal{R}}(x, y_{t-1}), \quad \text{and define } g_t := F(x_t),$$
$$y_t \leftarrow \arg\min_{x \in \mathcal{K}} g_t \cdot x + \frac{1}{\eta_t} \mathcal{D}_{\mathcal{R}}(x, y_{t-1})$$

end for
Output: $\bar{x}_T = \frac{1}{T} \sum_{t=1}^{T} x_t$

Solving variational inequalities. So far we have described the online setting where the loss and hint vectors may change arbitrarily. Here we focus on the case where there exists a gap function $\Delta : \mathcal{K} \times \mathcal{K} \mapsto \mathbb{R}$ that is compatible with a given monotone operator $F : \mathcal{K} \mapsto \mathbb{R}^d$ (see Definition 1). Recall that in this setting our goal is to minimize the duality gap induced by Δ. To do so, we choose g_t and M_t in each round as follows,

$$M_t = F(y_{t-1}) ; \quad \text{and} \quad g_t = F(x_t) , \tag{7}$$

These choices correspond to the extragradient (Korpelevich, 1976) and to Mirror-Prox (Nemirovski, 2004) methods.

In Alg. 1 we present our universal Mirror-Prox algorithm for solving variational inequalities. This algorithm combines the Mirror-Prox algorithm (i.e., combining Eq. (7) inside the optimistic OGD of Eq. (3)), together with the new adaptive scheme that we propose in Eq. (5).

Intuition. Before stating the guarantees of Alg. 1, let us give some intuition behind the learning rate that we suggest in Eq. (5). Note that the original Mirror-Prox algorithm employs two extreme learning rates for the non-smooth and smooth cases. In the smooth case the learning rate is *constant*, i.e., $\eta_t \propto 1/L$, and in the non-smooth case it is decaying, i.e., $\eta_t \propto D/(G\sqrt{t})$. Next we show how our adaptive learning rate seems to implicitly adapts to the smoothness of the problem.

For simplicity, let us focus on the convex optimization setting, where our goal is to minimize a convex function $f(\cdot)$, and therefore $F(x) := \nabla f(x)$. Also assume we use $\mathcal{R}(x) := \frac{1}{2} \|x\|_2^2$. In this case, optimistic OGD (Eq. (3)) is simply, $x_t \leftarrow \Pi_{\mathcal{K}}(y_{t-1} - \eta_t M_t)$, and $y_t \leftarrow \Pi_{\mathcal{K}}(y_{t-1} - \eta_t g_t)$, where $\Pi_{\mathcal{K}}$ is the orthogonal projection onto $\mathcal{K}$. Now, let $x^* = \arg\min_{x \in \mathcal{K}} f(x)$, and let us imagine two situations:
(i) If $f(\cdot)$ is non-smooth around x^*, then the norms of the gradients are not decaying as we approach x^*, and in this case the $\|Z_t\|$'s are lower bounded by some constant along all rounds. This implies that η_t will be proportional to $1/\sqrt{t}$.
(ii) Imagine that $f(\cdot)$ is smooth around x^*. If in addition $\nabla f(x^*) = 0$, this intuitively implies that the magnitudes $\|g_t\|$ and $\|M_t\|$ go to zero as we approach x^*, and therefore $\|Z_t\|$'s will also go to zero. This intuitively means that η_t tends to a constant when t tends to infinity. However, note that this behaviour can also be achieved by using an AdaGrad-like (Duchi et al., 2011) learning rate rule, i.e., $\eta_t \propto (\sum_{\tau=1}^{t} \|g_\tau\|^2 + \|M_\tau\|^2)^{-1/2}$. The reason that we employ the more complicated learning rate of Eq. (5) is in order to handle the case where $f(\cdot)$ is smooth yet $\|\nabla f(x^*)\| > 0$. In this case,

the norms $\|g_t\|$ and $\|M_t\|$ will not decay as we approach x^*; nevertheless the norms $\|Z_t\|$'s will intuitively go to zero, implying η_t tends to a constant. Thus, in a sense, our new learning rate rule can be seen as the appropriate adaptation of AdaGrad to the constrained case.

Guarantees. We are now ready to state our guarantees. We show that when the monotone operator F is smooth, then we minimize the *duality gap* in Eq. (2) at a fast rate of $O(1/T)$. Conversely, when the monotone operator is non-smooth, then we obtain a rate of $O(\sqrt{\log T/T})$. This is achieved without any prior knowledge regarding the smoothness of F. The next result addresses the smooth case (we provide a proof sketch in Sec. 3.3; the full proof appears in App. A.3),

Theorem 4 *Assume that F is L-smooth, and G-bounded. Then Alg. 1 used with the learning rate of Eq. (5) implies the following bound,*

$$\text{DualGap}(\bar{x}_T) := \max_{x \in \mathcal{K}} \Delta(\bar{x}_T, x) \le O\Big(\frac{\alpha GD + \alpha^2 LD^2 + LD^2 \log(LD/G_0)}{T}\Big) .$$

Recall that $\alpha := \max\{G/G_0, G_0/G\}$ measures the quality of our prior knowledge G_0 regarding the actual bound G on the norms of $F(\cdot)$. Next we present our guarantees for the non-smooth case.

Theorem 5 *Assume that F is G-bounded. Alg. 1 used with the learning rate of Eq. (5) implies,*

$$\text{DualGap}(\bar{x}_T) := \max_{x \in \mathcal{K}} \Delta(\bar{x}_T, x) \le O\Big(\alpha GD \sqrt{\log T}/\sqrt{T}\Big) .$$

Up to logarithmic terms, we recover the results from Juditsky et al. (2011), with a potential extra factor α, which is equal to 1 if we know a bound G on the norms of the values of F (but we do not require this value to obtain the correct dependence in T). The proof of Thm. 5 appears in App. A.4.

3.3. Proof Sketch of Theorem 4

Proof We shall require the following simple identity (see Rakhlin and Sridharan (2013)),

$$\|g_t - M_t\|_* \cdot \|x_t - y_t\| = \min_{\rho > 0} \Big\{ \frac{\rho}{2} \|g_t - M_t\|_*^2 + \frac{1}{2\rho} \|x_t - y_t\|^2 \Big\}.$$

Using the above with $\rho = 1/L$, together with $g_t := F(x_t), M_t = F(y_{t-1})$, and using the L-smoothness of F gives,

$$\|g_t - M_t\|_* \cdot \|x_t - y_t\| \le \frac{L}{2} \|x_t - y_{t-1}\|^2 + \frac{L}{2} \|x_t - y_t\|^2 .$$

Combining this inside the regret bound of Eq. (4) and re-arranging we obtain,

$$\text{Regret} \le \frac{D^2}{\eta_1} + \frac{D^2}{\eta_T} + \frac{5}{2} \sum_{t=1}^{T} \Big(L - \frac{1}{\eta_t} \Big) \eta_t^2 Z_t^2 , \tag{8}$$

where we have used, $Z_t^2 := \big(\|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2 \big) / 5\eta_t^2$.
Let us define $\tau_* := \max\{t \in [T] : 1/\eta_t \le 2L\}$, and divide the last term of the regret as follows,

$$\sum_{t=1}^{T} \Big(L - \frac{1}{\eta_t} \Big) \eta_t^2 Z_t^2 = \sum_{t=1}^{\tau_*} \Big(L - \frac{1}{\eta_t} \Big) \eta_t^2 Z_t^2 + \sum_{\tau_*+1}^{T} \Big(L - \frac{1}{\eta_t} \Big) \eta_t^2 Z_t^2 \le \sum_{t=1}^{\tau_*} L \eta_t^2 Z_t^2 - \frac{1}{2} \sum_{t=\tau_*+1}^{T} \eta_t Z_t^2 ,$$

where in the second inequality we use $2L \leq \frac{1}{\eta_t}$ which holds for $t > \tau_\star$; implying that $L - \frac{1}{\eta_t} \leq -\frac{1}{2\eta_t}$. Plugging the above back into Eq. (8) we obtain,

$$\text{Regret} \leq \frac{D^2}{\eta_1} + \underbrace{\frac{D^2}{\eta_T} - \frac{5}{4}\sum_{t=\tau_*+1}^{T} \eta_t Z_t^2}_{(A)} + \underbrace{\frac{5}{2}\sum_{t=1}^{\tau_*} L\eta_t^2 Z_t^2}_{(B)} . \tag{9}$$

Next we bound terms (A) and (B) above. To bound (A) we will require the following lemma,

Lemma *For any non-negative numbers $a_1, \ldots, a_n \in [0, a]$, and $a_0 \geq 0$, the following holds:*

$$\sqrt{a_0 + \sum_{i=1}^{n-1} a_i} - \sqrt{a_0} \leq \sum_{i=1}^{n} \frac{a_i}{\sqrt{a_0 + \sum_{j=1}^{i-1} a_j}} \leq \frac{2a}{\sqrt{a_0}} + 3\sqrt{a} + 3\sqrt{a_0 + \sum_{i=1}^{n-1} a_i} .$$

Recalling that $\eta_t = D/\sqrt{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2}$ (see Eq. (5)), and also recalling that $Z_\tau \in [0, G]$ we can use the above lemma to bound term (A),

$$(A): \quad = D\sqrt{G_0^2 + \sum_{t=1}^{T-1} Z_t^2} - \frac{5D}{4}\sum_{t=\tau_*+1}^{T} \frac{Z_t^2}{\sqrt{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2}}$$

$$\leq DG_0 + D\sum_{t=1}^{T} \frac{Z_t^2}{\sqrt{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2}} - \frac{5D}{4}\sum_{t=\tau_*+1}^{T} \frac{Z_t^2}{\sqrt{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2}}$$

$$\leq DG_0 + D\sum_{t=1}^{\tau_*} \frac{Z_t^2}{\sqrt{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2}}$$

$$\leq 3D(G + G_0) + 2DG^2/G_0 + 3D^2\frac{1}{\eta_{\tau_*}} \leq 3D(G + G_0) + 2DG^2/G_0 + 6LD^2 , \tag{10}$$

where we have used the definition of τ_* which implies $1/\eta_{\tau*} \leq 2L$.

Bounding term (B): In the full proof (Appendix A.3) we show that $(B) \leq O(LD^2 \log(LD/G_0))$.

Conclusion: Combining the bounds on (A) and (B) into Eq. (9) and using $\eta_1 = D/G_0$ implies,

$$\text{Regret} \leq O\left(\alpha DG + \alpha^2 LD^2 + LD^2 \log(LD/G_0)\right) ,$$

where we used the definition $\alpha := \max\{G/G_0, G_0/G\}$. Combining the above with the definition of $\bar{x}_T$ and using Jensen's inequality (recall Δ is convex in its first argument), as well as with the fact that Δ is a compatible gap function w.r.t. F concludes the proof. ∎

4. Stochastic Setting

In this section we present the stochastic variational inequality setting. Then we show that using the exact same universal Mirror-Prox algorithm (Alg. 1) that we have presented in the previous section, enables to provide the optimal guarantees for the stochastic setting. This is done without any prior knowledge regarding the smoothness or the stochasticity of the problem.

Setting. The stochastic setting is similar to the deterministic setting that we have described in Sec. 2. The only difference is that we do not have an access to the exact values of F. Instead, we assume that when querying a point $x \in \mathcal{K}$ we receive an unbiased noisy estimate of the exact monotone mapping $F(x)$. More formally, we have an access to an oracle $\tilde{F} : \mathcal{K} \mapsto \mathbb{R}^d$, such that for any $x \in \mathcal{K}$ we have,

$$\mathbf{E}[\tilde{F}(x)|x] = F(x) \ .$$

We also assume to have a bound G on the dual norms of $\tilde{F}$, i.e., almost surely, $\|\tilde{F}(x)\|_* \leq G; \ \forall x \in \mathcal{K}$. We are now ready to state our guarantees. Up to logarithmic terms, we recover the results from Juditsky et al. (2011) with an universal algorithm that does not need the knowledge of the various regularity constants. The first results regards the non-smooth noisy case.

Theorem 6 *Assume that we receive unbiased (noisy) estimates $\tilde{F}$ instead of F inside Alg. 1. Then Alg. 1 used with the learning rate of Eq. (5) ensures the following,*

$$\mathbf{E}\left[\mathrm{DualGap}(\bar{x}_T)\right] := \mathbf{E} \max_{x \in \mathcal{K}} \Delta(\bar{x}_T, x) \leq O\left(\alpha GD\sqrt{\log T}/\sqrt{T}\right) \ .$$

Next we further assume a bound on the variance of $\tilde{F}$, i.e., $\mathbf{E}\left[\|\tilde{F}(x) - F(x)\|_*^2 | x\right] \leq \sigma^2, \ \forall x \in \mathcal{K}$, but we do not assume any prior knowledge of σ. The next theorem regards the smooth noisy case.

Theorem 7 *Assume that F is L-smooth, and assume that we receive unbiased (noisy) estimates $\tilde{F}$ instead of F inside Alg. 1. Then Alg. 1 used with the learning rate of Eq. (5) ensures the following,*

$$\mathbf{E}\left[\mathrm{DualGap}(\bar{x}_T)\right] := \mathbf{E} \max_{x \in \mathcal{K}} \Delta(\bar{x}_T, x) \leq O\left(\frac{\alpha GD + \alpha^2 LD^2 + LD^2 \log(LD/G_0)}{T} + \frac{\alpha\sigma D\sqrt{\log T}}{\sqrt{T}}\right) \ .$$

4.1. Proof Sketch of Theorem 6

Proof Let us denote by $\tilde{g}_t$ the noisy estimates of $g_t := F(x_t)$. Following the exact steps as in the proof of Theorem 5 implies the following holds w.p. 1: $\sum_{t=1}^T \tilde{g}_t \cdot (x_t - x) \leq O(\alpha GD\sqrt{T \log T})$. Recalling the definition of $\bar{x}_T$, and using Jensen's inequality implies that for any $x \in \mathcal{K}$,

$$
\begin{aligned}
T \cdot \Delta(\bar{x}_T, x) \ &\leq \ \sum_{t=1}^T \Delta(x_t, x) \ \leq \ \sum_{t=1}^T g_t \cdot (x_t - x) \\
&= \ \sum_{t=1}^T \tilde{g}_t \cdot (x_t - x) - \sum_{t=1}^T \zeta_t \cdot (x_t - x) \ \leq \ O(\alpha GD\sqrt{T \log T}) - \sum_{t=1}^T \zeta_t \cdot (x_t - x) \ ,
\end{aligned}
$$

$$(11)$$

where we denote, $\zeta_t := \tilde{g}_t - g_t$. And clearly $\{\zeta_t\}_t$ is a martingale difference sequence.

Let $x^* := \arg\max_{x \in \mathcal{K}} \Delta(\bar{x}_t, x)$. Taking $x = x^*$ and taking expectation over Eq. (11) gives,

$$T \cdot \mathbf{E}\Delta(\bar{x}_t, x^*) \ \leq \ O(\alpha GD\sqrt{T \log T}) - \mathbf{E}\sum_{t=1}^T \zeta_t \cdot (x_t - x^*) \ = \ O(\alpha GD\sqrt{T \log T}) + \mathbf{E}\sum_{t=1}^T \zeta_t \cdot x^* \ .$$

To establish the proof we are left to show that $\mathbf{E}\sum_{t=1}^T \zeta_t \cdot x^* \leq O(GD\sqrt{T})$. This is challenging since x^*, by its definition, is a random variable that may depend on $\{\zeta_t\}_t$, implying that $\zeta_t \cdot x^*$ is not zero-mean. Nevertheless, we are able to make use of the martingale difference property of $\{\zeta_t\}_t$ in order to bound $\mathbf{E}\sum_{t=1}^T \zeta_t \cdot x^*$ (see Proposition 11 in the appendix). ∎

5. Applications

Here we show how the general Variational inequalities setting described in Section 2 captures two important settings, namely convex optimization and convex-concave zero-sum games.

5.1. Convex Optimization

Assume that $\mathcal{K}$ is a convex set, and $f : \mathcal{K} \mapsto \mathbb{R}$ is convex over $\mathcal{K}$. In the convex optimization setting our goal is to minimize f, i.e.,

$$\min_{x \in \mathcal{K}} f(x) .$$

We assume that we may query (sub)gradients of f. Next we show how this setting is captured by the variational inequality setting. Let us define a gap function and an operator F as follows,

$$\Delta(x, y) := f(x) - f(y), \quad \forall x, y, \in \mathcal{K} , \qquad \& \qquad F(x) := \nabla f(x), \quad \forall x \in \mathcal{K}.$$

Then by the (sub)gradient inequality for convex functions, it immediately follows that Δ is a compatible gap function with respect to F. Also, it is clear that $\Delta(x, y)$ is convex with respect to x. Finally, note that the duality gap in this case is the natural sub-optimality measure, i.e.,

$$\text{DualGap}(x) := \max_{y \in \mathcal{K}} \Delta(x, y) = f(x) - \min_{y \in \mathcal{K}} f(y) .$$

Moreover, if f is L-smooth w.r.t. a norm $\| \cdot \|$, then F is smooth with respect to the same norm.

5.2. Convex-Concave Zero-sum Games

Let $\phi : \mathcal{U} \times \mathcal{V} \mapsto \mathbb{R}$, where $\phi(u, v)$ is convex in u and concave in v, and $\mathcal{U} \subseteq \mathbb{R}^{d_1}, \mathcal{V} \subseteq \mathbb{R}^{d_2}$, are compact convex sets. The convex-concave zero-sum game induced by ϕ is defined as follows,

$$\min_{u \in \mathcal{U}} \max_{v \in \mathcal{V}} \phi(u, v) .$$

The performance measure for such games is the duality gap which is defined as,

$$\text{DualGap}(u, v) = \max_{v \in \mathcal{V}} \phi(u, v) - \min_{u \in \mathcal{U}} \phi(u, v) . \tag{12}$$

The duality gap is always non-negative, and we seek an (approximate) equilibrium, i.e., a point (u^*, v^*) such that $\text{DualGap}(u^*, v^*) = 0$.

This setting can be classically described as a variational inequality problem. Let us denote,

$$x := (u, v) \in \mathcal{U} \times \mathcal{V} ; \quad \text{and} \quad \mathcal{K} := \mathcal{U} \times \mathcal{V} .$$

For any $x = (u, v), x_0 = (u_0, v_0) \in \mathcal{K}$, define a gap function and an operator $F : \mathcal{K} \mapsto \mathbb{R}^{d_1 + d_2}$, as follows,

$$\Delta(x, x_0) := \phi(u, v_0) - \phi(u_0, v) , \quad \text{and} \quad F(x) := (\nabla_u \phi(u, v), -\nabla_v \phi(u, v)) .$$

It is immediate to show that this gap function, Δ, induces the duality gap appearing in Eq. (12), i.e., $\text{DualGap}(x) := \max_{x_0 \in \mathcal{K}} \Delta(x, x_0)$. Also, from the convex-concavity of ϕ it immediately follows that $\Delta(x, x_0)$ is convex in x. The next lemma from Nemirovski (2004) shows that Δ is a gap function compatible with F (for completeness we provide its proof in Appendix A.1),

Lemma 8 *The following applies for any $x := (u, v), x_0 := (u_0, v_0) \in \mathcal{U} \times \mathcal{V}$:*

$$\Delta(x, x_0) := \phi(u, v_0) - \phi(u_0, v) \leq F(x) \cdot (x - x_0) .$$

Mirror Map for Zero-sum Games. In this work, our variational inequality method employs a mirror-map over $\mathcal{K}$. For the case of zero-sum games $\mathcal{K} := \mathcal{U} \times \mathcal{V}$, and we usually have separate mirror-map terms, $\mathcal{R}_{\mathcal{U}} : \mathcal{U} \mapsto \mathbb{R}$, and $\mathcal{R}_{\mathcal{V}} : \mathcal{V} \mapsto \mathbb{R}$. Juditsky and Nemirovski (2011) have found a way to appropriately define a mirror-map over $\mathcal{K}$ using $\mathcal{R}_{\mathcal{U}}, \mathcal{R}_{\mathcal{V}}$. We hereby describe it.

Assume that the separate mirror-maps are 1-strongly convex w.r.t. norms $\|\cdot\|_{\mathcal{U}}$ and $\|\cdot\|_{\mathcal{V}}$, and let $\|\cdot\|_{\mathcal{U}}^*$ and $\|\cdot\|_{\mathcal{V}}^*$ be the respective dual norms. Also, define $D_{\mathcal{U}}^2 := \max_{u \in \mathcal{U}} \mathcal{R}_{\mathcal{U}}(u) - \min_{u \in \mathcal{U}} \mathcal{R}_{\mathcal{U}}(u)$, and similarly define $D_{\mathcal{V}}^2$. Juditsky and Nemirovski (2011) suggest to employ,

$$\mathcal{R}_{\mathcal{K}}(x) = \frac{1}{D_{\mathcal{U}}^2}\mathcal{R}_{\mathcal{U}}(u) + \frac{1}{D_{\mathcal{V}}^2}\mathcal{R}_{\mathcal{V}}(v) \,; \qquad \forall x := (u, v) \in \mathcal{K} \,,$$

and to define,

$$\|x\|_{\mathcal{K}} := \sqrt{\|u\|_{\mathcal{U}}^2/D_{\mathcal{U}}^2 + \|v\|_{\mathcal{V}}^2/D_{\mathcal{V}}^2} \,; \qquad \forall x := (u, v) \in \mathcal{K} \,. \tag{13}$$

In this case $\mathcal{R}_{\mathcal{K}}$ is 1-strongly-convex w.r.t. $\|\cdot\|_{\mathcal{K}}$. Also, the dual norm of $\|\cdot\|_{\mathcal{K}}$ in this case is,

$$\|x\|_{\mathcal{K}}^* := \sqrt{D_{\mathcal{U}}^2(\|u\|_{\mathcal{U}}^*)^2 + D_{\mathcal{V}}^2(\|v\|_{\mathcal{V}}^*)^2} \,; \qquad \forall x := (u, v) \in \mathbb{R}^{d_1} \times \mathbb{R}^{d_2} \,. \tag{14}$$

Smooth Zero-sum Games. It can be shown that if the gradient mapping $\nabla_u \phi(u, v)$, and $\nabla_v \phi(u, v)$ are Lipschitz-continuous with respect to both u and v, then the monotone operator F defined through $F(x) := (\nabla_u \phi(u, v), -\nabla_v \phi(u, v))$ is also smooth. Concretely, let $\|\cdot\|_{\mathcal{U}}$, and $\|\cdot\|_{\mathcal{V}}$ be norms over $\mathcal{U}$ and $\mathcal{V}$, and let $\|\cdot\|_{\mathcal{U}}^*$, and $\|\cdot\|_{\mathcal{V}}^*$ be their respective dual norms. Juditsky and Nemirovski (2011) show that if the following holds $\forall u, u' \in \mathcal{U}, v, v' \in \mathcal{V}$,

$$\|\nabla_u \phi(u, v) - \nabla_u \phi(u', v)\|_{\mathcal{U}}^* \leq L_{11}\|u - u'\|_{\mathcal{U}} \,; \qquad \|\nabla_v \phi(u, v) - \nabla_v \phi(u, v')\|_{\mathcal{V}}^* \leq L_{22}\|v - v'\|_{\mathcal{V}} \,;$$
$$\|\nabla_u \phi(u, v) - \nabla_u \phi(u, v')\|_{\mathcal{V}}^* \leq L_{12}\|v - v'\|_{\mathcal{V}} \,; \qquad \|\nabla_v \phi(u, v) - \nabla_v \phi(u', v)\|_{\mathcal{U}}^* \leq L_{21}\|u - u'\|_{\mathcal{U}} \,;$$

Then it can be shown that $\forall x, x' \in \mathcal{K} \; \|F(x) - F(x')\|_{\mathcal{K}}^* \leq L\|x - x'\|_{\mathcal{K}}$, where $\|\cdot\|_K$, and $\|\cdot\|_{\mathcal{K}}^*$ are defined in Equations (13) and (14), and, $L := 2\max\{L_{11}D_{\mathcal{U}}^2, L_{22}D_{\mathcal{V}}^2, L_{12}D_{\mathcal{U}}D_{\mathcal{V}}, L_{21}D_{\mathcal{U}}D_{\mathcal{V}}\}$.

6. Conclusion

In this paper, we have presented a universal algorithm for variational inequalities, that can adapt to smoothness and noise, leading, with a single algorithm with very little knowledge of the problem to the best convergence rates in all these set-ups (up to logarithmic factors). There are several avenues worth exploring: (a) an extension to a Matrix-AdaGrad-like algorithm (McMahan and Streeter, 2010; Duchi et al., 2011) where a matrix gain is employed rather than a scalar step-size, (b) an extension that could handle composite problems through additional proximal operators, (c) extensions of adaptivity to all Hölder-continuous mappings (Dvurechensky et al., 2018), and finally (d) the inclusion of deterministic error terms to allow biased operator evaluations.

References

Heinz H. Bauschke and Patrick L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*, volume 408. Springer, 2011.

Chao-Kai Chiang, Tianbao Yang, Chia-Jung Lee, Mehrdad Mahdavi, Chi-Jen Lu, Rong Jin, and Shenghuo Zhu. Online optimization with gradual variations. In *Conference on Learning Theory*, pages 6–1, 2012.

Constantinos Daskalakis, Alan Deckelbaum, and Anthony Kim. Near-optimal no-regret algorithms for zero-sum games. In *Proceedings of the Twenty-Second Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 235–254, 2011.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Pavel Dvurechensky, Alexander Gasnikov, Fedor Stonyakin, and Alexander Titov. Generalized mirror prox: Solving variational inequalities with monotone operator, inexact oracle, and unknown Hölder parameters. Technical Report 1806.05140, arXiv, 2018.

Yoav Freund, Robert E Schapire, et al. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Anatoli Juditsky and Arkadi Nemirovski. First order methods for nonsmooth convex large-scale optimization, ii: utilizing problems structure. *Optimization for Machine Learning*, pages 149–183, 2011.

Anatoli Juditsky and Arkadi Nemirovski. Solving variational inequalities with monotone operators on domains given by linear minimization oracles. *Mathematical Programming*, 156(1-2):221–256, 2016.

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

Anatoli Juditsky, Fatma Kılınç Karzan, and Arkadi Nemirovski. Randomized first order algorithms with applications to 1-minimization. *Mathematical Programming*, 142(1-2):269–310, 2013.

Anatoli B Juditsky and Arkadi S Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

Sham Kakade. Lecture notes in multivariate analysis, dimensionality reduction, and spectral methods. http://stat.wharton.upenn.edu/~skakade/courses/stat991_mult/lectures/MatrixConcen.pdf, April 2010.

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

G. M. Korpelevich. Extrapolational gradient methods and their connection with modified lagrangians. *Ehkon. Mat. Metody*, 19:694–703, 1983.

Kfir Levy. Online to offline conversions, universality and adaptive minibatch sizes. In *Advances in Neural Information Processing Systems*, pages 1612–1621, 2017.

Yehuda Kfir Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. In *Advances in Neural Information Processing Systems*, pages 6501–6510, 2018.

H. Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *COLT 2010*, page 244, 2010.

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

Arkadii Nemirovskii and David Borisovich Yudin. Problem complexity and method efficiency in optimization. 1983.

A. S. Nemirovsky. Information-based complexity of linear operator equations. *Journal of Complexity*, 8(2):153–175, 1992.

Yurii Nesterov. Universal gradient methods for convex optimization problems. *Mathematical Programming*, 152(1-2):381–404, 2015.

Muhammad Aslam Noor. New extragradient-type methods for general variational inequalities. *Journal of Mathematical Analysis and Applications*, 277(2):379–394, 2003.

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

Alp Yurtsever, Quoc Tran Dinh, and Volkan Cevher. A universal primal-dual convex optimization framework. In *Advances in Neural Information Processing Systems*, pages 3150–3158, 2015.

Appendix Description: In Appendix A, we provide the missing proofs related to the Deterministic Setting (Section 3): the proofs of Thm. 4 and Thm. 5 appear in App. A.3 and A.4. We also provide the proofs of Lemma 8 (see App A.1), and Lemma Lemma 3 (see App. A.2).

In Appendix B, we provide the missing proofs related to the Stochastic Setting (Section 4): the proofs of Thm. 6 and Thm. 7 appear in App. B.1 and B.2. And we also prove Proposition 11 (see App. B.3), which is a central tool in the proofs of the stochastic case. In Appendix C we provide the remaining proofs for the paper.

Appendix A. Proofs for the Deterministic Setting (Section 3)

A.1. Proof of Lemma 8

Proof Using convexity we get for any $u_0 \in \mathcal{U}, v \in \mathcal{V}$,

$$\phi(u, v) - \phi(u_0, v) \le \nabla_u \phi(u, v) \cdot (u - u_0) .$$

Similarly, using concavity we get for any $u \in \mathcal{U}, v_0 \in \mathcal{V}$,

$$-\phi(u, v) + \phi(u, v_0) \le -\nabla_v \phi(u, v) \cdot (v - v_0) .$$

Summing both of the above equations gives for any $x := (u, v), x_0 := (u_0, v_0)$,

$$\phi(u, v_0) - \phi(u_0, v) \le F(x) \cdot (x - x_0) .$$

where we used $F(x) := (\nabla_u \phi(u, v), -\nabla_v \phi(u, v))$. This concludes the proof. ∎

A.2. Proof of Lemma 3

Proof Here we show that $\|x_t - y_{t-1}\|/\eta_t \le G$; the proof of $\|y_t - y_{t-1}\|/\eta_t \le G$ follows the exact same steps.

Before we start, note that the following holds for Bregman Divergences,

$$\nabla_x \mathcal{D}_\mathcal{R}(x, y) = \nabla \mathcal{R}(x) - \nabla \mathcal{R}(y) .$$

Now, the following applies for any $x \in \mathcal{K}$ by the definition of x_t in Eq. (3),

$$\nabla_{x:=x_t} (\eta_t M_t \cdot x + \mathcal{D}_\mathcal{R}(x, y_{t-1})) \cdot (x - x_t) \ge 0 .$$

Taking $x = y_{t-1}$, the above implies,

$$(\eta_t M_t + \nabla \mathcal{R}(x_t) - \nabla \mathcal{R}(y_{t-1})) \cdot (y_{t-1} - x_t) \ge 0 .$$

Combing this with $\mathcal{R}$ being 1-strongly-convex w.r.t. $\|\cdot\|$ we obtain,

$$\eta_t M_t \cdot (y_{t-1} - x_t) \ge (\nabla \mathcal{R}(y_{t-1}) - \nabla \mathcal{R}(x_t)) \cdot (y_{t-1} - x_t) \ge \|x_t - y_{t-1}\|^2 .$$

Using Cauchy-Swartz immediately implies that,

$$\|x_t - y_{t-1}\|^2 \le \eta_t \|x_t - y_{t-1}\| \cdot \|M_t\|_* \le \eta_t \|x_t - y_{t-1}\| \cdot G .$$

Dividing the above equation by $\|x_t - y_{t-1}\|$ concludes the proof. ∎

A.3. Proof of Theorem 4

Proof We shall require the following simple identity (see Rakhlin and Sridharan (2013)),

$$\|g_t - M_t\|_* \cdot \|x_t - y_t\| = \min_{\rho > 0} \left\{ \frac{\rho}{2} \|g_t - M_t\|_*^2 + \frac{1}{2\rho} \|x_t - y_t\|^2 \right\}.$$

Using the above with $\rho = 1/L$ we get,

$$
\begin{aligned}
\|g_t - M_t\|_* \cdot \|x_t - y_t\| &\leq \frac{1}{2L} \|g_t - M_t\|_*^2 + \frac{L}{2} \|x_t - y_t\|^2 \\
&= \frac{1}{2L} \|F(x_t) - F(y_{t-1})\|_*^2 + \frac{L}{2} \|x_t - y_t\|^2 \\
&\leq \frac{L}{2} \|x_t - y_{t-1}\|^2 + \frac{L}{2} \|x_t - y_t\|^2 ,
\end{aligned}
\tag{15}
$$

where the last line uses the L-smooth of the operator F, i.e., $\|F(x_t) - F(y_{t-1})\|_* \leq L\|x_t - y_{t-1}\|$. Combining the above inside the regret bound of Eq. (4) we obtain,

$$
\begin{aligned}
\text{Regret} &\leq \frac{D^2}{\eta_1} + \frac{D^2}{\eta_T} + \frac{1}{2} \sum_{t=1}^{T} \left(L - \frac{1}{\eta_t} \right) \left(\|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2 \right) \\
&= \frac{D^2}{\eta_1} + \frac{D^2}{\eta_T} + \frac{5}{2} \sum_{t=1}^{T} \left(L - \frac{1}{\eta_t} \right) \eta_t^2 Z_t^2 ,
\end{aligned}
\tag{16}
$$

where we have used, $Z_t^2 := \left(\|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2 \right) / 5\eta_t^2$.

Now let us define $\tau_* := \max\{t \in [T] : 1/\eta_t \leq 2L\}$. We can now divide the last term of the regret according to τ_*,

$$
\begin{aligned}
\sum_{t=1}^{T} \left(L - \frac{1}{\eta_t} \right) \eta_t^2 Z_t^2 &= \sum_{t=1}^{\tau_*} \left(L - \frac{1}{\eta_t} \right) \eta_t^2 Z_t^2 + \sum_{\tau_*+1}^{T} \left(L - \frac{1}{\eta_t} \right) \eta_t^2 Z_t^2 \\
&\leq \sum_{t=1}^{\tau_*} L \eta_t^2 Z_t^2 - \frac{1}{2} \sum_{t=\tau_*+1}^{T} \eta_t Z_t^2 .
\end{aligned}
$$

where in the second line we use $2L \leq \frac{1}{\eta_t}$ which holds for $t > \tau_\star$, implying that $L - \frac{1}{\eta_t} \leq -\frac{1}{2\eta_t}$. Plugging the above back into Eq. (16) we obtain,

$$
\text{Regret} \leq \frac{D^2}{\eta_1} + \underbrace{\frac{D^2}{\eta_T} - \frac{5}{4} \sum_{t=\tau_*+1}^{T} \eta_t Z_t^2}_{(A)} + \underbrace{\frac{5}{2} \sum_{t=1}^{\tau_*} L \eta_t^2 Z_t^2}_{(B)} .
\tag{17}
$$

Next we bound terms (A) and (B) above,

Bounding term (A): We will require the following lemma which we prove in Appendix C.2,

Lemma 9 *For any non-negative numbers $a_1, \ldots, a_n \in [0, a]$, and $a_0 \geq 0$, the following holds:*

$$
\sqrt{a_0 + \sum_{i=1}^{n-1} a_i} - \sqrt{a_0} \leq \sum_{i=1}^{n} \frac{a_i}{\sqrt{a_0 + \sum_{j=1}^{i-1} a_j}} \leq \frac{2a}{\sqrt{a_0}} + 3\sqrt{a} + 3\sqrt{a_0 + \sum_{i=1}^{n-1} a_i} \, .
$$

Recalling the learning rate rule that we use, $\eta_t = D/\sqrt{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2}$ (see Eq. (5)), and also recalling that $Z_\tau \in [0, G]$ we can use the above lemma to bound term (A),

$$
\begin{aligned}
(A) &:= \frac{D^2}{\eta_T} - \frac{5}{4} \sum_{t=\tau_*+1}^{T} \eta_t Z_t^2 \\
&= D\sqrt{G_0^2 + \sum_{t=1}^{T-1} Z_t^2} - \frac{5}{4} \sum_{\tau_*+1}^{T} \eta_t Z_t^2 \\
&\leq DG_0 + D\sum_{t=1}^{T} \frac{Z_t^2}{\sqrt{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2}} - \frac{5}{4} \sum_{t=\tau_*+1}^{T} \eta_t Z_t^2 \\
&= DG_0 + \sum_{t=1}^{T} \eta_t Z_t^2 - \frac{5}{4} \sum_{t=\tau_*+1}^{T} \eta_t Z_t^2 \\
&\leq DG_0 + \sum_{t=1}^{\tau_*} \eta_t Z_t^2 \\
&= DG_0 + D\sum_{t=1}^{\tau_*} \frac{Z_t^2}{\sqrt{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2}} \\
&\leq DG_0 + \frac{2DG^2}{G_0} + 3DG + 3D\sqrt{G_0^2 + \sum_{t=1}^{\tau^*-1} Z_t^2} \\
&\leq 3D(G + G_0) + \frac{2DG^2}{G_0} + 3D^2 \frac{1}{\eta_{\tau_*}} \\
&\leq 3D(G + G_0) + \frac{2DG^2}{G_0} + 6LD^2 \, ,
\end{aligned}
\tag{18}
$$

where we have used the definition of τ_* which implies $1/\eta_{\tau*} \leq 2L$.

Bounding term (B): We will require the following lemma (proof is found in Appendix C.4),

Lemma 10 *For any non-negative real numbers $a_1, \ldots, a_n \in [0, a]$, and $a_0 \geq 0$,*

$$
\sum_{i=1}^{n} \frac{a_i}{a_0 + \sum_{j=1}^{i-1} a_j} \leq 2 + \frac{4a}{a_0} + 2\log\left(1 + \sum_{i=1}^{n-1} a_i/a_0\right) \, .
$$

Recalling the learning rate rule that we use, $\eta_t = D/\sqrt{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2}$ (see Eq. (5)), and also recalling that $Z_\tau \in [0, G]$ we can use the above lemma to bound term (B),

$$
\begin{aligned}
(\text{B}) &= \frac{L}{2} \sum_{t=1}^{\tau_*} \eta_t^2 Z_t^2 \\
&= \frac{LD^2}{2} \sum_{t=1}^{\tau_*} \frac{Z_t^2}{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2} \\
&\leq LD^2 + 2LD^2 \frac{G^2}{G_0^2} + LD^2 \log \left(\frac{G_0^2 + \sum_{t=1}^{\tau_*-1} Z_t^2}{G_0^2} \right) \\
&\leq 3LD^2 \max\{1, G^2/G_0^2\} + LD^2 \log \left(\frac{(D/G_0)^2}{\eta_{\tau_*}^2} \right) \\
&= 3LD^2 \max\{1, G^2/G_0^2\} + 2LD^2 \log \left(2LD/G_0 \right) ,
\end{aligned}
\tag{19}
$$

where we have used the definition of τ_* which implies $1/\eta_{\tau*} \leq 2L$.

Conclusion: Combining Equations (18) and (19) into Eq. (17) and using $\eta_1 = D/G_0$ implies the following regret bound for Alg. 1,

$$
\text{Regret} \leq O\left(\alpha DG + \alpha^2 LD^2 + LD^2 \log(LD/G_0) \right) ,
$$

where we used the definition $\alpha := \max\{G/G_0, G_0/G\}$. Combining the above with the definition of $\bar{x}_T$ and using Jensen's inequality implies,

$$
\Delta(\bar{x}_T, x) \leq O\left(\frac{\alpha DG + \alpha^2 LD^2 + LD^2 \log(LD/G_0)}{T} \right), \qquad \forall x \in \mathcal{K} .
$$

where we also used the fact that $\forall x_t, x \in \mathcal{K}$ the following holds,

$$
\Delta(x_t, x) \leq F(x) \cdot (x_t - x) := g_t \cdot (x_t - x) .
$$

■

A.4. Proof of Theorem 5

Proof Recall the regret bound of Eq. (4),

$$
\text{Regret} \le \underbrace{\frac{D^2}{\eta_1} + \frac{D^2}{\eta_T}}_{(A)} \underbrace{- \frac{1}{2} \sum_{t=1}^{T} \eta_t^{-1} \left(\|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2 \right)}_{(B)} + \underbrace{\sum_{t=1}^{T} \|g_t - M_t\|_* \cdot \|x_t - y_t\|}_{(C)} .
$$

$$(20)$$

Next we separately bound each of the above terms,

Bounding (A): Recalling the learning rate that we employ (see Eq. (5)) we have,

$$
(A) := \frac{D^2}{\eta_1} + \frac{D^2}{\eta_T} = DG_0 + D \sqrt{G_0^2 + \sum_{t=1}^{T-1} Z_t^2} .
$$

$$(21)$$

Bounding (B): Recalling Lemma 9 we may bound this term as follows,

$$
\begin{aligned}
(B) &:= \frac{1}{2} \sum_{t=1}^{T} \eta_t^{-1} \left(\|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2 \right) \\
&= \frac{5}{2} \sum_{t=1}^{T} \eta_t Z_t^2 \\
&= \frac{5D}{2} \sum_{t=1}^{T} \frac{Z_t^2}{\sqrt{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2}} \\
&\ge \frac{5D}{2} \sqrt{G_0^2 + \sum_{t=1}^{T-1} Z_t^2} - \frac{5DG_0}{2} ,
\end{aligned}
$$

$$(22)$$

where the second line uses $Z_t^2 := \left(\|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2 \right) / 5\eta_t^2$.

Bounding (C): Lets us recall Lemma 10,

Lemma *[Lemma 10] For any non-negative real numbers $a_1, \dots, a_n \in [0, a]$, and $a_0 \ge 0$,*

$$
\sum_{i=1}^{n} \frac{a_i}{a_0 + \sum_{j=1}^{i-1} a_j} \le 2 + \frac{4a}{a_0} + 2 \log \left(1 + \sum_{i=1}^{n-1} a_i / a_0 \right) .
$$

Recalling that $Z_t \in [0, G]$ we may use the above lemma to bound term (C),

$$
\begin{aligned}
(\mathrm{C}) &:= \sum_{t=1}^{T} \|g_t - M_t\|_* \cdot \|x_t - y_t\| \\
&\leq \sum_{t=1}^{T} 2G \cdot \sqrt{\|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2} \\
&= 2\sqrt{5}G \sum_{t=1}^{T} \eta_t Z_t \\
&\leq 5G\sqrt{T} \sqrt{\sum_{t=1}^{T} \eta_t^2 Z_t^2} \\
&\leq 5GD\sqrt{T} \sqrt{\sum_{t=1}^{T} \frac{Z_t^2}{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2}} \\
&\leq 5GD\sqrt{T} \sqrt{2 + 4(G^2/G_0^2) + 2\log\left(\frac{G_0^2 + \sum_{t=1}^{T-1} Z_t^2}{G_0^2}\right)} \\
&\leq 5GD\sqrt{T} \sqrt{2 + 4(G^2/G_0^2) + 2\log(1 + TG^2/G_0^2)} \,,
\end{aligned}
\tag{23}
$$

where the second line uses the bound G on the magnitude of the gradients (in the dual norm). The third line uses $Z_t^2 := \left(\|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2\right)/5\eta_t^2$.

Conclusion: Combining Equations (21) (22) and (23) into Eq. (20) implies the following regret bound for Alg. 1,

$$
\text{Regret} \leq 4DG_0 + 5GD\sqrt{T}\sqrt{2 + 4(G^2/G_0^2) + 2\log(1 + TG^2/G_0^2)} \,.
$$

Combining the above with the definition of $\bar{x}_T$ and using Jensen's inequality implies,

$$
\Delta(\bar{x}_T, x) \leq O\left(\frac{\alpha GD\sqrt{\log T}}{\sqrt{T}}\right) \,, \qquad \forall x \in \mathcal{K} \,.
$$

where we used the notation $\alpha := \max\{G/G_0, G_0/G\}$. We also used the fact that $\forall x_t, x \in \mathcal{K}$ the following holds,

$$
\Delta(x_t, x) \leq F(x) \cdot (x_t - x) := g_t \cdot (x_t - x) \,.
$$

∎

Appendix B. Proofs for the Stochastic Setting (Section 4)

B.1. Proof of Theorem 6

Proof Let us denote by $\tilde{g}_t$ the noisy estimates of $g_t := F(x_t)$, and by $\tilde{M}_t$ the noisy estimates of $M_t := F(y_{t-1})$. Recalling the regret bound of Eq. (4) implies that for any $x \in \mathcal{K}$,

$$\sum_{t=1}^{T} \tilde{g}_t \cdot (x_t - x)$$

$$\leq \frac{D^2}{\eta_1} + \frac{D^2}{\eta_T} + \sum_{t=1}^{T} \|\tilde{g}_t - \tilde{M}_t\|_* \cdot \|x_t - y_t\| - \frac{1}{2} \sum_{t=1}^{T} \eta_t^{-1} \left(\|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2 \right) .$$

$$(24)$$

Now following the exact steps as in the proof of Theorem 5 implies the following holds w.p. 1,

$$\sum_{t=1}^{T} \tilde{g}_t \cdot (x_t - x) \leq O(\alpha G D \sqrt{T \log T}) .$$

Recalling the definition of $\bar{x}_T$, and using Jensen's inequality we obtain for any $x \in \mathcal{K}$,

$$T \cdot \Delta(\bar{x}_T, x) \leq \sum_{t=1}^{T} \Delta(x_t, x)$$

$$\leq \sum_{t=1}^{T} g_t \cdot (x_t - x)$$

$$= \sum_{t=1}^{T} \tilde{g}_t \cdot (x_t - x) - \sum_{t=1}^{T} \zeta_t \cdot (x_t - x)$$

$$\leq O(\alpha G D \sqrt{T \log T}) - \sum_{t=1}^{T} \zeta_t \cdot (x_t - x) , \quad (25)$$

where we have used the following notation,

$$\zeta_t := \tilde{g}_t - g_t .$$

and clearly we have $\mathbf{E}[\zeta_t | x_t] = 0$, and $\{\zeta_t\}_t$ is a martingale difference sequence.

Let $x^* := \arg\max_{x \in \mathcal{K}} \Delta(\bar{x}_t, x)$. Taking $x = x^*$ and taking expectation over Eq. (25) gives,

$$T \cdot \mathbf{E}\Delta(\bar{x}_t, x^*) \leq O(\alpha G D \sqrt{T \log T}) - \mathbf{E} \sum_{t=1}^{T} \zeta_t \cdot (x_t - x^*)$$

$$= O(\alpha G D \sqrt{T \log T}) + \mathbf{E} \sum_{t=1}^{T} \zeta_t \cdot x^* . \quad (26)$$

Thus, to establish the proof we are left to show that $\mathbf{E} \sum_{t=1}^{T} \zeta_t \cdot x^* \leq O(G D \sqrt{T \log T})$. We will require the following proposition (its proof appears in Appendix B.3),

Proposition 11 *Let $\mathcal{K} \subseteq \mathbb{R}^d$ be a convex set, and $\mathcal{R} : \mathcal{K} \mapsto \mathbb{R}$ be a 1-strongly-convex function w.r.t. a norm $\| \cdot \|$ over $\mathcal{K}$. Also assume that $\forall x \in \mathcal{K}; \ \mathcal{R}(x) - \min_{x \in \mathcal{K}} \mathcal{R}(x) \leq \frac{1}{2} D^2$. Then for any martingale difference sequence $(Z_i)_{i=1}^n \in \mathbb{R}^d$, and any random vector X defined over $\mathcal{K}$, we have,*

$$\mathbf{E}\Big[\Big(\sum_{i=1}^n Z_i\Big)^\top X\Big] \leq \frac{D}{2} \sqrt{\sum_{i=1}^n \mathbf{E}\|Z_i\|_*^2} \,,$$

where $\| \cdot \|_*$ is the dual norm of $\| \cdot \|$.

We stress that the theorem applies for random vectors X which might even dependend of the martingale difference sequence $(Z_i)_{i=1}^n$.

Applying the above lemma with $Z_t \leftrightarrow \zeta_t$, and $X \leftrightarrow x^*$ we obtain,

$$\mathbf{E}\sum_{t=1}^T \zeta_t \cdot x^* \leq \frac{D}{2} \sqrt{\sum_{t=1}^T \mathbf{E}\|\zeta_t\|_*^2} \leq \frac{D}{2}\sqrt{4G^2 T} = DG\sqrt{T} \,,$$

where we used $\zeta_t = \tilde{g}_t - g_t$, together with the bound on the dual norms of $\tilde{g}_t, g_t$. Combining the above bound inside Eq. (26) concludes the proof.

∎

B.2. Proof of Theorem 7

Let us denote by $\tilde{g}_t$ the noisy estimates of $g_t := F(x_t)$, and by $\tilde{M}_t$ the noisy estimates of $M_t := F(y_{t-1})$.

Recall Equation (15) from the proof of Theorem 4 which states (see Section A.3),

$$\|g_t - M_t\|_* \cdot \|x_t - y_t\| \leq \frac{L}{2}\|x_t - y_{t-1}\|^2 + \frac{L}{2}\|x_t - y_t\|^2 \,.$$

Using the above together with the triangle inequality we get,

$$\|\tilde{g}_t - \tilde{M}_t\|_* \cdot \|x_t - y_t\| \leq \|\xi_t\|_* \cdot \|x_t - y_t\| + \frac{L}{2}\|x_t - y_{t-1}\|^2 + \frac{L}{2}\|x_t - y_t\|^2 \,,$$

where we define,

$$\xi_t := \tilde{g}_t - g_t - (\tilde{M}_t - M_t) \,.$$

Now, following the exact same analysis as in the proof of Theorem 4 (see Section A.3) shows the following applies to any $x \in \mathcal{K}$ when using Alg. 1,

$$\sum_{t=1}^T \tilde{g}_t \cdot (x_t - x) \leq O(\alpha DG + \alpha^2 LD^2 + LD^2 \log(LD/G_0)) + \underbrace{\Big[\sum_{t=1}^T \|\xi_t\|_* \cdot \|x_t - y_t\|\Big]}_{(D)} \,. \quad (27)$$

Now similarly to the proof of Theorem 6 we can show the following to hold for $x^* := \arg\max_{x \in \mathcal{K}} \Delta(\bar{x}_t, x^*)$ (see Eq. (25) and (26)),

$$
\begin{aligned}
T \cdot \mathbf{E}\Delta(\bar{x}_t, x^*) & \\
\leq \mathbf{E} \sum_{t=1}^{T} \tilde{g}_t \cdot (x_t - x) &- \mathbf{E} \sum_{t=1}^{T} \zeta_t \cdot (x_t - x) \\
\leq O(\alpha DG + \alpha^2 LD^2 + LD^2 \log(LD/G_0)) &+ \underbrace{\mathbf{E}\left[\sum_{t=1}^{T} \|\xi_t\|_* \cdot \|x_t - y_t\|\right]}_{(D)} + \underbrace{\mathbf{E} \sum_{t=1}^{T} \zeta_t \cdot x^*}_{(E)},
\end{aligned}
$$
(28)

where we define $\zeta_t := \tilde{g}_g - g_t$. Next, we will show that both (D) and (E) above are bounded by $O(\sigma D\sqrt{T \log T})$. This will conclude the proof.

Bounding (D): Using Cauchy-Schwarz we have,

$$
\sum_{t=1}^{T} \|\xi_t\|_* \cdot \|x_t - y_t\| \leq \sqrt{\sum_{t=1}^{T} \|\xi_t\|_*^2} \sqrt{\sum_{t=1}^{T} \|x_t - y_t\|^2}.
$$
(29)

Next we show that the sum in the second root is bounded by $O(\log T)$. Recalling the definition of Z_t we get,

$$
\begin{aligned}
\sum_{t=1}^{T} \|x_t - y_t\|^2 &\leq \sum_{t=1}^{T} \left(\|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2 \right) \\
&= 5 \sum_{t=1}^{T} \eta_t^2 Z_t^2 \\
&= 5D^2 \sum_{t=1}^{T} \frac{Z_t^2}{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2} \\
&\leq 30D^2 \max\{1, G^2/G_0^2\} + 10D^2 \log(1 + G^2 T/G_0^2),
\end{aligned}
$$
(30)

where we used $\eta_t = D/\sqrt{G_0^2 + \sum_{\tau=1}^{t-1} Z_\tau^2}$, we have also used $\forall t; \ Z_t \in [0, G]$ together with Lemma 10, which we remind below.

Lemma *[Lemma 10] For any non-negative real numbers $a_1, \ldots, a_n \in [0, a]$, and $a_0 \geq 0$,*

$$
\sum_{i=1}^{n} \frac{a_i}{a_0 + \sum_{j=1}^{i-1} a_j} \leq 2 + \frac{4a}{a_0} + 2\log\left(1 + \sum_{i=1}^{n-1} a_i/a_0\right).
$$

Thus, combining Eq. (30) inside Eq. (29), and taking expectation we conclude that,

$$
\begin{aligned}
(D) :&= \mathbf{E}\left[\sum_{t=1}^{T} \|\xi_t\|_* \cdot \|x_t - y_t\|\right] \leq 6D \max\{1, G_0/G_0\} \sqrt{1 + \log T} \cdot \mathbf{E}\sqrt{\sum_{t=1}^{T} \|\xi_t\|_*^2} \\
&\leq 12\alpha D\sigma \sqrt{T(1 + \log T)}.
\end{aligned}
$$
(31)

23

where we have used Jensen's inequality with respect to the $H(u) := \sqrt{u}$, as well as $\mathbf{E}\|\xi_t\|_*^2 \leq 4\sigma^2$. We also used $\alpha := \max\{G/G_0, G_0/G\}$.

Bounding (E): Using Proposition 11 (see section B.1), and taking $Z_t \leftrightarrow \zeta_t$, and $X \leftrightarrow x^*$ we obtain,

$$\mathbf{E}\sum_{t=1}^{T} \zeta_t \cdot x^* \leq \frac{D}{2}\sqrt{\sum_{t=1}^{T}\mathbf{E}\|\zeta_t\|_*^2} \leq \frac{D}{2}\sqrt{\sigma^2 T} = D\sigma\sqrt{T} \,, \tag{32}$$

where we used $\zeta_t = \tilde{g}_t - g_t$, together with the bound on the variance of $\tilde{g}_t$.

Concluding: Using the bounds in Eq. (31) and (32), inside Eq. (28) gives,

$$
\begin{aligned}
T\cdot\mathbf{E}&\Delta(\bar{x}_t, x^*) \\
&\leq O(\alpha DG + \alpha^2 LD^2 + LD^2\log(LD/G_0)) + O(\alpha\sigma D\sqrt{T\log T}) \,,
\end{aligned}
\tag{33}
$$

which concludes the proof.

B.3. Proof of Theorem 11

Proof Let us denote by $\mathcal{R}^*$ the Fenchel dual of $\mathcal{R}$. Thus the 1-strong-convexity of $\mathcal{R}$ w.r.t. $\|\cdot\|$ implies that $\mathcal{R}^*$ is 1-smooth w.r.t. $\|\cdot\|_*$.

We also denote by $x_0 \in \mathcal{K}$ be the minimizer of $\mathcal{R}$ on $\mathcal{K}$, which we assume to be in the relative interior of $\mathcal{K}$, so that $\nabla\mathcal{R}(x_0) = 0$. Without loss of generality, we also assume that $\mathcal{R}(x_0) = 0$. This implies that $\mathcal{R}^*(0) := \max_{x\in\mathbb{R}^d}\{x \cdot 0 - \mathcal{R}(x)\} = 0$.

Now recall the Fenchel-Young inequality,

Lemma 12 (Fenchel-Young inequality) *Let $f : \mathcal{K} \mapsto \mathbb{R}$ be a convex function, and $f^* : \mathbb{R}^d \mapsto \mathbb{R}$ be its Fenchel dual, then*

$$f(x) + f^*(y) \geq x^\top y, \qquad \forall x \in dom(f), y \in dom(f^*) \,.$$

Using this inequality and taking $y = s\sum_{i=1}^{n} Z_i$ and $x = X$, we have,

$$\Big(\sum_{i=1}^{n} Z_i\Big)^\top X = \frac{1}{s}\Big(s\sum_{i=1}^{n} Z_i\Big)^\top X \leq \frac{1}{s}\Big(\underbrace{\mathcal{R}(X)}_{(A)} + \mathcal{R}^*\big(\underbrace{s\sum_{i=1}^{n} Z_i}_{(B)}\big)\Big) \,. \tag{34}$$

Next we bound the two terms in the above inequality.
Bounding (A): Using $\mathcal{R}(x_0) = 0$ together with the boundedness of $\mathcal{R}$, gives,

$$(A) := \mathcal{R}(X) = \mathcal{R}(X) - \mathcal{R}(x_0) \leq \frac{1}{2}D^2 \,.$$

Bounding (B): To bound this term we first require the following lemma that can be found in Juditsky and Nemirovski (2008); Kakade (2010) (for completeness we provide a proof in Appendix B.3.1),

Lemma 13 *Let $f : \mathbb{R}^d$ be a L-smooth function w.r.t. a norm $\|\cdot\|_*$, such that $f(0) = 0$. Also let $(M_i)_i$ be a martingale difference sequence. Then the following holds,*

$$\mathbf{E}f\Big(\sum_{i=1}^n M_i\Big) \leq \frac{L}{2}\sum_{i=1}^n \mathbf{E}\|M_i\|_*^2 \,.$$

Applying the above lemma with $f \leftrightarrow \mathcal{R}^*$, and $M_i \leftrightarrow sZ_i$ gives,

$$\mathbf{E}\mathcal{R}^*\Big(s\sum_{i=1}^n Z_i\Big) \leq \frac{s^2}{2}\sum_{i=1}^n \mathbf{E}\|Z_i\|_*^2 \,.$$

Combining the bounds on (A) and (B) inside Eq. (34) and taking expectation gives,

$$\mathbf{E}\Big[\Big(\sum_{i=1}^n Z_i\Big)^\top X\Big] \leq \frac{1}{s}\Big(\frac{1}{2}D^2 + \frac{s^2}{2}\sum_{i=1}^n \mathbf{E}\|Z_i\|_*^2\Big) = \frac{D^2}{2s} + \frac{s}{2}\sum_{i=1}^n \mathbf{E}\|Z_i\|_*^2 \,.$$

Taking $s = \dfrac{D}{\sqrt{\sum_{i=1}^n \mathbf{E}\|Z_i\|_*^2}}$, we get,

$$\mathbf{E}\Big[\Big(\sum_{i=1}^n Z_i\Big)^\top X\Big] \leq \frac{D}{2}\sqrt{\sum_{i=1}^n \mathbf{E}\|Z_i\|_*^2} \,,$$

which concludes the proof. ∎

B.3.1. PROOF OF LEMMA 13

Proof We will prove the lemma by induction over n. For the base case $n = 1$, we may use the smoothness of f to get,

$$f(Z_1) = f(Z_1) - f(0) \leq \nabla f(0)^\top Z_1 + \frac{L}{2}\|Z_1\|_*^2 \,.$$

Taking expectation and using $\mathbf{E}Z_1 = 0$ the lemma follows.

Now for the induction step, assume that $\mathbf{E}f(\sum_{i=1}^{n-1} Z_i) \leq \frac{L}{2}\sum_{i=1}^{n-1}\mathbf{E}\|Z_i\|_*^2$. Using the smoothness of f gives,

$$f(\sum_{i=1}^n Z_i) = f(Z_n + \sum_{i=1}^{n-1} Z_i)$$

$$\leq f(\sum_{i=1}^{n-1} Z_i) + \nabla f\left(\sum_{i=1}^{n-1} Z_i\right)^\top Z_n + \frac{L}{2}\|Z_n\|_*^2 \,.$$

Taking expectation and using $\mathbf{E}[Z_n|Z_1,\ldots,Z_{n-1}] = 0$, as well as the induction assumption establishes the lemma. ∎

Appendix C. Additional Proofs

C.1. Proof of Lemma 2

Proof For any $x^* \in \mathcal{K}$,

$$
\begin{aligned}
g_t \cdot (x_t - x^*) &= g_t \cdot (x_t - y_t) + g_t \cdot (y_t - x^*) \\
&= (g_t - M_t) \cdot (x_t - y_t) + M_t \cdot (x_t - y_t) + g_t \cdot (y_t - x^*) .
\end{aligned} \tag{35}
$$

Moreover, by Cauchy-Schwarz,

$$
(g_t - M_t) \cdot (x_t - y_t) \leq \|g_t - M_t\|_* \|x_t - y_t\| . \tag{36}
$$

Also, any update of the form $a^* = \arg\min_{a \in A} a \cdot z + \mathcal{D}_{\mathcal{R}}(a, c)$, satisfies for any $b \in A$,

$$
z \cdot (a^* - b) \leq \mathcal{D}_{\mathcal{R}}(b, c) - \mathcal{D}_{\mathcal{R}}(b, a^*) - \mathcal{D}_{\mathcal{R}}(a^*, c) .
$$

Combining this with the Optimistic OGD learning rule (Eq. (3)) gives, (taking $a^* \leftrightarrow x_t$, $b \leftrightarrow y_t$, $c \leftrightarrow y_{t-1}$, $z \leftrightarrow M_t$),

$$
M_t \cdot (x_t - y_t) \leq \frac{1}{\eta_t} \left(\mathcal{D}_{\mathcal{R}}(y_t, y_{t-1}) - \mathcal{D}_{\mathcal{R}}(y_t, x_t) - \mathcal{D}_{\mathcal{R}}(x_t, y_{t-1}) \right) , \tag{37}
$$

as well as (taking $a^* \leftrightarrow y_t$, $b \leftrightarrow x^*$, $c \leftrightarrow y_{t-1}$, $z \leftrightarrow g_t$),

$$
g_t \cdot (y_t - x^*) \leq \frac{1}{\eta_t} \left(\mathcal{D}_{\mathcal{R}}(x^*, y_{t-1}) - \mathcal{D}_{\mathcal{R}}(x^*, y_t) - \mathcal{D}_{\mathcal{R}}(y_t, y_{t-1}) \right) . \tag{38}
$$

Combining Equations (36)-(38) inside Eq. (35) we obtain,

$$
\begin{aligned}
& g_t \cdot (x_t - x^*) \\
& \leq \|g_t - M_t\|_* \|x_t - y_t\| + \frac{1}{\eta_t} \left(\mathcal{D}_{\mathcal{R}}(x^*, y_{t-1}) - \mathcal{D}_{\mathcal{R}}(x^*, y_t) - \mathcal{D}_{\mathcal{R}}(y_t, x_t) - \mathcal{D}_{\mathcal{R}}(x_t, y_{t-1}) \right) \\
& \leq \|g_t - M_t\|_* \|x_t - y_t\| + \frac{1}{\eta_t} \left(\mathcal{D}_{\mathcal{R}}(x^*, y_{t-1}) - \mathcal{D}_{\mathcal{R}}(x^*, y_t) - \frac{1}{2}\|x_t - y_t\|^2 - \frac{1}{2}\|x_t - y_{t-1}\|^2 \right) ,
\end{aligned}
$$

where the last line uses the 1-strong-convexity of $\mathcal{R}$, implying that $\forall x, y \in \mathcal{K}; \; \mathcal{D}_{\mathcal{R}}(x, y) \geq \frac{1}{2}\|x - y\|^2$. Summing over $t \in [T]$ we obtain that for any $x^* \in \mathcal{K}$,

$$
\sum_{t=1}^{T} g_t \cdot (x_t - x^*)
$$

$$
\leq \frac{1}{\eta_1} \mathcal{D}_{\mathcal{R}}(x^*, y_0) + \sum_{t=2}^{T} \mathcal{D}_{\mathcal{R}}(x^*, y_{t-1}) \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right)
$$

$$
+ \sum_{t=1}^{T} \|g_t - M_t\|_* \cdot \|x_t - y_t\| - \frac{1}{2} \sum_{t=1}^{T} \eta_t^{-1} \left(\|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2 \right)
$$

$$
\leq \frac{D^2}{\eta_1} + D^2 \sum_{t=2}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right)
$$

$$
+ \sum_{t=1}^{T} \|g_t - M_t\|_* \cdot \|x_t - y_t\| - \frac{1}{2} \sum_{t=1}^{T} \eta_t^{-1} \left(\|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2 \right)
$$

$$
\leq \frac{D^2}{\eta_1} + \frac{D^2}{\eta_T}
$$

$$
+ \sum_{t=1}^{T} \|g_t - M_t\|_* \cdot \|x_t - y_t\| - \frac{1}{2} \sum_{t=1}^{T} \eta_t^{-1} \left(\|x_t - y_t\|^2 + \|x_t - y_{t-1}\|^2 \right) ,
$$

where in the second inequality we use the fact that η_t is monotonically non-increasing and thus $\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \geq 0$, as well as $\forall x \in \mathcal{K}; \; \mathcal{D}_{\mathcal{R}}(x, y_0) \in [0, D^2]$. This concludes the proof. ∎

C.2. Proof of Lemma 9

Proof

First direction: Here we actually prove a stronger result which is the following,

$$
\sqrt{a_0 + \sum_{i=1}^{n} a_i} - \sqrt{a_0} \leq \sum_{i=1}^{n} \frac{a_i}{\sqrt{a_0 + \sum_{j=1}^{i-1} a_j}} . \tag{39}
$$

The above combined with $a_n \geq 0$ immediately implies the first part of Lemma 9.

We will prove this Eq. (39) by induction. The base case, $n = 1$, holds since in this case,

$$
\sum_{i=1}^{1} \frac{a_i}{\sqrt{a_0 + \sum_{j=1}^{i-1} a_j}} = \frac{a_1}{\sqrt{a_0}} \geq \sqrt{a_0 + a_1} - \sqrt{a_0} ,
$$

The above is equivalent to,

$$
(a_0 + a_1)^2 \geq a_0(a_0 + a_1) ,
$$

which is holds true since $a_1, a_0 \geq 0$.

For the induction step assume that the Eq. (39) holds for $n-1$ and let us show it holds for n. By the induction assumption,

$$\sum_{i=1}^{n} \frac{a_i}{\sqrt{a_0 + \sum_{j=1}^{i-1} a_j}} \geq \sqrt{a_0 + \sum_{i=1}^{n-1} a_i} - \sqrt{a_0} + \frac{a_n}{\sqrt{a_0 + \sum_{i=1}^{n-1} a_i}}$$

$$= \sqrt{Z - x} + \frac{x}{\sqrt{Z - x}} - \sqrt{a_0} \, ,$$

where we denote $x := a_n$ and $Z = a_0 + \sum_{i=1}^{n} a_i$ (note that $x < Z$). Thus, in order to prove the lemma it is sufficient to show that,

$$\sqrt{Z - x} + \frac{x}{\sqrt{Z - x}} \geq \sqrt{Z} \, .$$

Looking at the function $H(x) := \sqrt{Z - x} + \frac{x}{\sqrt{Z-x}}$ it is immediate to validate that $H(\cdot)$ is monotonically increasing for any $x \in [0, Z]$ (since its derivative is non-negative in this line segment) and therefore for any $x \in [0, Z]$ we have,

$$\sqrt{Z - x} + \frac{x}{\sqrt{Z - x}} = H(x) \geq H(0) = \sqrt{Z} \, .$$

This establishes Eq. (39) which in turn concludes the first part of the proof.

Second direction: For this part of the proof we will need the following lemma which we prove in Section C.3,

Lemma 14 *For any non-negative numbers $a_1, \ldots, a_n \in [0, a]$, the following holds:*

$$\sum_{i=1}^{n} \frac{a_i}{\sqrt{a + \sum_{j=1}^{i-1} a_j}} \leq 2\sqrt{a} + 2\sqrt{a + \sum_{i=1}^{n-1} a_i} \, .$$

Now let us divide into two cases. **Assume** $a \leq a_0$, in this case it is clear that,

$$\sum_{i=1}^{n} \frac{a_i}{\sqrt{a_0 + \sum_{j=1}^{i-1} a_j}} \leq \sum_{i=1}^{n} \frac{a_i}{\sqrt{a + \sum_{j=1}^{i-1} a_j}} \leq 2\sqrt{a} + 2\sqrt{a + \sum_{i=1}^{n-1} a_i} \leq 2\sqrt{a} + 2\sqrt{a_0 + \sum_{i=1}^{n-1} a_i} \, ,$$

where the second inequality holds by Lemma 14. We are therefore left to analyze the case **where** $a_0 \leq a$. In this case, let us denote the following,

$$N_0 = \min \left\{ i \in [n] : \sum_{j=1}^{i-1} a_j \geq a \right\} \, .$$

Next we divide the relevant sum according to N_0,

$$
\begin{aligned}
\sum_{i=1}^{n} \frac{a_i}{\sqrt{a_0 + \sum_{j=1}^{i-1} a_j}} &= \sum_{i=1}^{N_0-1} \frac{a_i}{\sqrt{a_0 + \sum_{j=1}^{i-1} a_j}} + \sum_{i=N_0}^{n} \frac{a_i}{\sqrt{a_0 + \sum_{j=1}^{i-1} a_j}} \\
&\leq \frac{1}{\sqrt{a_0}} \sum_{i=1}^{N_0-1} a_i + \sum_{i=N_0}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i-1} a_j}} \\
&\leq \frac{2a}{\sqrt{a_0}} + \sum_{i=N_0}^{n} \frac{a_i}{\sqrt{\frac{1}{2}a + \frac{1}{2}\sum_{j=1}^{i-1} a_j}} \\
&\leq \frac{2a}{\sqrt{a_0}} + \sqrt{2} \sum_{i=N_0}^{n} \frac{a_i}{\sqrt{a + \sum_{j=N_0}^{i-1} a_j}} \\
&\leq \frac{2a}{\sqrt{a_0}} + 2 \cdot \sqrt{2}\sqrt{a} + 2 \cdot \sqrt{2}\sqrt{a + \sum_{j=N_0}^{n-1} a_j} \\
&\leq \frac{2a}{\sqrt{a_0}} + 3\sqrt{a} + 3\sqrt{a_0 + \sum_{j=1}^{n-1} a_j} \, ,
\end{aligned}
$$

where we used $\sum_{j=1}^{N_0-2} a_j \leq a$, as well as $\forall i \geq N_0; \; \sum_{j=1}^{i-1} a_j \geq a$. Both follow by the definition of N_0. The last line uses $a \leq \sum_{j=1}^{N_0-1} a_j$, and the line before last (i.e., the fifth line) uses Lemma 14. This concludes the second part of the proof. ∎

C.3. Proof of Lemma 14

Proof We will require the following lemma from McMahan and Streeter (2010) (appear as Lemma 7 therein).

Lemma 15 *For any non-negative numbers $a_1, \dots, a_n$ the following holds:*

$$
\sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}} \leq 2\sqrt{\sum_{i=1}^{n} a_i} \, .
$$

Using the above lemma together with $\forall i;\ a_i \leq a$ we have,

$$\sum_{i=1}^{n} \frac{a_i}{\sqrt{a + \sum_{j=1}^{i-1} a_j}} \leq \sum_{i=1}^{n} \frac{a_i}{\sqrt{\sum_{j=1}^{i} a_j}}$$

$$\leq 2\sqrt{\sum_{i=1}^{n} a_i}$$

$$\leq 2\sqrt{\sum_{i=1}^{n-1} a_i + 2\sqrt{a}}$$

$$\leq 2\sqrt{a + \sum_{i=1}^{n-1} a_i + 2\sqrt{a}}\,,$$

where we used $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ which hold $\forall a, b \geq 0$. This concludes the second part of the proof. $\blacksquare$

C.4. Proof of Lemma 10

Proof Let us denote the following:

$$N_0 = \min\left\{ i \in [n] : \sum_{j=1}^{i-1} a_j \geq a \right\}.$$

Next we divide the relevant sum according to N_0,

$$\sum_{i=1}^{n} \frac{a_i}{a_0 + \sum_{j=1}^{i-1} a_j} = \sum_{i=1}^{N_0-1} \frac{a_i}{a_0 + \sum_{j=1}^{i-1} a_j} + \sum_{i=N_0}^{n} \frac{a_i}{a_0 + \sum_{j=1}^{i-1} a_j}$$

$$\leq \frac{1}{a_0} \sum_{i=1}^{N_0-1} a_i + \sum_{i=N_0}^{n} \frac{a_i}{\frac{1}{2}a_0 + \frac{1}{2}a + \frac{1}{2}\sum_{j=1}^{i-1} a_j}$$

$$\leq \frac{2a}{a_0} + 2\sum_{i=N_0}^{n} \frac{a_i}{a_0 + a_i + \sum_{j=N_0}^{i-1} a_j}$$

$$= \frac{2a}{a_0} + 2\sum_{i=N_0}^{n} \frac{a_i/a_0}{1 + \sum_{j=N_0}^{i} a_j/a_0}$$

$$\leq \frac{2a}{a_0} + 2 + 2\log\left(1 + \sum_{i=N_0}^{n} a_i/a_0\right)$$

$$\leq \frac{2a}{a_0} + 2 + 2\log\left(1 + \sum_{i=1}^{n} a_i/a_0\right)\,, \tag{40}$$

where we used $\sum_{j=1}^{N_0-2} a_j \leq a$, as well as $\forall i \geq N_0$; $\sum_{j=1}^{i-1} a_j \geq a$. Both follow by the definition of N_0. And the fourth line uses the following lemma which we borrow from Levy et al. (2018) (appears as Lemma **A.3** therein),

Lemma 16 *For any non-negative real numbers $b_1, \ldots, b_n$,*

$$\sum_{i=1}^{n} \frac{b_i}{1 + \sum_{j=1}^{i} b_j} \leq 1 + \log\left(1 + \sum_{i=1}^{n} b_i\right) .$$

Now notice that,

$$\log\left(1 + \sum_{i=1}^{n} a_i/a_0\right) = \log\left(1 + \sum_{i=1}^{n-1} a_i/a_0\right) + \log\left(1 + \frac{a_n/a_0}{1 + \sum_{i=1}^{n-1} a_i/a_0}\right)$$

$$\leq \log\left(1 + \sum_{i=1}^{n-1} a_i/a_0\right) + \frac{a_n/a_0}{1 + \sum_{i=1}^{n-1} a_i/a_0}$$

$$\leq a/a_0 + \log\left(1 + \sum_{i=1}^{n-1} a_i/a\right) , \tag{41}$$

where the third line uses $\forall x \geq 0$, $\log(1+x) \leq x$, and the last line uses $a_n \leq a$. Combining Equations (40) and (41) concludes the proof. ∎