Universal methods for variational inequalities: deterministic and stochastic cases

Anton Klimza^a, Alexander Gasnikov^{b,a,c}, Fedor Stonyakin^{a,d} and Mohammad Alkousa^a

^{a}*Moscow Institute of Physics and Technology, Department of Applied Mathematics and Computer Science, Institutsky lane 9, Dolgoprudny, 141701, Russian Federation*
^{b}*Innopolis University, Universitetskaya Str., Innopolis, 420500, Russian Federation*
^{c}*Institute for Information Transmission Problems RAS, Bolshoy Karetny per. 19, build.1, Moscow, 127051, Russian Federation*
^{d}*V. I. Vernadsky Crimean Federal University, 4 Academician Vernadsky Avenue, Simferopol, 295007, Republic of Crimea, Russian Federation*

ABSTRACT

In this paper, we propose universal proximal mirror methods to solve the variational inequality problem with Hölder-continuous operators in both deterministic and stochastic settings. The proposed methods automatically adapt not only to the oracle's noise (in the stochastic setting of the problem) but also to the Hölder continuity of the operator without having prior knowledge of either the problem class or the nature of the operator information. We analyzed the proposed algorithms in both deterministic and stochastic settings and obtained estimates for the required number of iterations to achieve a given quality of a solution to the variational inequality. We showed that, without knowing the Hölder exponent and Hölder constant of the operators, the proposed algorithms have the least possible in the worst-case sense complexity for the considered class of variational inequalities. We also compared the resulting stochastic algorithm with other popular optimizers for the task of image classification.

1. Introduction

Variational inequalities (VIs) often arise in a wide variety of mathematical problems, such as optimization, optimal control, partial differential equations, mechanics, and finance. For a comprehensive overview, please refer to Konnov (2007); Facchinei and Pang (2003); Elliott (1987); Nagurney (1993). Variational inequalities cover as a special case many optimization problems such as minimization problems, saddle point problems, and fixed point problem (see Examples 1, 2 and 3 below). They play a key role in solving equilibrium and complementarity problems Facchinei and Pang (2003); Harker and Pang (1990), and have many applications in smooth and non-smooth optimization problems Nesterov (2005), also in saddle point problems that are an important part of the machine learning research Arjovsky and Chintala (2017); Kniaz, Knyaz, Mizginov, Papazyan, Fomin and Grodzitsky (2021), generative adversarial networks Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville and Bengio (2020), supervised/unsupervised learning Bach, Jenatton, Mairal, Obozinski et al. (2012); Bach, Mairal and Ponce (2008); Esser, Zhang and Chan (2010); Joachims (2005); Xu, Neufeld, Larson and Schuurmans (2004), reinforcement learning Jin and Sidford (2020); Omidshafiei, Pazis, Amato, How and Vian (2017), adversarial training Madry, Makelov, Schmidt, Tsipras and Vladu (2017), and generative models Daskalakis, Ilyas, Syrgkanis and Zeng (2017); Goodfellow, Pouget-Abadie, Mirza, Xu, Warde-Farley, Ozair, Courville and Bengio (2014); Gidel, Berard, Vignoud, Vincent and Lacoste-Julien (2018).

Numerous researchers have dedicated their efforts to exploring theoretical aspects related to the existence and stability of solutions and constructing iterative methods for solving VIs. A significant contribution to the development of numerical methods for solving VIs was made in the 1970's, when the extragradient method was proposed in Korpelevich (1976). More recently, Nemirovski in his seminal work Nemirovski (2004) proposed a non-Euclidean variant of this method, called Mirror Prox algorithm, which can be applied to Lipschitz continuous operators. Different methods with similar complexity were also proposed in Auslender and Teboulle (2005); Gasnikov, Dvurechensky, Stonyakin and Titov (2019); Monteiro and Svaiter (2010); Nesterov (2007); Solodov and Svaiter (1999). Besides that, in Nesterov (2007), Nesterov proposed a method for variational inequalities with a bounded variation of the operator, i.e., with

✉ klimza.aa@phystech.edu (A. Klimza); gasnikov@yandex.ru (A. Gasnikov); fedyor@mail.ru (F. Stonyakin); mohammad.alkousa@phystech.edu (M. Alkousa)
ORCID(s): 0009-0001-6839-0194 (A. Klimza); 0000-0002-7386-039X (A. Gasnikov); 0000-0002-9250-4438 (F. Stonyakin); 0000-0001-5470-0182 (M. Alkousa)

a non-smooth operator. He also raised the question of whether it is possible to propose a method that automatically "adjusts to the actual level of smoothness of the current problem instance." One of the goals of this paper is to propose such an algorithm. There is also extensive literature on variations of extragradient method that avoid taking two steps or two gradient computations per iteration, and so on (see for example Hsieh, Iutzeler, Malick and Mertikopoulos (2019); Malitsky and Tam (2020)).

Furthermore, the rising significance of stochastic methods in large-scale computations has driven interest in stochastic extragradient methods. In Juditsky, Nemirovski and Tauvel (2011), the authors studied stochastic Mirror-Prox methods for VIs with compact convex feasible sets, and in Mishchenko, Kovalev, Shulgin, Richtárik and Malitsky (2020) it improved the stochastic Mirror-Prox method for VIs by using a single sample per iteration. In Hsieh et al. (2019), the authors showed stochastic variants of the single-call type extragradient methods. Yet, with a constant stepsize, the iterates of these methods only converge to a neighborhood of the solution set. Diminishing stepsizes do ensure convergence, but they empirically slow down the performance. For more information about stochastic VIs, see Beznosikov, Polyak, Gorbunov, Kovalev and Gasnikov (2023) and references therein.

Similarly to a more general class of optimization problems with so-called Hölder-continuous functions, which cover both smooth and non-smooth problems, and for which authors proposed so-called universal algorithms that do not use the information about the Hölder exponent and Hölder constant Kamzolov, Dvurechensky and Gasnikov (2021); Baimurzina, Gasnikov, Gasnikova, Dvurechensky, Ershov, Kubentaeva and Lagunovskaya (2019); Ghadimi, Lan and Zhang (2019); Gasnikov and Nesterov (2018); Nesterov (2015) (the main advantage of the universal method is that it adjusts itself to the smoothness of the problem and does not require input parameters), there are many works for VIs with a more general class of operators, called Hölder-continuous, which covers both cases of operators with bounded variation and Lipschitz-continuous operators. VIs with Hölder-continuous monotone operators were considered in Nemirovski (2004), where a special choice of the stepsize for the Mirror Prox algorithm led to the optimal complexity for this class of problems Nemirovskii and Yudin (1983). The authors of Dang and Lan (2015) consider variational inequalities with non-monotone Hölder-continuous operators. Both papers use the Hölder constant and exponent to define the stepsize of the proposed methods. In Bach and Levy (2019), for solving monotonic stochastic variational inequalities, authors proposed a universal method based on the proximal mirror method. Essentially, the standard proximal mirror method is used, in which the Lipshitz constant is proposed to be selected in a special way, similar to the method used in Adagrad. However, this method is not fully adaptive because, like Adagrad, the step selection strategy makes significant use of information about the dimension of the problem. A fully adaptive method for solving smooth stochastic monotonic variational inequalities was constructed (with minor reservations) in the work Iusem, Jofré, Oliveira and Thompson (2019) and in Stonyakin, Gasnikov, Dvurechensky, Titov and Alkousa (2022); Stonyakin, Tyurin, Gasnikov, Dvurechensky, Agafonov, Dvinskikh, Alkousa, Pasechnyuk, Artamonov and Piskunova (2021) for the inexact oracle.

Recently, in Rodomanov, Kavis, Wu, Antonakopoulos and Cevher (2024) the authors propose a universal gradient descent for stochastic convex optimization problems. In this paper, we propose an analog of this method for deterministic and stochastic variational inequalities. We propose new universal algorithms, called "Universal Mirror Proximal Method", for variational inequality problem (1), in deterministic and stochastic settings.

Our main contributions can be summarized as follows:

1.1. Contributions

- We propose universal algorithms for the variational inequality problem in both deterministic and stochastic settings. These algorithms automatically adapt not only to the oracle's noise (in the stochastic setting of the problem) but also to the Hölder continuity of the operator without having prior knowledge of either the problem class or the nature of the operator information.

- We analyze the proposed algorithm for the problems in the deterministic setting with Hölder continuous operators and conclude its optimal convergence rate.

- We analyze the proposed stochastic algorithm for the problems with Hölder continuous unbiased operators and conclude its optimal convergence rate.

- We conduct some numerical experiments to test the proposed stochastic algorithm and compare its performance with other algorithms, such as SGD and Adam for the task of image classification.

1.2. Paper Organization

The paper consists of an introduction and 3 main sections. In Sect. 2, we mentioned the statement of the considered problem, some examples that can be covered by the problem under consideration, and the connected fundamental concepts. In Sect. 3, we proposed new algorithms called "Universal Mirror Proximal Method (UMP)" to solve the variational inequality problem in deterministic and stochastic settings. We analyze the proposed algorithms for the problems with Hölder continuous operators (and unbiased when the setting of the problem is stochastic) and conclude its optimal convergence rate. Sect. 4 devoted to the numerical experiments, to test the proposed stochastic algorithm and show its performance by comparison with other algorithms such as Stochastic Gradient Descent (SGD) and Adam for the task of image classification.

2. Preliminaries and Problem Statement

We will use $\langle x, y \rangle := \sum_{i=1}^{n} x_i y_i$ to denote the standard inner product of $x = (x_1, \ldots, x_n), y = (y_1, \ldots, y_n) \in \mathbb{R}^n$, and $\|x\| = \sqrt{\langle x, x \rangle}$ to denote the Euclidean norm. The operator $\mathbb{E}[\cdot]$ denotes the full mathematical expectation. We also use the notation $|x|_+ := \max\{0, x\}$ for any $x \in \mathbb{R}$.

Let $Q \subseteq \mathbb{R}^n$ be a compact and convex subset, bounded by a constant $D \geq \max_{x,y \in Q} \|x - y\|$. In this paper, we consider the following Minty variational inequality

$$\text{Find} \quad x^* \in Q : \quad \langle g(x), x^* - x \rangle \leq 0 \quad \forall x \in Q, \tag{1}$$

where $g : Q \longrightarrow \mathbb{R}^n$ is a continuous, monotone operator, i.e.,

$$\langle g(x) - g(y), x - y \rangle \geq 0 \quad \forall x, y \in Q. \tag{2}$$

We called that the operator g satisfies the Hölder condition on Q (or g is a Hölder-continuous operator), if for some $\nu \in [0, 1]$ and $L_\nu \geq 0$, it holds the following inequality

$$\|g(x) - g(y)\| \leq L_\nu \|x - y\|^\nu \quad \forall x, y \in Q, \tag{3}$$

where we refer to ν as Hölder exponent and to L_ν as Hölder constant.

Under the assumption of continuity and monotonicity of the operator g, the problem (1) is equivalent to a Stampacchia Giannessi (1998) (or strong Nesterov (2007)) variational inequality, in which the goal is to find $x^* \in Q$ such that

$$\langle g(x^*), x^* - x \rangle \leq 0, \quad \forall x \in Q. \tag{4}$$

To emphasize the extensiveness of the problem (1) (or (4)), Let us consider three common use cases for VIs.

Example 1 (Minimization problem). *Let us consider the minimization problem*

$$\min_{x \in Q} f(x), \tag{5}$$

and assume that $g(x) = \nabla f(x)$, where $\nabla f(x)$ denotes to the (sub)gradient of f at x. Then, if f is convex, it can be proved that $x^* \in Q$ is a solution to (4) if and only if $x^* \in Q$ is a solution to (5).

Example 2 (Saddle point problem). *Let us consider the saddle point problem*

$$\min_{u \in Q_u} \max_{v \in Q_v} f(u, v), \tag{6}$$

and assume that $g(x) := g(u, v) = \left(\nabla_u f(u, v), -\nabla_v f(u, v) \right)^\top$, where $Q = Q_u \times Q_v$ with $Q_u \subseteq \mathbb{R}^{n_u}, Q_v \subseteq \mathbb{R}^{n_v}$. Then if f is convex in u and concave in v, it can be proved that $x^* \in Q$ is a solution to (4) if and only if $x^* = (u^*, v^*) \in Q$ is a solution to (6).

Example 3 (Fixed point problem). *Let us consider the fixed point problem*

$$\text{find} \ \ x^* \in Q \ \ \text{such that} \quad F(x^*) = x^*, \tag{7}$$

where $F : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ is an operator. By taking $g(x) = x - F(x)$, it can be proved that $x^* \in Q = \mathbb{R}^n$ is a solution to (4) if $g(x^*) = 0$, i.e., x^* is a solution to (7).

Following Nesterov (2007); Antonakopoulos, Belmega and Mertikopoulos (2020), to assess the quality of a candidate solution $\widehat{x}$, we use the following restricted gap (or merit) function

$$\text{Gap}(\widehat{x}) = \max_{u \in Q} \langle g(u), \widehat{x} - u \rangle. \tag{8}$$

Thus, our goal is to find an approximate solution to the problem, that is, a point $\widehat{x} \in Q$ such that the following inequality holds

$$\text{Gap}(\widehat{x}) = \max_{u \in Q} \langle g(u), \widehat{x} - u \rangle \le \varepsilon, \tag{9}$$

for some $\varepsilon > 0$.

As already mentioned, Nemirovski (2004) proposed Mirror Prox algorithm under the assumption of L_1-Lipschitz continuity of the operator, i.e., g satisfies (3) with $v = 1$ and L_1. This method has complexity $O\left(L_1 D^2 / \varepsilon\right)$, where D characterizes the diameter of the set Q and ε is the desired accuracy. By complexity we mean the number of iterations of an algorithm to find a point $\widehat{x} \in Q$ such that (9) holds. For the case of variational inequalities with bounded variation of the operator g, i.e., g satisfying (3) with $v = 0$ and L_0 Nesterov (2007) proposed a method with complexity $O\left(L_0^2 D^2 / \varepsilon^2\right)$. The method for variational inequalities with Hölder-continuous monotone operator Nemirovski (2004) has the complexity $O\left(\left(L_v / \varepsilon\right)^{2/(1+v)} D^2\right)$, which is optimal for for the case of $v = 1$ and for the case of $v = 0$ Nemirovskii and Yudin (1983); Ouyang and Xu (2021).

For the stochastic setting of the problem (1) (or (4)), we consider the following operator

$$g(x) = \mathbb{E}_{\xi}\left[g(x, \xi)\right], \tag{10}$$

where ξ is a random variable, and $g(\cdot, \xi) : Q \longrightarrow \mathbb{R}^n$ is a stochastic operator, which satisfies the following assumption.

Assumption 1 (Bounded variance). *For all ξ with any distribution, and $x \in Q$, the unbiased operator $g(\cdot, \xi)$ has uniformly bounded variance, i.e., it holds the following inequality*

$$\mathbb{E}_{\xi}\left[\|g(x, \xi) - g(x)\|^2\right] \le \sigma^2, \tag{11}$$

for some $\sigma > 0$.

3. Universal proximal mirror method

In this section, for problem (1), we introduce new algorithms called Universal Mirror Proximal (UMP) for both deterministic and stochastic settings. For the deterministic setting, the algorithm is listed below as Algorithm 1, whereas for the stochastic setting, the algorithm is listed as Algorithm 2. These algorithms automatically adapt not only to the oracle's noise (in the stochastic setting of the problem) but also to the Hölder continuity of the operator g without having prior knowledge of either the problem class or the nature of the operator information, i.e., without knowing neither the Hölder exponent v, nor the corresponding Hölder constant L_v.

3.1. Analysis for deterministic setting

In this subsection, we provide an analysis for Algorithm 1, when the setting of the problem (1) is deterministic.

Algorithm 1 Universal Mirror Proximal Method (UMP) (for the deterministic setting of the problem).

Inputs: $z_0 = \arg\min_{u \in Q} d(u)$, $L_0 = \|g(z_0)\|$, and diameter $D > 0$.

1: **for** $k = 0, 1, \dots$ **do**

2:      $w_k = \arg\min_{x \in Q} \left(\langle g(z_k), x - z_k \rangle + \frac{L_k}{2} \|z_k - x\|^2 \right),$

3:      $z_{k+1} = \arg\min_{x \in Q} \left(\langle g(w_k), x - w_k \rangle + \frac{L_k}{2} \|z_k - x\|^2 \right),$

4:      $L_{k+1} = L_k + \max\left\{ 0, \frac{2 \langle g(w_k), w_k - z_{k+1} \rangle - L_k \|z_k - z_{k+1}\|^2}{D^2 + \|z_k - z_{k+1}\|^2} \right\}.$

5: **end for**

At the first, let us prove the following lemma.

Lemma 1. *Let $Q \subseteq \mathbb{R}^n$ be a convex set and $g : Q \longrightarrow \mathbb{R}^n$ be an operator. The function $h_k : Q \longrightarrow \mathbb{R}$, which defined on Q as*

$$h_k(x) = \langle g(w), x - w \rangle + \frac{L_k}{2} \|z - x\|^2 \quad \forall x, w, z \in Q, \tag{12}$$

is L_k-strongly convex, for some $L_k > 0$.

Proof. Let $\alpha \in [0, 1]$ and $x, y \in Q$. We have

$$h_k(\alpha x + (1 - \alpha)y) = \langle g(w), \alpha x + (1 - \alpha)y - w \rangle + \frac{L_k}{2} \|\alpha x + (1 - \alpha)y - z\|^2$$

$$= \alpha \langle g(w), x - w \rangle + (1 - \alpha) \langle g(w), y - w \rangle + \frac{L_k}{2} \|\alpha x + (1 - \alpha)y - z\|^2$$

$$= \alpha h_k(x) + (1 - \alpha)h_k(y) + \frac{L_k}{2} \|\alpha x + (1 - \alpha)y - z\|^2 - \alpha \frac{L_k}{2} \|x - z\|^2 - (1 - \alpha)\frac{L_k}{2} \|y - z\|^2$$

$$= \alpha h_k(x) + (1 - \alpha)h_k(y) + L_k \frac{\alpha^2 - \alpha}{2} \|x\|^2 + L_k \frac{(1 - \alpha)^2 - (1 - \alpha)}{2} \|y\|^2 + L_k \alpha(1 - \alpha) \langle x, y \rangle$$

$$= \alpha h_k(x) + (1 - \alpha)h_k(y) - \frac{L_k \alpha(1 - \alpha)}{2} \|x - y\|^2.$$

Thus, we get the following inequality

$$h_k(\alpha x + (1 - \alpha)y) \leq \alpha h_k(x) + (1 - \alpha)h_k(y) - \frac{L_k \alpha(1 - \alpha)}{2} \|x - y\|^2.$$

This means that the function h_k is L_k-strongly convex. $\square$

Now, for z_k and w_k in Algorithm 1, for every $k \geq 0$, let us set

$$h_k(x) = \langle g(w_k), x - w_k \rangle + \frac{L_k}{2} \|z_k - x\|^2 \quad \forall x \in Q.$$

Then, from item 3 in Algorithm 1, we have

$$z_{k+1} = \arg\min_{x \in Q} h_k(x) \quad \text{for} \quad w = w_k, \ z = z_k, \tag{13}$$

and, by Lemma 1 and (13) we obtain the following inequality

$$h_k(y) \geq h_k(z_{k+1}) + \frac{L_k}{2} \|y - z_{k+1}\|^2, \quad \forall y \in Q. \tag{14}$$

Thus, for $w_k, z_k, \forall k \geq 0$ in Algorithm 1, from (14) for any $y \in Q$, we get

$$\langle g(w_k), y - w_k \rangle + \frac{L_k}{2} \|z_k - y\|^2 \geq \langle g(w_k), z_{k+1} - w_k \rangle + \frac{L_k}{2} \|z_k - z_{k+1}\|^2 + \frac{L_k}{2} \|z_{k+1} - y\|^2,$$

i.e.,

$$\langle g(w_k), w_k - y \rangle \leq \langle g(w_k), w_k - z_{k+1} \rangle - \frac{L_k}{2} \|z_k - z_{k+1}\|^2 - \frac{L_k}{2} \|z_{k+1} - y\|^2 + \frac{L_k}{2} \|z_k - y\|^2. \tag{15}$$

By adding $\frac{L_{k+1}}{2} \|z_{k+1} - y\|^2$ to the both sides of (15), we get

$$\langle g(w_k), w_k - y \rangle + \frac{L_{k+1}}{2} \|z_{k+1} - y\|^2 \leq \langle g(w_k), w_k - z_{k+1} \rangle - \frac{L_k}{2} \|z_k - z_{k+1}\|^2 + \frac{L_{k+1} - L_k}{2} \|z_{k+1} - y\|^2$$
$$+ \frac{L_k}{2} \|z_k - y\|^2. \tag{16}$$

Since $D \geq \max_{x,y \in Q} \|x - y\|$, and $L_{k+1} \geq L_k$ (see item 4 in Algorithm 1) we get

$$\langle g(w_k), w_k - y \rangle + \frac{L_{k+1}}{2} \|z_{k+1} - y\|^2 \leq \langle g(w_k), w_k - z_{k+1} \rangle - \frac{L_k}{2} \|z_k - z_{k+1}\|^2 + \left(L_{k+1} - L_k\right) \frac{D^2}{2}$$

$$+ \frac{L_k}{2} \|z_k - y\|^2.$$

We will look for L_{k+1} such that

$$\left(L_{k+1} - L_k \right) \frac{D^2}{2} = \left| \left\langle g(w_k), w_k - z_{k+1} \right\rangle - \frac{L_k}{2} \|z_k - z_{k+1}\|^2 \right|_+ . \tag{17}$$

Then, we get

$$\langle g(w_k), w_k - y \rangle + \frac{L_{k+1}}{2} \|z_{k+1} - y\|^2 \leq D^2 \left(L_{k+1} - L_k \right) + \frac{L_k}{2} \|z_k - y\|^2. \tag{18}$$

Through the telescopic sum, we get

$$\sum_{i=0}^{k} \langle g(w_i), w_i - y \rangle + \frac{L_{k+1}}{2} \|z_{k+1} - y\|^2 \leq D^2 \left(L_{k+1} - L_0 \right) + \frac{L_0}{2} \|z_0 - y\|^2.$$

Thus, we have

$$\frac{1}{k} \sum_{i=0}^{k} \langle g(w_i), w_i - y \rangle \leq \frac{D^2 L_{k+1}}{k}. \tag{19}$$

Let us set $\widehat{w} := \frac{1}{k} \sum_{i=0}^{k} w_i$, then we have

$$\text{Gap}(\widehat{w}) = \max_{y \in Q} \left\langle g(y), \widehat{w} - y \right\rangle = \max_{y \in Q} \frac{1}{k} \sum_{i=0}^{k} \left(\langle g(w_i), w_i - y \rangle + \langle g(y) - g(w_i), w_i - y \rangle \right). \tag{20}$$

Since g is a monotone operator, then $\langle g(y) - g(w_i), w_i - y \rangle \leq 0$, and as a result from (19) and (20), we get

$$\text{Gap}(\widehat{w}) \leq \frac{D^2 L_{k+1}}{k}.$$

We have obtained an estimate of the convergence rate of the method for a given choice of L_{k+1}.

Let us find an upper bound to L_{k+1} to formulate the convergence rate depending on k. Using (18) we could not get a sufficient upper bound for L_{k+1}, thus instead (17) we take a slightly different formula, which allows us to estimate the upper bound for L_{k+1} but we will get a slightly worse estimation for $\text{Gap}(\widehat{w})$. This formula has the following form

$$\left(L_{k+1} - L_k \right) \frac{D^2}{2} = \left| \left\langle g(w_k), w_k - z_{k+1} \right\rangle - \frac{L_{k+1}}{2} \|z_k - z_{k+1}\|^2 \right|_+ . \tag{21}$$

Thus, for any $y \in Q$, we get

$$\langle g(w_k), w_k - y \rangle + \frac{L_{k+1}}{2} \|z_{k+1} - y\|^2 \leq \left(D^2 + \|z_{k+1} - z_k\|^2 \right) \left(L_{k+1} - L_k \right) + \frac{L_k}{2} \|z_k - y\|^2$$

$$\leq 2 D^2 \left(L_{k+1} - L_k \right) + \frac{L_k}{2} \|z_k - y\|^2.$$

From this, in a similar way as in the previous, we get the following estimate

$$\text{Gap}(\widehat{w}) \leq \frac{2 D^2 L_{k+1}}{k}. \tag{22}$$

Lemma 2. *Assume that the operator g is monotone and Hölder continuous with constant L_ν for some $\nu \in (0, 1)$, given on a convex compact subset Q bounded by a constant D. Then for problem (1), by Algorithm 1, the following inequality holds*

$$L_{k+1} \leq \left(\frac{8k}{D^2} \right)^{\frac{1-\nu}{2}} L_\nu.$$

Proof. The recalculation of L_{k+1} in Algorithm 1 is performed using (21). If $\left\langle g(w_k), w_k - z_{k+1} \right\rangle - \frac{L_{k+1}}{2} \|z_k - z_{k+1}\|^2 < 0$ (see (21)), then we get $L_{k+1} = L_k$, otherwise

$$\left(L_{k+1} - L_k\right) \frac{D^2}{2} = \left\langle g(w_k), w_k - z_{k+1} \right\rangle - \frac{L_{k+1}}{2} \|z_k - z_{k+1}\|^2. \tag{23}$$

By using (3), we get

$$\left\langle g(w_k) - g(z_k), w_k - z_{k+1} \right\rangle \leq \left| \left\langle g(w_k) - g(z_k), w_k - z_{k+1} \right\rangle \right| \leq L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\|.$$

Thus,

$$\left\langle g(w_k), w_k - z_{k+1} \right\rangle \leq \left\langle g(z_k), w_k - z_{k+1} \right\rangle + L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\|. \tag{24}$$

From item 2 in Algorithm 1, we have

$$w_k = \arg \min_{x \in Q} h_k(x) \quad \text{for} \quad w = z = z_k, \tag{25}$$

and, by Lemma 1 and (25) we obtain the following inequality

$$h_k(y) \geq h_k(w_k) + \frac{L_k}{2} \|y - w_k\|^2, \quad \forall y \in Q. \tag{26}$$

Thus, for $w_k, z_k, \forall k \geq 0$ and $y = z_{k+1}$ in Algorithm 1, from (26), we get

$$\left\langle g(z_k), z_{k+1} - z_k \right\rangle + \frac{L_k}{2} \|z_k - z_{k+1}\|^2 \geq \left\langle g(z_k), w_k - z_k \right\rangle + \frac{L_k}{2} \|z_k - w_k\|^2 + \frac{L_k}{2} \|z_{k+1} - w_k\|^2,$$

i.e,

$$\left\langle g(z_k), w_k - z_{k+1} + z_k - w_k \right\rangle - \frac{L_k}{2} \|z_k - z_{k+1}\|^2 \leq \left\langle g(z_k), z_k - w_k \right\rangle - \frac{L_k}{2} \|z_k - w_k\|^2 - \frac{L_k}{2} \|z_{k+1} - w_k\|^2.$$

Thus, we get

$$\left\langle g(z_k), w_k - z_{k+1} \right\rangle \leq \frac{L_k}{2} \|z_k - z_{k+1}\|^2 - \frac{L_k}{2} \|z_k - w_k\|^2 - \frac{L_k}{2} \|z_{k+1} - w_k\|^2. \tag{27}$$

From (24) and (27), we find

$$\left\langle g(w_k), w_k - z_{k+1} \right\rangle \leq L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\| + \frac{L_k}{2} \|z_k - z_{k+1}\|^2 - \frac{L_k}{2} \|z_k - w_k\|^2 - \frac{L_k}{2} \|z_{k+1} - w_k\|^2. \tag{28}$$

Thus, from (23) and (28), we get the following

$$\left(L_{k+1} - L_k\right) \frac{D^2}{2} \leq L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\| + \frac{L_k - L_{k+1}}{2} \|z_k - z_{k+1}\|^2 - \frac{L_k}{2} \|z_k - w_k\|^2 - \frac{L_k}{2} \|z_{k+1} - w_k\|^2.$$

Thus,

$$\frac{L_{k+1} - L_k}{2} \left(D^2 + \|z_k - z_{k+1}\|^2 - \|z_k - w_k\|^2 - \|z_{k+1} - w_k\|^2\right) \leq L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\|$$
$$- \frac{L_{k+1}}{2} \left(\|z_k - w_k\|^2 + \|z_{k+1} - w_k\|^2\right).$$

Let's denote $R = D^2 + \|z_k - z_{k+1}\|^2 - \|z_k - w_k\|^2 - \|z_{k+1} - w_k\|^2$. Then, we have

$$\frac{L_{k+1} - L_k}{2} R \leq L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\| - \frac{L_{k+1}}{2} \left(\|z_k - w_k\|^2 + \|z_{k+1} - w_k\|^2\right). \tag{29}$$

Let us define the function $f(x, y) := L_\nu x^\nu y - \frac{1}{2} L_{k+1}(x^2 + y^2)$ on $\mathbb{R}^2$, which is concave for any $L_\nu > 0, L_{k+1} > 0$ and $\nu \in (0, 1)$. This function attains its maximal value at the point $(x, y) \in \mathbb{R}^2$, such that

$$\frac{\partial f}{\partial x} = \nu L_\nu x^{\nu-1} y - L_{k+1} x = 0, \quad \text{and} \quad \frac{\partial f}{\partial y} = L_\nu x^\nu - L_{k+1} y = 0.$$

From these equations, we get

$$y = \frac{L_{k+1} x^{2-\nu}}{\nu L_\nu} \quad \text{and} \quad y = \frac{L_\nu x^\nu}{L_{k+1}}.$$

Thus, we have

$$\frac{L_{k+1} x^{2-\nu}}{\nu L_\nu} = \frac{L_\nu x^\nu}{L_{k+1}} \implies x = \left(\frac{\sqrt{\nu} L_\nu}{L_{k+1}} \right)^{\frac{1}{1-\nu}}, \quad y = \left(\frac{\nu^{\nu/2} L_\nu}{L_{k+1}} \right)^{\frac{1}{1-\nu}},$$

and therefore we get

$$f(x, y) \le L_\nu \left(\frac{\sqrt{\nu} L_\nu}{L_{k+1}} \right)^{\frac{\nu}{1-\nu}} \left(\frac{\nu^{\nu/2} L_\nu}{L_{k+1}} \right)^{\frac{1}{1-\nu}} - \frac{1}{2} L_{k+1} \left(\left(\frac{\sqrt{\nu} L_\nu}{L_{k+1}} \right)^{\frac{2}{1-\nu}} + \left(\frac{\nu^{\nu/2} L_\nu}{L_{k+1}} \right)^{\frac{2}{1-\nu}} \right)$$

$$= \frac{\nu^{\frac{\nu}{1-\nu}} L_\nu^{\frac{2}{1-\nu}}}{L_{k+1}^{\frac{1+\nu}{1-\nu}}} \left(1 - \frac{\nu^{1/\nu}}{2} - \frac{1}{2} \right) = \frac{\nu^{\frac{\nu}{1-\nu}} \left(1 - \nu^{1/\nu} \right) L_\nu^{\frac{2}{1-\nu}}}{2 L_{k+1}^{\frac{1+\nu}{1-\nu}}}. \tag{30}$$

By setting $x := \|w_k - z_k\|$, $y := \|w_k - z_{k+1}\|$ (in $f(x, y)$) and $p = \frac{1+\nu}{1-\nu}$, then from (30) we find the following inequality

$$L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\| - \frac{L_{k+1}}{2} \left(\|z_k - w_k\|^2 + \|z_{k+1} - w_k\|^2 \right) \le \frac{\nu^{\frac{\nu}{1-\nu}} \left(1 - \nu^{1/\nu} \right) L_\nu^{p+1}}{2 L_{k+1}^p}.$$

Thus, from (29) we find

$$\frac{L_{k+1} - L_k}{2} R \le \frac{\nu^{\frac{\nu}{1-\nu}} \left(1 - \nu^{1/\nu} \right) L_\nu^{p+1}}{2 L_{k+1}^p},$$

i.e.,

$$(p+1) L_{k+1}^p \left(L_{k+1} - L_k \right) \le \frac{\nu^{\frac{\nu}{1-\nu}} \left(1 - \nu^{1/\nu} \right) (p+1) L_\nu^{p+1}}{R} := \alpha.$$

Note that, since $L_{k+1} \ge L_k$, we have

$$(p+1) L_{k+1}^p (L_{k+1} - L_k) \ge (p+1) \int_{L_k}^{L_{k+1}} t^p dt = L_{k+1}^{p+1} - L_k^{p+1}.$$

Then

$$L_{k+1}^{p+1} - L_k^{p+1} \le \alpha,$$

and from the telescopic sum, we get

$$L_{k+1}^{p+1} \le k\alpha + L_0^{p+1}. \tag{31}$$

For a sufficiently large k, we have $L_0^{p+1} \le k\alpha$, then from (31) with $p = \frac{1+\nu}{1-\nu}$, we get

$$L_{k+1} \le (2k\alpha)^{\frac{1}{p+1}} = \left(2k \frac{\nu^{\frac{\nu}{1-\nu}} \left(1 - \nu^{1/\nu} \right) (p+1) L_\nu^{p+1}}{R} \right)^{\frac{1}{p+1}} = k^{\frac{1-\nu}{2}} \frac{\nu^{\nu/2} \left(1 - \nu^{1/\nu} \right)^{\frac{1-\nu}{2}} L_\nu}{\left(\frac{1-\nu}{4} \right)^{\frac{1-\nu}{2}} R^{\frac{1-\nu}{2}}} \le \left(\frac{4k}{R} \right)^{\frac{1-\nu}{2}} L_\nu. \tag{32}$$

By the parallelogram identity $2\|a\|^2 + 2\|b\|^2 = \|a+b\|^2 + \|a-b\|^2$ with $a = z_k - w_k$, $b = z_{k+1} - w_k$, we find

$$R = D^2 + \|z_k - z_{k+1}\|^2 - \|z_k - w_k\|^2 - \|z_{k+1} - w_k\|^2 = D^2 + \frac{1}{2} \|z_k - z_{k+1}\|^2 - \frac{1}{2} \|z_k + z_{k+1} - 2w_k\|^2 \ge \frac{D^2}{2}. \tag{33}$$

Thus, from (32) and (33), we get the following upper bound of the L_{k+1}

$$L_{k+1} \leq \left(\frac{8k}{D^2} \right)^{\frac{1-v}{2}} L_v.$$

$\square$

By combining (22) with the result of Lemma 2, we have come to the following result, which indicates the convergence rate of the method in the deterministic case.

Theorem 1. *Assume that the operator g is monotone and Hölder continuous with constant L_v for some $v \in (0, 1)$, given on a convex compact subset Q bounded by a constant D. Then for problem* (1), *by Algorithm* 1, *it holds the following inequality*

$$\text{Gap}(\hat{w}) \leq \frac{16 L_v D^{1+v}}{(8k)^{\frac{1+v}{2}}},$$

where $\hat{w} = \frac{1}{k} \sum_{i=0}^{k} w_i$. Consequently, to achieve an ε-solution to (1), *i.e., $\text{Gap}(\hat{w}) \leq \varepsilon$, it needs to make*

$$k = O \left(\inf_{v \in [0,1]} \left(\frac{16 L_v}{\varepsilon} \right)^{\frac{2}{1+v}} \frac{D^2}{8} \right) = O \left(\inf_{v \in [0,1]} \left(\frac{L_v}{\varepsilon} \right)^{\frac{2}{1+v}} 2^{\frac{5-3v}{1+v}} D^2 \right).$$

oracle calls of Algorithm 1.

3.2. Analysis for stochastic setting

In this subsection, we provide an analysis for Algorithm 2, when the setting of the problem (1) is stochastic. Algorithm 2 is adaptive concerning the Hölder exponent v, Hölder constant L_v, and variance of the stochastic oracle σ. We mention that the adaptivity according to the parameter σ is somewhat different from the adaptivity according to the parameter L. The adaptivity according to σ can be understood in the sense that the σ absences in the steps of Algorithm 2, but it is not selected in the method as the selection of the parameter L (see item 4 in Algorithm 2).

Algorithm 2 Universal Mirror Proximal Method (UMP) (for the stochastic setting of the problem).

Inputs: $z_0 = \arg\min_{u \in Q} d(u)$, $L_0 = \|g(z_0, \xi_{z_0})\|$ for a generated random variable ξ_{z_0}, and diameter $D > 0$.

1: **for** $k = 0, 1, \dots$ **do**

2: $\quad w_k = \arg\min_{x \in Q} \left(\left\langle g(z_k, \xi_{z_k}), x - z_k \right\rangle + \frac{L_k}{2} \|z_k - x\|^2 \right),$

3: $\quad z_{k+1} = \arg\min_{x \in Q} \left(\left\langle g(w_k, \xi_{w_k}), x - w_k \right\rangle + \frac{L_k}{2} \|z_k - x\|^2 \right),$

4: $\quad L_{k+1} = L_k + \max \left\{ 0, \frac{2 \left\langle g(w_k, \xi_{w_k}), w_k - z_{k+1} \right\rangle - L_k \|z_k - z_{k+1}\|^2}{D^2 + \|z_k - z_{k+1}\|^2} \right\}.$

5: **end for**

We have $g(x) = \mathbb{E}_{\xi_x}[g(x, \xi_x)]$, but we can only get the recalculation of L_{k+1}, w_k and z_{k+1} using noisy operator $g(x, \xi_x)$.

Let us repeat all estimates from (12) to (16) for the operator $g(x, \xi_x)$ instead of $g(x)$ and obtain:

$$\left\langle g(w_k, \xi_{w_k}), w_k - y \right\rangle + \frac{L_{k+1}}{2} \|z_{k+1} - y\|^2 \leq \left\langle g(w_k, \xi_{w_k}), w_k - z_{k+1} \right\rangle - \frac{L_k}{2} \|z_k - z_{k+1}\|^2$$
$$+ \frac{L_{k+1} - L_k}{2} \|z_{k+1} - y\|^2 + \frac{L_k}{2} \|z_k - y\|^2.$$

We again look for L_{k+1} such that:

$$\frac{(L_{k+1} - L_k) D^2}{2} = \left| \left\langle g(w_k, \xi_{w_k}), w_k - z_{k+1} \right\rangle - \frac{L_{k+1}}{2} \|z_k - z_{k+1}\|^2 \right|_+.$$

Using this formula for L_{k+1} we get

$$\left\langle g(w_k, \xi_{w_k}), w_k - y \right\rangle + \frac{L_{k+1}}{2} \|z_{k+1} - y\|^2 \le 2D^2(L_{k+1} - L_k) + \frac{L_k}{2} \|z_k - y\|^2. \tag{34}$$

Since ξ_{w_k} is independent with w_k, then $\mathbb{E}_{\xi_{w_k}} \left[\left\langle g(w_k, \xi_{w_k}), w_k - y \right\rangle \right] = \langle g(w_k), w_k - y \rangle$. Thus, by applying the expected value to both parts of the inequality (34), we get

$$\langle g(w_k), w_k - y \rangle + \mathbb{E}_{\xi_{w_k}} \left[\frac{L_{k+1}}{2} \|z_{k+1} - y\|^2 \right] \le 2D^2 \left(\mathbb{E}_{\xi_{w_k}} \left[L_{k+1} \right] - L_k \right) + \frac{L_k}{2} \|z_k - y\|^2.$$

and

$$\mathbb{E}_{\xi_{w_{k-1}}} \left[\langle g(w_k), w_k - y \rangle \right] + \mathbb{E}_{\xi_{w_k}} \left[\frac{L_{k+1}}{2} \|z_{k+1} - y\|^2 \right] \le 2D^2 \left(\mathbb{E}_{\xi_{w_k}} \left[L_{k+1} \right] - \mathbb{E}_{\xi_{w_{k-1}}} \left[L_k \right] \right)$$

$$+ \mathbb{E}_{\xi_{w_{k-1}}} \left[\frac{L_k}{2} \|z_k - y\|^2 \right].$$

Using the telescopic sum, we get

$$\mathbb{E}_{\xi} \left[\frac{1}{k} \sum_{i=0}^{k} \langle g(w_i), w_i - y \rangle \right] \le \frac{2D^2 \mathbb{E}_{\xi} \left[L_{k+1} \right]}{k}. \tag{35}$$

For $\hat{w} = \frac{1}{k} \sum_{i=0}^{k} w_i$, we have

$$\text{Gap}(\hat{w}) = \max_{y \in Q} \left\langle g(y), \hat{w} - y \right\rangle = \max_{y \in Q} \frac{1}{k} \sum_{i=0}^{k} \left(\langle g(w_i), w_i - y \rangle + \langle g(y) - g(w_i), w_i - y \rangle \right). \tag{36}$$

Since g is a monotone operator, then $\langle g(y) - g(w_i), w_i - y \rangle \le 0$, and as a result from (35) and (36), we get

$$\mathbb{E}_{\xi} \left[\text{Gap}(\hat{w}) \right] \le \frac{2D^2 \mathbb{E}_{\xi} \left[L_{k+1} \right]}{k}. \tag{37}$$

Our goal now is finding an upper bound to $\mathbb{E}_{\xi} \left[L_{k+1} \right]$. To this goal, let us first prove the following lemma.

Lemma 3. *Given a sequence of positive numbers x_k, α_k, β_k. Let us assume that for any $k \ge 0$ there exists $p, q \ge 1$ such that at least one of the following two conditions is satisfied*

$$x_{k+1}^p \le \alpha_k + x_k^p, \quad and \quad x_{k+1}^q \le \beta_k + x_k^q.$$

Then

$$x_{k+1} \le \left(\sum_{i=0}^{k} \alpha_i \right)^{1/p} + \left(\sum_{i=0}^{k} \beta_i \right)^{1/q} + x_0. \tag{38}$$

Proof. Let us prove (38) by induction on $k \ge 0$.

- Let $x_1^p \le \alpha_0 + x_0^p$, then $x_1 \le \left(\alpha_0 + x_0^p \right)^{1/p} \le \alpha_0^{1/p} + \beta_0^{1/q} + x_0$.

- Let $x_1^q \le \beta_0 + x_0^q$, then $x_1 \le \left(\beta_0 + x_0^q \right)^{1/q} \le \alpha_0^{1/p} + \beta_0^{1/q} + x_0$.

Let us assume that $x_k \le \left(\sum_{i=0}^{k-1} \alpha_i \right)^{1/p} + \left(\sum_{i=0}^{k-1} \beta_i \right)^{1/q} + x_0$. Then,

If $x_{k+1}^p \leq \alpha_k + x_k^p$, then we get

$$x_{k+1} \leq \left(\alpha_k + \left[\left(\sum_{i=0}^{k-1} \alpha_i \right)^{1/p} + \left(\sum_{i=0}^{k-1} \beta_i \right)^{1/q} + x_0 \right]^p \right)^{1/p}. \tag{39}$$

Let us consider the vectors $(\alpha_k, 0, 0)$ and $\left(\sum_{i=0}^{k-1} \alpha_i, \left(\sum_{i=0}^{k-1} \beta_i \right)^{p/q}, x_0^p \right)$. Then, by the inverse triangle inequality for Minkowski space, we have

$$\alpha_k + \left[\left(\sum_{i=0}^{k-1} \alpha_i \right)^{1/p} + \left(\sum_{i=0}^{k-1} \beta_i \right)^{1/q} + x_0 \right]^p = \left(\alpha_k^{1/p} + 0^{1/p} + 0^{1/p} \right)^p + \left[\left(\sum_{i=0}^{k-1} \alpha_i \right)^{1/p} + \left(\sum_{i=0}^{k-1} \beta_i \right)^{p/q \cdot 1/p} + \left(x_0^p \right)^{1/p} \right]^p$$

$$\leq \left[\left(\alpha_k + \sum_{i=0}^{k-1} \alpha_i \right)^{1/p} + \left(\sum_{i=0}^{k-1} \beta_i \right)^{p/q \cdot 1/p} + \left(x_0^p \right)^{1/p} \right]^p.$$

Thus, from (39), we find

$$x_{k+1} \leq \left(\sum_{i=0}^{k} \alpha_i \right)^{1/p} + \left(0 + \sum_{i=0}^{k-1} \beta_i \right)^{1/q} + x_0 \leq \left(\sum_{i=0}^{k} \alpha_i \right)^{1/p} + \left(\sum_{i=0}^{k} \beta_i \right)^{1/q} + x_0.$$

If $x_{k+1}^q \leq \beta_k + x_k^q$, then we get

$$x_{k+1} \leq \left(\beta_k + \left[\left(\sum_{i=0}^{k-1} \alpha_i \right)^{1/p} + \left(\sum_{i=0}^{k-1} \beta_i \right)^{1/q} + x_0 \right]^q \right)^{1/q}. \tag{40}$$

Let us consider the vectors $(0, \beta_k, 0)$ and $\left(\left(\sum_{i=0}^{k-1} \alpha_i \right)^{q/p}, \sum_{i=0}^{k-1} \beta_i, x_0^q \right)$. Then, by the inverse triangle inequality for Minkowski space, we have

$$\beta_k + \left[\left(\sum_{i=0}^{k-1} \alpha_i \right)^{1/p} + \left(\sum_{i=0}^{k-1} \beta_i \right)^{1/q} + x_0 \right]^q = \left(0^{1/q} + \beta_k^{1/q} + 0^{1/q} \right)^q + \left[\left(\sum_{i=0}^{k-1} \alpha_i \right)^{q/p \cdot 1/q} + \left(\sum_{i=0}^{k-1} \beta_i \right)^{1/q} + \left(x_0^q \right)^{1/q} \right]^q$$

$$\leq \left[\left(\sum_{i=0}^{k-1} \alpha_i \right)^{q/p \cdot 1/q} + \left(\beta_k + \sum_{i=0}^{k-1} \beta_i \right)^{1/q} + \left(x_0^q \right)^{1/q} \right]^q.$$

Thus, from (40), we find

$$x_{k+1} \leq \left(0 + \sum_{i=0}^{k-1} \alpha_i \right)^{1/p} + \left(\sum_{i=0}^{k} \beta_i \right)^{1/q} + x_0 \leq \left(\sum_{i=0}^{k} \alpha_i \right)^{1/p} + \left(\sum_{i=0}^{k} \beta_i \right)^{1/q} + x_0.$$

$\square$

Now, let us prove the following lemma, which gives an upper bound to $\mathbb{E}_\xi \left[L_{k+1} \right]$.

Lemma 4. *Assume that the operator g is monotone and Hölder continuous with constant L_ν for some $\nu \in (0, 1)$, given on a convex compact subset Q bounded by a constant D. Then for problem* (1)*, in the stochastic setting under Assumption* 1*, by Algorithm* 2*, the following inequality holds*

$$\mathbb{E}_\xi \left[L_{k+1} \right] \leq 2 \left(\frac{8k}{D^2} \right)^{\frac{1-\nu}{2}} L_\nu + \frac{\sqrt{8k}}{D} \sigma + L_0.$$

Proof. If the right side is negative, then we get $L_{k+1} = L_k$, otherwise

$$
\frac{(L_{k+1} - L_k) D^2}{2} = \left\langle g(w_k, \xi_{w_k}), w_k - z_{k+1} \right\rangle - \frac{L_{k+1}}{2} \|z_k - z_{k+1}\|^2
$$

$$
= \left\langle g(w_k, \xi_{w_k}) - g(w_k), w_k - z_{k+1} \right\rangle + \left\langle g(w_k), w_k - z_{k+1} \right\rangle - \frac{L_{k+1}}{2} \|z_k - z_{k+1}\|^2.
$$

Similar to Lemma 2 from (23) to (28), we obtain

$$
(L_{k+1} - L_k) \frac{R}{2} \le L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\| - \frac{L_{k+1}}{2} \left(\|z_k - w_k\|^2 + \|z_{k+1} - w_k\|^2 \right)
$$

$$
+ \left\langle g(w_k, \xi_{w_k}) - g(w_k), w_k - z_{k+1} \right\rangle.
$$

Let's denote $\delta_k := \|g(w_k, \xi_{w_k}) - g(w_k)\|$, then

$$
(L_{k+1} - L_k) \frac{R}{2} \le L_\nu \|w_k - z_k\|^\nu \|w_k - z_{k+1}\| - \frac{L_{k+1}}{2} \left(\|z_k - w_k\|^2 + \|z_{k+1} - w_k\|^2 \right) + \delta_k \|w_k - z_{k+1}\|.
$$

Let us define the function $f(x, y) := L_\nu x^\nu y + \delta_k y - \frac{L_{k+1}}{2} \left(x^2 + y^2 \right)$ on $\mathbb{R}^2$, which is concave for any $L_\nu > 0, L_{k+1} > 0, \delta_k$ and $\nu \in (0, 1)$. This function attains its maximal value at the point $(x, y) \in \mathbb{R}^2$, such that

$$
\frac{\partial f}{\partial x} = \nu L_\nu x^{\nu-1} y - L_{k+1} x = 0, \quad \text{and} \quad L_\nu x^\nu - L_{k+1} y + \delta_k = 0.
$$

From these equations, we get

$$
y = \frac{L_{k+1} x^{2-\nu}}{\nu L_\nu} \quad \text{and} \quad y = \frac{L_\nu x^\nu + \delta_k}{L_{k+1}}.
$$

Thus, we have

$$
\frac{L_{k+1} x^{2-\nu}}{\nu L_\nu} = \frac{L_\nu x^\nu}{L_{k+1}} + \frac{\delta_k}{L_{k+1}}.
$$

Let's substitute $x = C_k \left(\frac{\sqrt{\nu} L_\nu}{L_{k+1}} \right)^{\frac{1}{1-\nu}}$, then $y = C_k^{2-\nu} \left(\frac{\nu^{\nu/2} L_\nu}{L_{k+1}} \right)^{\frac{1}{1-\nu}}$ and it holds the following inequality

$$
\frac{L_{k+1} C_k^{2-\nu}}{\nu L_\nu} \left(\frac{L_{k+1}}{\nu^{\nu/2} L_\nu} \right)^{\frac{2-\nu}{1-\nu}} = \frac{L_\nu C_k^\nu}{L_{k+1}} \left(\frac{L_{k+1}}{\nu^{\nu/2} L_\nu} \right)^{\frac{\nu}{1-\nu}} + \frac{\delta_k}{L_{k+1}}.
$$

Thus,

$$
C_k^{2-\nu} - C_k^\nu - \frac{\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{\nu^{\nu/2} L_\nu} \right)^{\frac{1}{1-\nu}} = 0.
$$

Since $\delta_k \ge 0$, the maximum of the function $f(x, y)$ is achieved at $y \ge 0$, then $C_k \ge 0$.

Now let's substitute $x = C_k \left(\frac{\sqrt{\nu} L_\nu}{L_{k+1}} \right)^{\frac{1}{1-\nu}}$ and $y = C_k^{2-\nu} \left(\frac{\nu^{\nu/2} L_\nu}{L_{k+1}} \right)^{\frac{1}{1-\nu}}$ in $f(x, y)$, then we get

$$
f(x, y) = C_k^2 L_\nu \left(\frac{\sqrt{\nu} L_\nu}{L_{k+1}} \right)^{\frac{\nu}{1-\nu}} \left(\frac{\nu^{\nu/2} L_\nu}{L_{k+1}} \right)^{\frac{1}{1-\nu}} - \frac{L_{k+1}}{2} \left(C_k^2 \left(\frac{\sqrt{\nu} L_\nu}{L_{k+1}} \right)^{\frac{2}{1-\nu}} + C_k^{4-2\nu} \left(\frac{\nu^{\nu/2} L_\nu}{L_{k+1}} \right)^{\frac{2}{1-\nu}} \right)
$$

$$
+ \delta_k C_k^{2-\nu} \left(\frac{\nu^{\nu/2} L_\nu}{L_{k+1}} \right)^{\frac{1}{1-\nu}}
$$

$$
= \frac{\nu^{\frac{\nu}{1-\nu}} L_\nu^{\frac{2}{1-\nu}}}{L_{k+1}^{\frac{1+\nu}{1-\nu}}} C_k^2 \left(1 - \frac{\nu^{1/\nu}}{2} - \frac{C_k^{2-2\nu}}{2} \right) + \delta_k C_k^{2-\nu} \left(\frac{\nu^{\nu/2} L_\nu}{L_{k+1}} \right)^{\frac{1}{1-\nu}}
$$

$$= \frac{v^{\frac{v}{1-v}} L_v^{\frac{2}{1-v}}}{2L_{k+1}^{\frac{1+v}{1-v}}} C_k^2 \left(1 - v^{1/v} - \frac{\frac{\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{v^{v/2} L_v} \right)^{\frac{1}{1-v}}}{C_k^v} \right) + \delta_k C_k^{2-v} \left(\frac{v^{v/2} L_v}{L_{k+1}} \right)^{\frac{1}{1-v}}$$

$$= \frac{v^{\frac{v}{1-v}} L_v^{\frac{2}{1-v}} \left(1 - v^{1/v} \right)}{2L_{k+1}^{\frac{1+v}{1-v}}} C_k^2 + \frac{\delta_k C_k^{2-v}}{2} \left(\frac{v^{v/2} L_v}{L_{k+1}} \right)^{\frac{1}{1-v}}.$$

Let's consider two cases.

1. **First case:** Let $\frac{\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{v^{\frac{v}{2}} L_v} \right)^{\frac{1}{1-v}} \leq C_k^v$, then

$$0 = C_k^{2-v} - C_k^v - \frac{\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{v^{\frac{v}{2}} L_v} \right)^{\frac{1}{1-v}} \geq C_k^{2-v} - 2C_k^v.$$

Thus $C_k \leq 2^{\frac{1}{2-2v}}$. This means that

$$f(x,y) = \frac{v^{\frac{v}{1-v}} L_v^{\frac{2}{1-v}} \left(1 - v^{1/v} \right)}{2L_{k+1}^{\frac{1+v}{1-v}}} C_k^2 + \frac{\delta_k C_k^{2-v}}{2} \left(\frac{v^{v/2} L_v}{L_{k+1}} \right)^{\frac{1}{1-v}}$$

$$\leq \frac{(2v)^{\frac{v}{1-v}} L_v^{\frac{2}{1-v}} \left(1 - v^{1/v} \right)}{L_{k+1}^{\frac{1+v}{1-v}}} + 2^{\frac{v}{1-v}} L_{k+1} \left(\frac{v^{v/2} L_v}{L_{k+1}} \right)^{\frac{2}{1-v}}$$

$$= \frac{(2v)^{\frac{v}{1-v}} L_v^{\frac{2}{1-v}}}{L_{k+1}^{\frac{1+v}{1-v}}} \left(2 - v^{1/v} \right).$$

Let $p = \frac{1+v}{1-v}$, then, similarly to Lemma 2, $L_{k+1}^{p+1} \leq \alpha_k + L_k^{p+1}$, where

$$\alpha_k := \frac{2(p+1)(2v)^{\frac{v}{1-v}} L_v^{p+1}}{R} \left(2 - v^{1/v} \right).$$

2. **Second case:** Let $\frac{\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{v^{v/2} L_v} \right)^{1/(1-v)} \geq C_k^v$, then

$$0 = C_k^{2-v} - C_k^v - \frac{\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{v^{v/2} L_v} \right)^{1/(1-v)} \geq C_k^{2-v} - \frac{2\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{v^{v/2} L_v} \right)^{1/(1-v)},$$

therefore $C_k \leq \left(\frac{2\delta_k}{L_{k+1}} \left(\frac{L_{k+1}}{v^{v/2} L_v} \right)^{1/(1-v)} \right)^{1/(2-v)}$. This means that

$$f(x,y) = \frac{v^{v/(1-v)} L_v^{2/(1-v)} \left(1 - v^{1/v} \right)}{2L_{k+1}^{\frac{1+v}{1-v}}} C_k^2 + \frac{\delta_k C_k^{2-v}}{2} \left(\frac{v^{v/2} L_v}{L_{k+1}} \right)^{1/(1-v)}$$

$$\leq \frac{\delta_k C_k^{2-v} \left(2 - v^{1/v} \right)}{2} \left(\frac{v^{v/2} L_v}{L_{k+1}} \right)^{1/(1-v)}$$

$$\leq \frac{\delta_k^2 \left(2 - \nu^{1/\nu}\right)}{L_{k+1}}.$$

Let $q = 1$, then, similarly to Lemma 2, $L_{k+1}^{q+1} \leq \beta_k + L_k^{q+1}$, where

$$\beta_k := \frac{2\delta_k^2 \left(2 - \nu^{1/\nu}\right)}{R}.$$

By Lemma 3 we get

$$L_{k+1} \leq \left(k \frac{2(p+1)(2\nu)^{\frac{\nu}{1-\nu}} L_\nu^{p+1}}{R} \left(2 - \nu^{1/\nu}\right) \right)^{\frac{1}{p+1}} + \sqrt{\frac{2\left(2 - \nu^{1/\nu}\right)}{R} \sum_{i=1}^{k} \delta_i^2} + L_0$$

$$\leq 2^{\frac{2-\nu}{2}} k^{\frac{1-\nu}{2}} \frac{\nu^{\nu/2} L_\nu}{(1-\nu)^{\frac{1-\nu}{2}} R^{\frac{1-\nu}{2}}} \left(2 - \nu^{1/\nu}\right)^{\frac{1-\nu}{2}} + \frac{2}{\sqrt{R}} \sqrt{\sum_{i=1}^{k} \delta_i^2} + L_0$$

$$\leq 2 \left(\frac{4k}{R}\right)^{\frac{1-\nu}{2}} L_\nu + \frac{2}{\sqrt{R}} \sqrt{\sum_{i=1}^{k} \delta_i^2} + L_0.$$

According to Jensen's inequality, we have $\mathbb{E}_\xi \left[\left(\sum_{i=1}^{k} \delta_i^2 \right)^{1/2} \right] \leq \left(\sum_{i=1}^{k} \mathbb{E}_\xi \left[\delta_i^2\right] \right)^{1/2} = \sigma \sqrt{k}$. Thus, we get the following

$$\mathbb{E}_\xi \left[L_{k+1}\right] \leq 2 \left(\frac{8k}{D^2}\right)^{\frac{1-\nu}{2}} L_\nu + \frac{\sqrt{8k}}{D} \sigma + L_0.$$

$\square$

By combining (37) with the result of Lemma 4, we have come to the following result, which indicates the convergence rate of the method in the stochastic case.

Theorem 2. *Assume that the operator g, as in* (10), *is monotone and Hölder continuous with constant L_ν for some $\nu \in (0, 1)$, given on a convex compact subset Q bounded by a constant D. Then for problem* (1), *in the stochastic setting under Assumption 1, by Algorithm 2, it holds the following inequality*

$$\mathbb{E}_\xi \left[\text{Gap}(\hat{w})\right] \leq 32 \left(\frac{D^2}{8k}\right)^{\frac{1+\nu}{2}} L_\nu + \frac{32D}{\sqrt{8k}} \sigma + \frac{2D^2 L_0}{k},$$

where $\hat{w} = \frac{1}{k} \sum_{i=0}^{k} w_i$ and L_0 is an initial guess of the Hölder constant.

4. Numerical experiments

To compare the performance of Algorithm 2 (UMP) with other optimizers, let's train a convolutional neural network resnet18 implemented in the PyTorch library based on articles He, Zhang, Ren and Sun (2016) for image classification dataset CIFAR10 also provided by the PyTorch library.

CIFAR10 is a dataset of object samples from 10 classes (airplane, car, bird, cat, deer, dog, frog, horse, ship, and truck), containing 50,000 train and 10,000 test color pictures of size 32×32, each signed by the corresponding class.

As optimizers, we consider SGD, Adam, AdamW, UMP (Algorithm 2), and UMP without changing the variable L as a baseline. We select the optimal learning rate for each optimizer and our method uses it as initialization for variable L.

The most important graph is the comparison of the optimizers losses for training when training resnet18 on the CIFAR10 dataset. It shows the speed of the CrossEntropyLoss minimization which is the problem that the optimizers solve in the experiment. We see that compared to the baseline, changing the variable L in the UMP method improves the speed of the loss minimization considerably. We also see that the UMP achieves results similar to SGD for training and similar to Adam and AdamW for validation when training resnet18 on the CIFAR10 dataset.
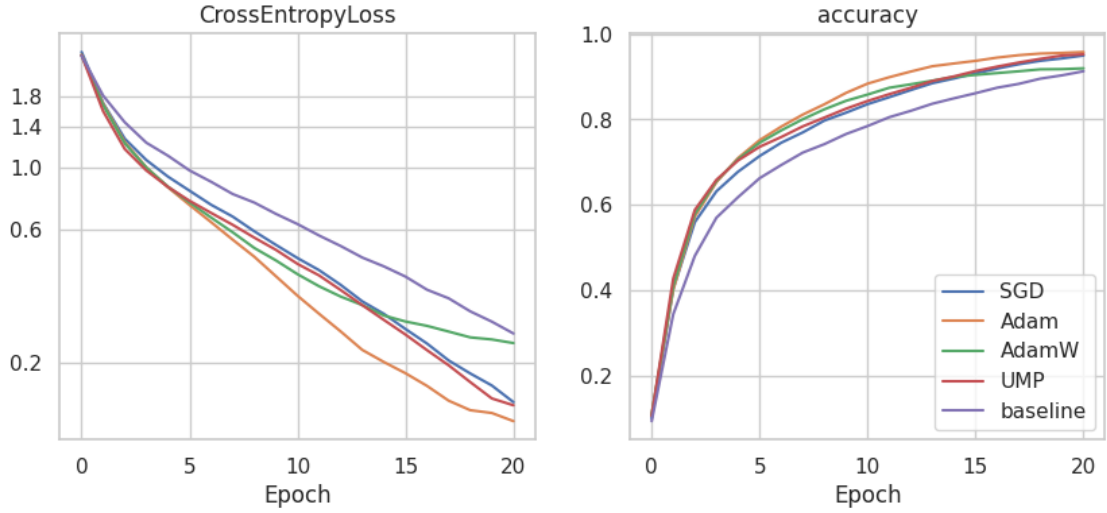
Figure 1: Comparison of the optimizers metrics for training when training resnet18 on the CIFAR10 dataset.
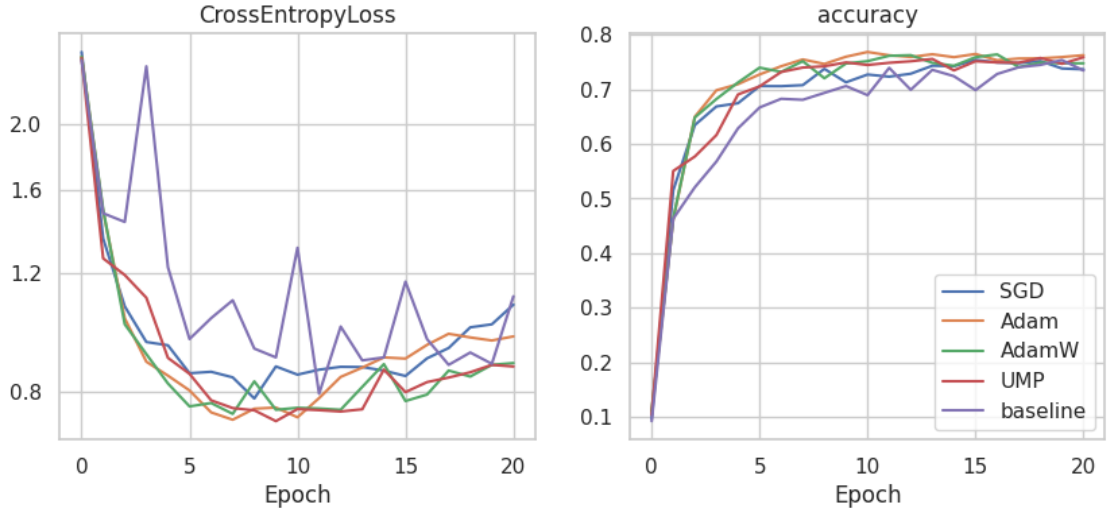

Figure 2: Comparison of the optimizers metrics for validation when training resnet18 on the CIFAR10 dataset.

5. Conclusion

In this paper, we show known improvements to the proximal mirror method Nemirovski (2004) and propose our universal version in both deterministic and stochastic settings of the problem, based on the ideas of universal gradient descent Rodomanov et al. (2024). We prove the theoretical convergence of the proposed universal proximal mirror methods and obtain an estimate of the convergence rate in the deterministic case

$$O\left(\inf_{\nu\in[0,1]}\left(\frac{L_\nu}{\varepsilon}\right)^{\frac{2}{1+\nu}}2^{\frac{5-3\nu}{1+\nu}}D^2\right).$$

In the stochastic case, we obtain the following estimate for the convergence rate

$$\mathbb{E}_\xi \left[\text{Gap}\left(\frac{1}{k} \sum_{i=0}^{k} w_i \right) \right] \le 32 \left(\frac{D^2}{8k} \right)^{\frac{1+\nu}{2}} L_\nu + \frac{32D}{\sqrt{8k}}\sigma + \frac{2D^2 L_0}{k}.$$

We also compare the stochastic method's performance with popular optimizers on the classical minimization problem. The experiment shows that the proposed method achieves similar results for loss minimization.

In the future, we plan to conduct more experiments with other models and compare the performance of the universal proximal mirror methods with different methods for saddle-point problems, namely the problem of training generative adversarial networks. It is also planned to study the behavior of the proposed methods with restarts to accelerate its convergence rate for variational inequalities with strongly monotone operators. Generalize the proposed algorithms and the conducted analysis to be convenient with any norm (i.e., with an arbitrary prox function and Bregman divergence) not necessarily Euclidean. Also, we plan to investigate adaptive batch size strategies (see eg. Lau, Liu and Kolar (2024)) to solve the variational inequality problem in the stochastic setting for more general operators.

Acknowledgements

References

Antonakopoulos, K., Belmega, E.V., Mertikopoulos, P., 2020. Adaptive extra-gradient methods for min-max optimization and games. arXiv preprint arXiv:2010.12100 .

Arjovsky, M., Chintala, S., 2017. Bottou. wasserstein gan. arXiv preprint arXiv:1701.07875 7.

Auslender, A., Teboulle, M., 2005. Interior projection-like methods for monotone variational inequalities. Mathematical programming 104, 39–68.

Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al., 2012. Optimization with sparsity-inducing penalties. Foundations and Trends® in Machine Learning 4, 1–106.

Bach, F., Levy, K.Y., 2019. A universal algorithm for variational inequalities adaptive to smoothness and noise, in: Conference on learning theory, PMLR. pp. 164–194.

Bach, F., Mairal, J., Ponce, J., 2008. Convex sparse matrix factorizations. arXiv preprint arXiv:0812.1869 .

Baimurzina, D., Gasnikov, A., Gasnikova, E., Dvurechensky, P., Ershov, E., Kubentaeva, M., Lagunovskaya, A., 2019. Universal method of searching for equilibria and stochastic equilibria in transportation networks. Computational Mathematics and Mathematical Physics 59, 19–33.

Beznosikov, A., Polyak, B., Gorbunov, E., Kovalev, D., Gasnikov, A., 2023. Smooth monotone stochastic variational inequalities and saddle point problems: A survey. European Mathematical Society Magazine , 15–28.

Dang, C.D., Lan, G., 2015. On the convergence properties of non-euclidean extragradient methods for variational inequalities with generalized monotone operators. Computational Optimization and applications 60, 277–310.

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

Elliott, C., 1987. Variational and quasivariational inequalities applications to free–boundary problems.(claudio baiocchi and antonio capelo). SIAM Review 29, 314.

Esser, E., Zhang, X., Chan, T.F., 2010. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. SIAM Journal on Imaging Sciences 3, 1015–1046.

Facchinei, F., Pang, J.S., 2003. Finite-dimensional variational inequalities and complementarity problems. Springer.

Gasnikov, A.V., Dvurechensky, P.E., Stonyakin, F.S., Titov, A.A., 2019. An adaptive proximal method for variational inequalities. Computational Mathematics and Mathematical Physics 59, 836–841.

Gasnikov, A.V., Nesterov, Y.E., 2018. Universal method for stochastic composite optimization problems. Computational Mathematics and Mathematical Physics 58, 48–64.

Ghadimi, S., Lan, G., Zhang, H., 2019. Generalized uniformly optimal methods for nonlinear programming. Journal of Scientific Computing 79, 1854–1881.

Giannessi, F., 1998. On minty variational principle. New Trends in Mathematical Programming: Homage to Steven Vajda , 93–99.

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

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

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2020. Generative adversarial networks. Communications of the ACM 63, 139–144.

Harker, P.T., Pang, J.S., 1990. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. Mathematical programming 48, 161–220.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Hsieh, Y.G., Iutzeler, F., Malick, J., Mertikopoulos, P., 2019. On the convergence of single-call stochastic extra-gradient methods. Advances in Neural Information Processing Systems 32.

Iusem, A.N., Jofré, A., Oliveira, R.I., Thompson, P., 2019. Variance-based extragradient methods with line search for stochastic variational inequalities. SIAM Journal on Optimization 29, 175–206.

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

Joachims, T., 2005. A support vector method for multivariate performance measures, in: Proceedings of the 22nd international conference on Machine learning, pp. 377–384.

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

Kamzolov, D., Dvurechensky, P., Gasnikov, A.V., 2021. Universal intermediate gradient method for convex problems with inexact oracle. Optimization Methods and Software 36, 1289–1316.

Kniaz, V.V., Knyaz, V.A., Mizginov, V., Papazyan, A., Fomin, N., Grodzitsky, L., 2021. Adversarial dataset augmentation using reinforcement learning and 3d modeling, in: Advances in Neural Computation, Machine Learning, and Cognitive Research IV: Selected Papers from the XXII International Conference on Neuroinformatics, October 12-16, 2020, Moscow, Russia, Springer. pp. 316–329.

Konnov, I., 2007. Equilibrium models and variational inequalities. Elsevier.

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

Lau, T.T.K., Liu, H., Kolar, M., 2024. Adadagrad: Adaptive batch size schemes for adaptive gradient methods. arXiv preprint arXiv:2402.11215 .

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A., 2017. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 .

Malitsky, Y., Tam, M.K., 2020. A forward-backward splitting method for monotone inclusions without cocoercivity. SIAM Journal on Optimization 30, 1451–1472.

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

Monteiro, R.D., Svaiter, B.F., 2010. On the complexity of the hybrid proximal extragradient method for the iterates and the ergodic mean. SIAM Journal on Optimization 20, 2755–2787.

Nagurney, A., 1993. Network Economics: A Variational Inequality Approach. Springer Science+Business Media New York.

Nemirovski, A., 2004. 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, 229–251.

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

Nesterov, Y., 2005. Smooth minimization of non-smooth functions. Mathematical programming 103, 127–152.

Nesterov, Y., 2007. Dual extrapolation and its applications to solving variational inequalities and related problems. Mathematical Programming 109, 319–344.

Nesterov, Y., 2015. Universal gradient methods for convex optimization problems. Mathematical Programming 152, 381–404.

Omidshafiei, S., Pazis, J., Amato, C., How, J.P., Vian, J., 2017. Deep decentralized multi-task multi-agent reinforcement learning under partial observability, in: International Conference on Machine Learning, PMLR. pp. 2681–2690.

Ouyang, Y., Xu, Y., 2021. Lower complexity bounds of first-order methods for convex-concave bilinear saddle-point problems. Mathematical Programming 185, 1–35.

Rodomanov, A., Kavis, A., Wu, Y., Antonakopoulos, K., Cevher, V., 2024. Universal gradient methods for stochastic convex optimization. arXiv preprint arXiv:2402.03210 .

Solodov, M.V., Svaiter, B.F., 1999. A hybrid approximate extragradient–proximal point algorithm using the enlargement of a maximal monotone operator. Set-Valued Analysis 7, 323–345.

Stonyakin, F., Gasnikov, A., Dvurechensky, P., Titov, A., Alkousa, M., 2022. Generalized mirror prox algorithm for monotone variational inequalities: Universality and inexact oracle. Journal of Optimization Theory and Applications 194, 988–1013.

Stonyakin, F., Tyurin, A., Gasnikov, A., Dvurechensky, P., Agafonov, A., Dvinskikh, D., Alkousa, M., Pasechnyuk, D., Artamonov, S., Piskunova, V., 2021. Inexact model: A framework for optimization and variational inequalities. Optimization Methods and Software 36, 1155–1201.

Xu, L., Neufeld, J., Larson, B., Schuurmans, D., 2004. Maximum margin clustering. Advances in neural information processing systems 17.