UNIVERSAL EXTRA-GRADIENT METHOD FOR STOCHASTIC VARIATIONAL INEQUALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the stochastic Extra-Gradient Method (EGM) for solving a stochastic variational inequality (VI) over a compact convex set. The main assumption is (δ, L)-continuity of the VI mapping, which includes Hölder mappings and their linear combinations. We develop new adaptive, line-search-free step-size rules for stochastic EGM. These rules can be interpreted as enhanced AdaGrad-type updates and require no prior knowledge of either the continuity modulus of the VI mapping or the variance of the stochastic oracle. The resulting methods are universal and attain the best complexity bound among the admissible continuity regimes. We prove convergence guarantees in terms of an expected residual function that upper bounds the relevant accuracy measures for several important problem classes, including the dual gap for monotone VIs, and the primal-dual gap for saddle-point and optimization problems. For Hölder-continuous mappings, our bounds recover the optimal dimension-independent complexity of first-order methods.

1 INTRODUCTION

Variational inequalities (VIs) provide a unified framework for a wide range of problems in optimization, games, and machine learning. Convex minimization, convex-concave saddle-point problems, monotone inclusions, and equilibrium problems in games can all be formulated as VIs; see, e.g., (Facchinei & Pang, 2003). This generality is especially relevant to modern machine learning, where adversarial and robust training (Goodfellow et al., 2020; Madry et al., 2018), multi-agent reinforcement learning, and constrained Markov decision processes (Jin & Sidford, 2020) naturally lead to saddle-point or game-theoretic formulations rather than the minimization of a single objective.

A VI is characterized by an underlying mapping, and methods that use the mapping evaluations only are attractive because they avoid Jacobian evaluations, which can be computationally expensive. However, simple projection methods may fail to converge even for monotone mappings. The extragradient method (EGM) (Korpelevich, 1976) overcomes this difficulty by evaluating the mapping at an intermediate point before performing the "main" update. This extra-gradient correction ensures convergence under suitable step-size choices and has led to several extensions, including the Mirror-Prox method in deterministic (Nemirovski, 2004) and stochastic settings (Juditsky et al., 2011).

A practical limitation of classical stochastic EGM is its dependence on problem parameters. Standard convergence guarantees typically require step sizes tuned to quantities such as the continuity modulus of the mapping, the Lipschitz or Hölder constants, the Hölder exponent, the diameter of the constraint set, or the variance of the stochastic oracle. In applications, these quantities are often unknown, expensive to estimate, or available only through conservative worst-case bounds. Moreover, a single problem may satisfy several continuity descriptions, and the most favorable one may not be apparent in advance. A step size tuned to the wrong regime can therefore lose much of the practical advantage promised by the theory.

Universal methods remove this dependence by adapting automatically to the regularity of the problem. In optimization, this idea goes back to Nesterov's universal gradient methods (Nesterov, 2015) and has since led to a broad literature on adaptive and line-search-based schemes for optimization (Yurtsever et al., 2015; Guminov et al., 2019; Kamzolov et al., 2021; Ito et al., 2023). For VIs, line-search variants have also been developed (Stonyakin et al., 2021; 2022), but they typically require additional mapping evaluations or repeated local tests. While such tests are suitable in deterministic problems,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

they become delicate in stochastic settings because noisy mapping evaluations can make acceptance conditions unreliable. Moreover, a line-search-free adaptive method is preferable when stochastic oracle calls are costly.

Apart from the line-search techniques, other adaptive algorithms for VIs have also been studied. One line of work develops Golden Ratio Algorithms for deterministic monotone VIs, including an adaptive variant for locally Lipschitz-continuous mappings (Malitsky, 2020). This variant estimates the local curvature from consecutive iterates and therefore requires neither prior knowledge of a Lipschitz constant nor a line search. Although its initial step size can be arbitrary in theory, this choice influences the subsequent step sizes in practice, and the original formulation also employs a prescribed upper bound on the step sizes. A method with a similar flavor was studied under Bregman continuity in (Antonakopoulos et al., 2019). It was shown to be universal by establishing rates for the bounded-variation and Lipschitz-continuous regimes. However, the method uses a "shrink ratio" hyperparameter, whose value can be sensitive in practice.

Another line of fully adaptive and universal methods uses AdaGrad- or AdaGrad-Norm-type updates. In (Antonakopoulos et al., 2021), universal methods for stochastic VIs were developed using variants of AdaGrad-Norm stepsizes. However, the selected stepsizes are always smaller than 1 because of suboptimal choices of the associated constants. Moreover, the analysis assumes that the stochastic mapping is uniformly bounded almost surely, which is a strong assumption. A similar bounded-mapping assumption was used in (Bach & Levy, 2019), where better constants allow the AdaGrad-Norm stepsizes to exceed 1. However, the method requires an arbitrary initial estimate of the mapping-norm bound, treated as a hyperparameter, as well as knowledge of the diameter of the underlying bounded constraint set. Its convergence rate also incurs an additional $\sqrt{\log T}$ factor in the stochastic term, making it suboptimal. Later work (Ene et al., 2021) developed an AdaGrad method for EGM that requires no hyperparameter other than an upper bound on the ℓ_∞-diameter of the underlying constraint set, which may be unknown in practice. Although this method is universal and addresses an important gap, its setting differs from ours since it uses coordinate-wise AdaGrad updates and an ℓ_∞-based analysis. Moreover, its stochastic rate depends on the problem dimension and incurs an additional $\sqrt{\log T}$ factor in the stochastic term. Inspired by work (Rodomanov et al., 2024b) in optimization, an attempt to close this gap by developing an adaptive, line-search-free, and universal stepsize for stochastic VIs was made in (Klimza et al., 2024). The work establishes rates for Hölder-continuous mappings but requires knowledge of the diameter of the underlying constraint set; some steps of its analysis appear to be incomplete, as discussed in Appendix A.

Universality has no single formal definition, and its meaning varies across the literature. In the stochastic VI literature, a method is often called universal if it attains optimal convergence rates for Lipschitz-continuous mappings and mappings with bounded variance, or, more generally, for Hölder-continuous mappings. A broader notion of universality is introduced in (Nesterov, 2025) for optimization through a Global Curvature Bound (GCB) that acts as a modulus of smoothness. Hölder-continuous functions and their linear combinations arise as special cases of this framework. Subsequently, (Doikov & Nesterov, 2025) extended GCB approach to universal higher-order methods for deterministic VIs using line search. The modulus of smoothness is similar in spirit to (δ, L)-smoothness in optimization (Rodomanov et al., 2024a). We extend this notion to mappings requiring that the constant L depends on δ, $\delta > 0$, (see Section 2). To the best of our knowledge, no universal method for stochastic VIs exists that avoids all problem parameters and hyperparameters, requires no line search, and adapts to a broader class of mappings encompassing several continuity regimes.

Contributions: Our main contributions are summarized below:

(1) We introduce the class of (δ, L)-continuous VI mappings, an approximate-continuity condition tailored to variational inequalities. This class includes Hölder-continuous mappings, linear combinations of mappings with different continuity regimes, and mappings arising from inexact computations. This framework provides a unified way to analyze stochastic VIs beyond a single prescribed class.

(2) We develop adaptive, line-search-free step-size rules for the stochastic Extra-Gradient Method. The proposed rules are derived from balance equations that control the key error terms in the EGM analysis. They can be viewed as enhanced AdaGrad-type updates and do not require prior knowledge of the continuity modulus, Hölder exponent, Lipschitz constant, stochastic-oracle variance, or the exact diameter of the constraint set.

(3) We prove a unified convergence guarantee for the proposed methods in terms of an expected residual function. This residual function upper bounds natural accuracy measures across several important settings, including the dual gap for monotone VIs, and the primal-dual gap for saddle-point and optimization problems. For Hölder-continuous mappings, our bounds recover the optimal dimension-independent complexity of first-order methods.

(4) We further develop distance-adaptive variants that reduce sensitivity to the initial diameter estimate. These variants preserve the same optimal convergence rates while improving the dependence of the constants on the quality of the initial estimate. Numerical experiments on optimization and min-max problems illustrate the practical effectiveness of the proposed adaptive rules.

Notation. For any vectors $x, y \in \mathbb{R}^n$, we denote $\langle x, y\rangle$ as their standard inner product. We denote the weighted Euclidean norm of a vector $x \in \mathbb{R}^n$ induced by a fixed positive definite matrix $B \in \mathbb{R}^{n\times n}$ as $\|x\| = \sqrt{\langle Bx, x\rangle}$. The dual norm is defined as $\|g\|_* = \max_{\|x\|\le 1}\langle g, x\rangle = \sqrt{\langle g, B^{-1}g\rangle}$. For the standard Euclidean norm, we have $B = I$, where I is the identity matrix. For a scalar $a \in \mathbb{R}$, we use the notation $[a]_+ = \max\{0, a\}$. We use $\mathbb{E}[\cdot]$ for the expectation mapping.

2 Problem Formulation

We consider the following Variational Inequality (VI) problem:

$$\text{Find } x^* \in X: \quad \langle F(x^*),\, x - x^*\rangle \ge 0 \qquad \forall x \in X, \tag{1}$$

where $X \subseteq \mathbb{R}^n$ is a nonempty closed convex set, and $F\colon X \to \mathbb{R}^n$ is a monotone mapping, i.e.,

$$\langle F(x) - F(y), x - y\rangle \ge 0 \qquad \forall x, y \in X. \tag{2}$$

We assume that the constraint set X is easy to project onto, so that projection onto X can be performed efficiently. We use the following assumption on the constraint set X.

Assumption 2.1. *The constraint set X in problem* (1) *is bounded, i.e., there exists a constant $D > 0$ such that* $\max_{x,y\in X}\|x - y\| \le D$.

A point x^* satisfying inequality (1) is usually called a *strong solution* of the VI corresponding to (X, F). In this regard, we would like to define also the concept of a weak (or Minty) solution:

$$\text{Find } x_* \in X: \quad \langle F(x),\, x - x_*\rangle \ge 0 \qquad \forall x \in X. \tag{3}$$

Assumption 2.1 ensures that the Minty solution exists by (Lassonde, 2015, Theorem 4); this result does not require continuity of F. If F is monotone, each strong solution (1) of the VI is also a weak solution. Conversely, if F is continuous, each weak solution is also a strong one (Minty, 1962), (Kinderlehrer & Stampacchia, 2000, Lemma 1.5, page 84). Therefore, when F is both monotone and continuous, the two solution concepts coincide.

2.1 (δ, L)-continuity of Mapping

We first state the notion of (δ, L)-smoothness in optimization, which has been studied in (Devolder et al., 2014; Nesterov, 2015; Rodomanov et al., 2024a; Savchuk et al., 2025), among others. Consider the optimization problem $\min_{x\in X_1} f(x)$, where f is convex and X_1 is a convex set. The function f is said to be (δ, L)-smooth over the set X_1 if

$$0 \le f(x) - (f(y) + \langle \nabla f(y), x - y\rangle) \le \frac{L}{2}\|x - y\|^2 + \delta \qquad \forall x, y \in X_1. \tag{4}$$

Work (Nesterov, 2025) generalized this definition using a modulus of smoothness. This notion is similar in spirit to (4), which can be imposed for every $\delta_1 > 0$, with some $L \equiv L(\delta_1)$ depending on δ_1. However for VIs, only mapping evaluations are available, and there is no corresponding notion of objective-function values. Therefore, the approximate-smoothness condition from optimization must be adapted to define approximate continuity for VI mappings. Previous works on VIs (Juditsky et al., 2011; Chakraborty & Nedić, 2025) consider mappings satisfying

$$\|F(x) - F(y)\|_* \le L\|x - y\| + \delta \qquad \forall x, y \in X.$$

In this work, we consider a reparameterized version of the preceding relation and say that F is (δ, L)-continuous on X if, for every $\delta > 0$, there exists $L(\delta) > 0$ such that

$$\|F(x) - F(y)\|_*^2 \leq L^2(\delta)\|x - y\|^2 + 4L(\delta)\delta \qquad \forall x, y \in X. \tag{5}$$

Definition (5) of (δ, L)-continuous mapping F is analogous to the one used in the optimization literature[1]. In what follows, we assume that F is (δ, L)-continuous mapping.

Assumption 2.2. *The mapping F is (δ, L)-continuous over the set X (cf. relation (5)).*

Assumption 2.2 may arise in settings where the mapping F is computed using an inexact oracle (see (Devolder et al., 2014) for examples in optimization). An important example of (δ, L)-continuous mapping is given by a Hölder-continuous mapping.

Lemma 2.3. *Let F be a Hölder continuous mapping, i.e., there exists $\nu \in [0, 1]$ and $H_\nu > 0$ such that*

$$\|F(x) - F(y)\|_* \leq H_\nu \|x - y\|^\nu \qquad \forall x, y \in X. \tag{6}$$

Then, F is (δ, L)-continuous $L(\delta) = \left(\frac{1-\nu}{4\delta}\right)^{\frac{1-\nu}{1+\nu}} \nu^{\frac{\nu}{1+\nu}} H_\nu^{\frac{2}{1+\nu}}$, with the convention that $0^0 = 1$.

The (δ, L)-continuity properties are preserved under the following calculus rules: (i) scalar multiplication, (ii) summation, and (iii) composition with an affine mapping (see Lemma C.2). In particular, the sum of Hölder mappings with different Hölder exponents is also (δ, L)-continuous, even though the resulting mapping may not be Hölder (Nesterov, 2025, Example 1). Thus, the class of (δ, L)-continuous mappings is wider than that of Hölder-continuous ones.

Remark 2.4. *In the context of smooth optimization, we have $F(\cdot) = \nabla f(\cdot)$, a different (δ, L) condition than that of relation (4) follows from Hölder-continuity of the gradient with $\tilde{L}(\delta) = \left(\frac{1-\nu}{2(1+\nu)\delta}\right)^{\frac{1-\nu}{1+\nu}} H_\nu^{\frac{2}{1+\nu}}$ (Rodomanov et al., 2024a, Remark F.1). For a Hölder-continuous mapping F, Lemma 2.3 derives $L(\delta)$ which satisfies $L(\delta) \leq \tilde{L}(\delta)$, as shown in Lemma C.1.*

2.2 Residual Function

Let $(x_t)_{t=1}^k$ be a sequence of iterates generated by an algorithm up to iteration $k \geq 1$, where $x_t \in X$ for all $1 \leq t \leq k$, and let $(\lambda_t)_{t=1}^k$ be a sequence of nonnegative weights. We introduce the accuracy certificate $(x_t, \lambda_t)_{t=1}^k$ as the main object of interest. Define $\Lambda_k = \sum_{t=1}^k \lambda_t$ and assume that $\Lambda_k > 0$ for all $k \geq 1$. We consider the residual function ϵ_k associated with the accuracy certificate $(x_t, \lambda_t)_{t=1}^k$ over the set X, defined as (Nesterov, 2009; Nemirovski et al., 2010; Rodomanov & Nesterov, 2023),

$$\epsilon_k \equiv \epsilon_k((x_t, \lambda_t)_{t=1}^k) := \max_{x \in X} \frac{1}{\Lambda_k} \sum_{t=1}^k \lambda_t \langle F(x_t), x_t - x \rangle. \tag{7}$$

In view of relation (3), we have $\epsilon_k \geq 0$. Our goal is to develop a universal and parameter-free algorithm that produces an accuracy certificate whose residual is bounded by a prescribed tolerance ε. Note that this residual is a weighted average of the well-known gap function (Facchinei & Pang, 2003, equation (1.5.15)) studied in the variational inequality literature. Hence, it serves as a meaningful measure for analyzing the convergence of an algorithm. Below, we provide several examples illustrating the applicability of this residual, similar to those studied in (Rodomanov & Nesterov, 2023).

i) **Convex minimization problem:** Let us consider the problem $f^* = \min_{x \in X} f(x)$, where $f : X \to \mathbb{R}$ is a differentiable convex function, X is a convex constraint set, and let $F = \nabla f$. Then, for $\widehat{x}_k = \frac{1}{\Lambda_k} \sum_{t=1}^k \lambda_t x_t$, we have

$$f(\widehat{x}_k) - f^* \leq \epsilon_k.$$

Therefore, establishing a convergence rate for the quantity ϵ_k directly implies the same convergence rate for the function values evaluated at the averaged iterates. Moreover, the problem admits the dual formulation $f^* = \max_{s \in \operatorname{dom} f_*} [-f_*(s) - \xi_X(-s)]$, where f_* is the conjugate of f and

[1] If $F = \nabla f$ and $X \subset X_1$, then relation (4) implies relation (5) (Devolder et al., 2014, Corollary 1).

$\xi_X(-s) := \max_{x\in X}\langle -s, x\rangle$. In this setting, ϵ_k is an upper bound for the corresponding primal-dual gap (Rodomanov & Nesterov, 2023), i.e., for all $k \geq 1$ and $\widehat{s}_k := \frac{1}{\Lambda_k}\sum_{t=1}^k \lambda_t \nabla f(x_t)$,

$$f(\widehat{x}_k) + f_*(\widehat{s}_k) + \xi_X(-\widehat{s}_k) \leq \epsilon_k.$$

ii) **Convex-concave min-max problem:** For convex and closed sets $U \subseteq \mathbb{R}^{n_1}$ and $V \subseteq \mathbb{R}^{n_2}$, consider the saddle-point problem $\min_{u\in U}\max_{v\in V} f(u,v)$, where the function $f : U \times V \to \mathbb{R}$ is convex in u for every fixed $v \in V$ and concave in v for every fixed $u \in U$. With $F(u,v) = (\nabla_u f(u,v),\ -\nabla_v f(u,v))$, and $\widehat{u}_k = \frac{1}{\Lambda_k}\sum_{t=1}^k \lambda_t u_t$ and $\widehat{v}_k = \frac{1}{\Lambda_k}\sum_{t=1}^k \lambda_t v_t$, we have

$$\max_{v\in V} f(\widehat{u}_k, v) - \min_{u\in U} f(u, \widehat{v}_k) \leq \epsilon_k.$$

iii) **Monotone Variational Inequality:** Consider VI problem (1), with a monotone F (cf. relation (2)), and its dual gap function (Auslender et al., 1974; Marcotte & Zhu, 1998; Nemirovski, 2004)

$$\bar{G}(y) = \max_{x\in X}\langle F(x), y - x\rangle.$$

Then, for $\widehat{x}_k = \frac{1}{\Lambda_k}\sum_{t=1}^k \lambda_t x_t$, we have for all $k \geq 1$,

$$\bar{G}(\widehat{x}_k) \leq \epsilon_k.$$

The dual gap function characterizes a weak (or Minty) solution (cf. relation (3)). It is always non-negative and takes zero value only when $y = x_*$ is a weak solution.

3 Algorithm for Stochastic Variational Inequalities

We consider a stochastic VI with the mapping $F(x) = \mathbb{E}\big[\widehat{F}(x,\omega)\big]$, $x \in X$, where ω is a random variable. We use the following assumption on the stochastic realizations of the mapping $\widehat{F}(x,\omega)$.

Assumption 3.1. *The mapping $\widehat{F}(x,\omega)$ is unbiased and, for some $\sigma > 0$, has a bounded variance*

$$\mathbb{E}\big[\widehat{F}(x,\omega)\big] = F(x) \quad \textit{and} \quad \mathbb{E}\big[\|\widehat{F}(x,\omega) - F(x)\|_*^2\big] \leq \sigma^2 \qquad \forall x \in X.$$

For solving stochastic VI problem (1), we consider the stochastic Extra-Gradient Method (EGM) (Korpelevich, 1976), with iterates given by for $k \geq 0$:

$$\begin{aligned} x_{k+1} &= \arg\min_{x\in X}\left\{\langle \widehat{F}(v_k,\omega_k),\ x - v_k\rangle + \frac{H_k}{2}\|x - v_k\|^2\right\}, \\ v_{k+1} &= \arg\min_{x\in X}\left\{\langle \widehat{F}(x_{k+1},\omega_{k+1}),\ x - x_{k+1}\rangle + \frac{H_k}{2}\|x - v_k\|^2\right\}, \end{aligned} \tag{8}$$

where $H_k \geq 0$ and $(\omega_k)_{k\geq 0}$ is a sequence of stochastic samples. The method starts with some random initial point $v_0 \in X$. Define

$$\begin{aligned} &Y_k = \widehat{F}(v_k,\omega_k) - F(v_k) \quad \text{and} \quad Z_k = \widehat{F}(x_k,\omega_k) - F(x_k) \qquad \forall k \geq 0, \\ &h_{k+1} = \arg\min_{x\in X}\left\{\langle -Z_{k+1}, x - h_k\rangle + \frac{\eta}{2}\|x - h_k\|^2\right\} \qquad \forall k \geq 0, \end{aligned} \tag{9}$$

for any $\eta \geq 0$. Then, under Assumption 2.1, for the method (8) we have for any $x \in X$ and $k \geq 0$ (cf. Lemma D.1) ,

$$\begin{aligned} \langle F(x_{k+1}), x_{k+1} - x\rangle \leq{}& \Delta_k(H_{k+1}) + \frac{3}{2}(H_{k+1} - H_k)D^2 + \langle Z_{k+1}, h_k - x_{k+1}\rangle + \frac{\|Z_{k+1}\|_*^2}{2\eta} \\ &+ \frac{\eta}{2}[\|x - h_k\|^2 - \|x - h_{k+1}\|^2] + \frac{H_k}{2}\|x - v_k\|^2 - \frac{H_{k+1}}{2}\|x - v_{k+1}\|^2, \end{aligned} \tag{10}$$

where $\Delta_k(H) \equiv \xi_k - \frac{H}{2}r_k^2$ with $r_k^2 = \|v_k - x_{k+1}\|^2 + \|v_{k+1} - x_{k+1}\|^2$ and

$$\begin{aligned} \xi_k &\equiv \langle \widehat{F}(x_{k+1},\omega_{k+1}), x_{k+1} - v_{k+1}\rangle - \frac{H_k}{2}\|v_k - v_{k+1}\|^2 + \frac{H_k}{2}r_k^2 \\ &\leq \langle \widehat{F}(x_{k+1},\omega_{k+1}) - \widehat{F}(v_k,\omega_k), x_{k+1} - v_{k+1}\rangle \equiv \xi_k'. \end{aligned}$$

The auxiliary sequence $(h_k)_{k\geq 0}$ is introduced to aid the analysis of the EGM iterates. The relation $\Delta_k(H_{k+1}) = \xi_k - \frac{H_{k+1}}{2} r_k^2$ is established using the optimality condition for the v_k-update and is valid for any sequence $(x_k)_{k\geq 0}$. The optimality conditions for both the v_k- and x_k-updates imply $\xi_k \leq \xi_k'$. Relation (10) is the key inequality guiding the choice of the sequence $(H_k)_{k\geq 0}$. Its left-hand side leads directly to the residual function in (7), which is the convergence measure used throughout the paper. On the right-hand side, the terms involving telescoping differences can be upper bounded by summing over k and using the boundedness of the set X (cf. Assumption 2.1). The terms involving stochastic noise Z_{k+1} are controlled using the unbiasedness and bounded variance of the sampled mapping (cf. Assumption 3.1). Thus, the only remaining term that requires an upper bound is the error term $\Delta_k(H_{k+1})$, which motivates the use of balance equations for updating H_{k+1}. In particular, by controlling either $\Delta_k(H_{k+1}) = \xi_k - \frac{H_{k+1}}{2} r_k^2$ or its upper bound $\xi_k' - \frac{H_{k+1}}{2} r_k^2$, we obtain two update rules for H_{k+1}, which we describe next.

We define the quantity Γ_k as the sum of the error term $\Delta_k(H_{k+1})$ and the term involving the diameter D on the right-hand side of relation (10), i.e., for all $k \geq 0$,

$$\Gamma_k = \Delta_k(H_{k+1}) + \frac{3}{2}(H_{k+1} - H_k) D^2. \tag{11}$$

Let $\widehat{D} > 0$ be an arbitrary estimate of the true diameter D. We update $(H_{k+1})_{k\geq 0}$ through a balance equation that controls Γ_k and yields, for all $k \geq 0$, the upper bound (cf. Lemma D.4)

$$\Gamma_k \leq \frac{\left[\min\{H_{k+1}, \bar{H}\} - H_k\right]_+}{2}(c_1 D^2 + c_2 \widehat{D}^2) + \frac{4L(\delta)\delta + \|Z_{k+1} - Y_k\|_*^2}{\bar{H}\widehat{D}^2}(3D^2 + \widehat{D}^2), \tag{12}$$

where $\bar{H} \geq \sqrt{2}\bar{L}(\delta)$, for $\delta > 0$, is a parameter that is a byproduct of the analysis and not required to implement the methods. We choose $\bar{H}$ optimally to obtain sharper convergence rates. Next, we present several examples that satisfy the preceding relation with different constants c_1, c_2, and $\bar{L}(\delta)$.

To control either $\Delta_k(H_{k+1})$ or its upper bound $\xi_k' - \frac{H_{k+1}}{2} r_k^2$, we choose H_{k+1} to satisfy one of the following balance equations: for all $k \geq 0$,

$$\frac{H_{k+1} - H_k}{2}\widehat{D}^2 = [\Delta_k(H_{k+1})]_+ \qquad \text{(Rule 1)}, \tag{13}$$

$$\frac{H_{k+1} - H_k}{2}\widehat{D}^2 = \left[\xi_k' - \frac{H_{k+1}}{2} r_k^2\right]_+ \qquad \text{(Rule 2)}. \tag{14}$$

The balance equations (13) and (14) yield relation (12) with constants $c_1 = 3$, $c_2 = 1$, $\bar{L}(\delta) = L(\delta)$. The explicit update and algorithmic details for Rules 1 and 2 are provided in Appendix D.3. In the context of optimization, a similar balance equation (14) appears with $\widehat{D}^2 = 2D^2$ in (Rodomanov et al., 2024b;a), where the knowledge of the set diameter D (cf. Assumption 2.1) is used. However, this requirement can be relaxed by modifying the analysis in (Rodomanov et al., 2024b;a), which allows the use of an arbitrary $\widehat{D} > 0$, instead. Along this line, an update rule simpler than the classic AdaGrad-norm can be obtained via the following balance equation: for all $k \geq 0$,

$$(H_{k+1}^2 - H_k^2)\widehat{D}^2 = \|\widehat{F}(x_{k+1}, \omega_{k+1}) - \widehat{F}(v_k, \omega_k)\|_*^2 \qquad \text{(Rule 3)}, \tag{15}$$

which also satisfies relation (12) with the constants $c_1 = 3$, $c_2 = 2$, and $\bar{L}(\delta) = L(\delta)\sqrt{1 + 3D^2/\widehat{D}^2}$.

4 Convergence Rate Analysis

In this section, a unified convergence rate analysis is provided for the iterates of the stochastic EGM in (8) with update rules given by balance equations (13)–(15).

Theorem 4.1. *Let Assumptions 2.1, 2.2, and 3.1 hold. Let $\widehat{D} > 0$ and $\delta > 0$ be arbitrary, and consider update Rules 1–3 (balance equations* (13)–(15)*). Then, the following relation holds for the expected residual function value*

$$\mathbb{E}[\epsilon_T] \leq \frac{\bar{L}(\delta)\bar{D}_1^2}{\sqrt{2}T} + \frac{2\sqrt{2L(\delta)\delta}}{\sqrt{T}}\left[\frac{\bar{D}_1\bar{D}_2}{\widehat{D}}\right] + \frac{\sigma}{\sqrt{T}}\left[D + \frac{2\bar{D}_1\bar{D}_2}{\widehat{D}}\right] \qquad \forall T \geq 1,$$

where ϵ_T is the residual function (cf. relation (7)), D is the diameter of the set X, $L(\delta)$ is the constant characterizing (δ, L)-continuity of F (cf. (5)), $\bar{D}_1^2 = c_1 D^2 + c_2 \widehat{D}^2$ and $\bar{D}_2^2 = 3D^2 + \widehat{D}^2$. The constants c_1, c_2, and $\bar{L}(\delta)$ depend on Rules 1–3, as follows: $c_1 = 3$, $c_2 = 1$, $\bar{L}(\delta) = L(\delta)$ for Rules 1 and 2, and $c_1 = 3$, $c_2 = 2$, $\bar{L}(\delta) = L(\delta)\sqrt{1 + 3D^2/\widehat{D}^2}$ for Rule 3.

Theorem 4.1 shows that the expected residual function is bounded by (δ, L)-continuity related term $\bar{L}(\delta)\bar{D}_1^2/T + \sqrt{L(\delta)\delta}\,\bar{D}_1\bar{D}_2/(\widehat{D}\sqrt{T})$ and the variance-based term $\frac{\sigma}{\sqrt{T}}\left[2\bar{D}_1\bar{D}_2/\widehat{D} + D\right]$, both decaying at the rate of $O(1/\sqrt{T})$. Our method is universal as it achieves the optimal rate for stochastic VIs under Assumption 2.2 without knowledge of any problem-dependent parameters. In particular, the error due to (δ, L)-continuity decays in the order of $O(1/\sqrt{T})$, rather than producing the non-vanishing $O(\delta)$ error floor observed in the optimization literature (Devolder et al., 2014; Rodomanov et al., 2024a), where a different approximate-smoothness condition is imposed on the objective function (cf. (4)). The bound in Theorem 4.1 can be optimized over $\delta > 0$, since it holds for any $\delta > 0$.

For a Hölder continuous mapping, using Lemma 2.3, we have the following corollary of Theorem 4.1.

Corollary 4.2. *Let Assumptions 2.1 and 3.1 hold, and let the mapping F be Hölder continuous (cf. relation (6)) for some $\nu \in [0, 1]$ and $H_\nu > 0$. Then, for any $\widehat{D} > 0$, the expected residual function (cf. relation (7)) satisfies the following inequality*

$$\mathbb{E}[\epsilon_T] \leq \frac{2^{\frac{1}{2}-\nu} H_\nu}{(1-\nu)^{\frac{1-\nu}{2}} \nu^{\frac{\nu}{2}}} \left[\frac{\bar{D}_2}{\widehat{D}}\right]^{1+\nu(a-1)} \left[\frac{\bar{D}_1^2}{T}\right]^{\frac{1+\nu}{2}} + \frac{\sigma}{\sqrt{T}}\left[\frac{2\bar{D}_1\bar{D}_2}{\widehat{D}} + D\right] \qquad \forall T \geq 1,$$

where $\bar{D}_1^2 = c_1 D^2 + c_2 \widehat{D}^2$ and $\bar{D}_2^2 = 3D^2 + \widehat{D}^2$, with the constants $c_1 = 3$, $c_2 = 1$, $a = 0$ for Rules 1 and 2, and $c_1 = 3$, $c_2 = 2$, $a = 1$ for Rule 3, and with the convention $0^0 = 1$.

Hence, the balance equations (13)–(15) prescribe rules that achieve the optimal convergence rate for Hölder-continuous mappings. The constant $\widehat{D}$ in both Theorem 4.1 and Corollary 4.2 may be chosen arbitrarily, although its relation to D affects the constants in the bound through the ratio $\bar{D}_2/\widehat{D} = \sqrt{1 + 3D^2/\widehat{D}^2}$.

5 Algorithm with Distance Adaptation

In this section, to obtain better constants in the convergence rate and eliminate the ratio $\bar{D}_2/\widehat{D}$ appearing in Theorem 4.1 and Corollary 4.2, we consider a modified EGM with distance adaptation. Specifically, an alternative to the use of a constant $\widehat{D} > 0$, the distance adaptation (Carmon & Hinder, 2022; Ivgi et al., 2023; Khaled et al., 2023) provides iterative estimates D_k of the set X diameter D. Starting with some initial $D_0 = \widehat{D} > 0$, the adaptive distance update is given by

$$D_{k+1} = \max\{D_k, \|v_{k+1} - v_0\|, r_k\} \qquad \forall k \geq 0, \tag{16}$$

where $r_k = \sqrt{\|v_k - x_{k+1}\|^2 + \|v_{k+1} - x_{k+1}\|^2}$.

In this case, the iterates of the Diatance-Adaptive EGM (DA EGM) are generated as: for all $k \geq 0$,

$$x_{k+1} = \arg\min_{x \in X} \left\{ D_k \langle \widehat{F}(v_k, \omega_k),\, x - v_k \rangle + \frac{H_k}{2}\|x - v_k\|^2 \right\},$$
$$v_{k+1} = \arg\min_{x \in X} \left\{ D_k \langle \widehat{F}(x_{k+1}, \omega_{k+1}),\, x - x_{k+1} \rangle + \frac{H_k}{2}\|x - v_k\|^2 \right\}.$$

starting with a random initial point $v_0 \in X$. At each iteration k, after performing the DA EGM updates, the diameter D_{k+1} is updated according to (16), followed by an appropriate error balance equation similar to those for Rules 1 and 2 (cf. relations (13) and (14)), but with an additional scaling factor of distances. We refer to such balance equations as "Methods 1 and 2" which is detailed in relations (53) and (54), later on in Appendix F.1. A detailed discussion about the algorithm, balance equations, and its related analysis is presented in Appendix F. Under Assumption 2.2, both Methods 1

and 2 are universal as they generate iterates such that for any $T \geq 1$,

$$\mathbb{E}\big[\epsilon_{\tau(T)}\big] \leq \rho^{1/T} \log(e\rho) \left[\frac{3\sqrt{2}L(\delta)D^2}{T} + \frac{12D\sqrt{L(\delta)\delta} + (6\sqrt{2}+5)\sigma D}{\sqrt{T}} \right] \quad \text{(Theorem F.3)},$$

where $\tau(T) := \underset{1\leq t\leq T}{\arg\min} \frac{D_t}{\sum_{k=0}^{t-1} D_k}$ and $\rho = \max\left\{1, \frac{D}{\widehat{D}}\right\}$. For Hölder continuous mapping, we have

$$\mathbb{E}\big[\epsilon_{\tau(T)}\big] \leq \rho^{1/T} \log(e\rho) \left[\frac{3 \times 2^{1-\nu/2} H_\nu D^{1+\nu}}{(1-\nu)^{(1-\nu)/2} T^{(1+\nu)/2}} + (6\sqrt{2}+5)\frac{\sigma D}{\sqrt{T}} \right] \quad \text{(Corollary F.4)},$$

with the convention $0^0 = 1$. In comparison to Theorem 4.1 and Corollary 4.2, the factor $D/\widehat{D}$ does not appear in the preceding results for distance-adaptive methods. If $\widehat{D} << D$, then the non-distance-adaptive bounds may contain a large factor proportional to $D/\widehat{D}$. In contrast, in the distance-adaptive methods the factor $\rho^{1/T}\log(e\rho)$ is milder, approaching $\log(e\rho)$ as T increases.

6 Experiments

We evaluate our algorithms on both optimization and min-max problems. For optimization, we consider (i) Hölder-smooth polyhedral feasibility, (ii) logistic regression, and (iii) softmax problems. We compare our methods EGM with Rules 1 and 2 (aka. Algorithm 1), EGM with Rule 3 (aka. Algorithm 2), and Distance-Adaptive EGM (DA EGM) Methods 1 and 2 (aka. Algorithm 3), with two variants of the Universal Gradient Method (UGM), corresponding to Equations (3.2) and (3.3) of Rodomanov et al. (2024a), which are denoted by UGM with AdaGrad and UGM, respectively. We also consider projected AdaGrad-Norm (Ward et al., 2020) and DoG (Ivgi et al., 2023) methods. Note that EGM with Rules 1–3 are provided in relations (13)–(15), whereas DA EGM Methods 1 and 2 are provided in relations (53) and (54) in Appendix F.1. For the min-max experiment, we consider the saddle-point formulation of a constrained ℓ_2-regression problem and compare our methods with EGM where step sizes are selected using grid-search, aka. EGM (grid-tuned), and also with Generalized Extra-Gradient (GEG) (Antonakopoulos et al., 2021).

In this section, we present experiments on Hölder-smooth polyhedral feasibility and saddle-point problems. Additional experiments, implementation details, and results for all problems are provided in Appendix G. All experiments were conducted on a 2021 MacBook Pro equipped with an Apple M1 Pro chip and 16 GB of RAM.

Hölder-smooth polyhedral feasibility problem. We consider the constrained optimization problem

$$f_q^\star := \min_{\|x\|_2 \leq R} f_q(x) \quad \text{where} \quad f_q(x) := \frac{1}{n}\sum_{i=1}^{n} [\langle a_i, x\rangle - b_i]_+^q \,, \tag{17}$$

where $a_i \in \mathbb{R}^d$, $b_i \in \mathbb{R}$, and $q \in [1,2]$. This objective belongs to the Hölder-smooth function class with Hölder exponent $\nu = q - 1 \in [0,1]$. Thus, $q = 1$ corresponds to the non-smooth endpoint, $q = 2$ corresponds to the Lipschitz-gradient case, and values $q \in (1,2)$ describe intermediate levels of smoothness. Problem (17) can also be viewed as a penalty formulation of the polyhedral feasibility problem $\langle a_i, x\rangle \leq b_i$ for all $i = 1, \ldots, n$. In particular, $f_q(x) = 0$ if and only if x satisfies all inequalities. The problem data are generated following the procedure in (Moshtaghifar et al., 2026, Appendix E). The construction ensures that x^* lies in the Euclidean ball and satisfies every inequality constraint. Consequently, $f_q(x^*) = f_q^* = 0$ for each tested value of q. The algorithms are run stochastically with batch size 1: every oracle call samples an index out of $\{1, \ldots, n\}$ uniformly at random with replacement and evaluates its (sub)gradient. Figure 1 shows that the methods exhibit comparable empirical performance. For $q = 1$ and $q = 1.5$, our DA EGM Method 1 and EGM Rule 1 perform slightly better, respectively. Whereas for $q = 2$, UGM with AdaGrad and DoG perform slightly better.

Min-max dual formulation of a constrained ℓ_2-regression problem. We consider the problem

$$\min_{\|x\|_2 \leq 1} \|Ax - b\|_2 = \min_{\|x\|_2 \leq 1} \max_{\|y\|_2 \leq 1} \langle Ax - b, y\rangle.$$

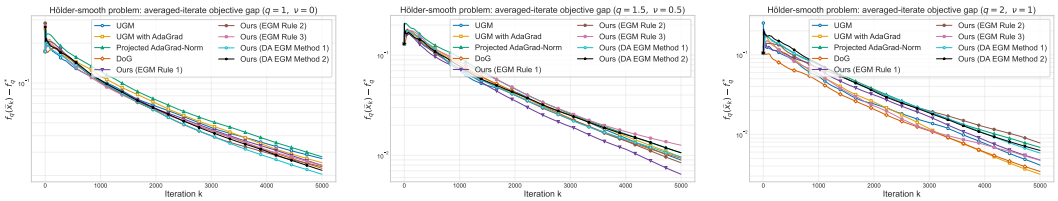
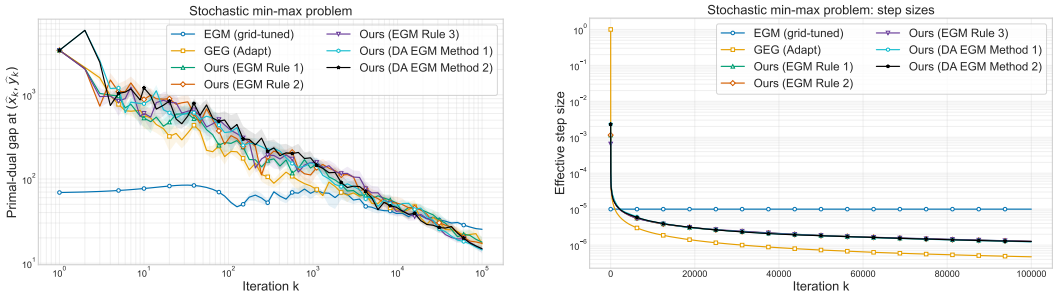


Figure 1: Results for the stochastic Hölder-smooth polyhedral feasibility problem. Averaged-iterate objective function gaps are presented for $q = 1$, $q = 1.5$, and $q = 2$, respectively.



(a) Primal-dual gap

(b) Step sizes

Figure 2: Results for the min-max dual formulation of constrained ℓ_2-regression problem.

The dual reformulation is a min-max problem with the associated VI mapping given by

$$F([x, y]) = \begin{bmatrix} A^\top y \\ b - Ax \end{bmatrix}.$$

If the dimension of matrix A is large, then computing the matrix-vector operation while computing $F([x, y])$ can be costly. We employ the randomized matrix-vector oracle described in (Nesterov & Nemirovski, 2013, Section 3.6.6) to construct the stochastic mapping oracle $\widehat{F}([x, y], \omega)$, where the randomness arises from sampling a column of the matrix. Further details are provided in Appendix G.

Figure 2 reports the primal-dual gap at the averaged iterate and the step sizes generated by the various methods. EGM (grid-tuned) uses a step size selected via grid search and therefore performs well during the initial iterations. However, later on, the performance of all methods becomes comparable. DA-EGM Methods 1 and 2, as well as EGM with Rule 3, perform slightly better in the final iterations.

7 Conclusion

We developed universal extra-gradient methods for stochastic VIs with approximately continuous mappings, a property we refer to as (δ, L)-continuity. We established calculus rules for (δ, L)-continuous mappings and showed that Hölder-continuous mappings and their linear combinations belong to this class. We derived several line-search-free step-size rules in the form of balance equations that adapt to both the continuity of the mapping and the stochastic-oracle variance without requiring prior knowledge of either quantity. Our analysis establishes optimal convergence guarantees through a unified residual function encompassing optimization, saddle-point, and monotone VI problems. For Hölder-continuous mappings, the resulting rates recover the optimal dimension-independent complexity. We also presented distance-adaptive variants that reduce sensitivity to the initial diameter estimate. These variants preserve the same optimal convergence rates while improving some of the constants relative to their non-distance-adaptive counterparts, whose bounds depend more strongly on the initial estimate $\widehat{D}$. Numerical experiments on optimization and min-max problems demonstrate the practical effectiveness of the proposed methods across different continuity and stochastic noise regimes. Future work may extend these methods to non-Euclidean geometries and to unbounded constraint sets.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing the experimental methods, including writing, debugging, and editing portions of the experimental code. We also used generative AI tools to correct grammar and improve the readability of author-written text.

We did not use generative AI tools to develop the theoretical or conceptual framework, formulate mathematical claims, assist with proofs, propose or refine hypotheses, design the research methodology or experiments, generate synthetic datasets, analyze or interpret experimental results, or draw the scientific conclusions of this work. The remaining tasks requiring disclosure were not applicable to this work.

All AI-assisted code was manually reviewed and tested by the authors before being used to produce the reported results. All AI-assisted language edits were reviewed to ensure that they preserved the authors' intended technical meaning and claims. We take responsibility for the final content of this work, including all text, claims, results, and artifacts produced with the aid of generative AI.

REFERENCES

Kimon Antonakopoulos, Veronica Belmega, and Panayotis Mertikopoulos. An adaptive mirror-prox algorithm for variational inequalities with singular operators. *Advances in Neural Information Processing Systems*, 32, 2019.

Kimon Antonakopoulos, Thomas Pethick, Ali Kavis, Panayotis Mertikopoulos, and Volkan Cevher. Sifting through the noise: Universal first-order methods for stochastic variational inequalities. *Advances in Neural Information Processing Systems*, 34:13099–13111, 2021.

A Auslender, M Gourgand, and A Guillet. Resolution numerique d'inegalites variationnelles. In *Analyse Convexe et Ses Applications: Comptes Rendus, Janvier 1974*, pp. 1–40. Springer, 1974.

Francis Bach and Kfir Y Levy. A universal algorithm for variational inequalities adaptive to smoothness and noise. In *Conference on Learning Theory*, pp. 164–194. PMLR, 2019.

Yair Carmon and Oliver Hinder. Making SGD parameter-free. In *Conference on Learning Theory*, pp. 2360–2389. PMLR, 2022.

Abhishek Chakraborty and Angelia Nedić. Popov mirror-prox method for variational inequalities. *arXiv preprint arXiv:2507.23395*, 2025.

Abhishek Chakraborty and Angelia Nedić. Randomized feasibility methods for constrained optimization with adaptive step sizes. In *Forty-third International Conference on Machine Learning*, 2026.

Olivier Devolder, Francois Glineur, and Yurii Nesterov. First-order methods of smooth convex optimization with inexact oracle. *Mathematical Programming*, 146(1):37–75, 2014.

Nikita Doikov and Yurii Nesterov. Universal reduced-operator method and high-order global curvature bounds. *arXiv preprint arXiv:2511.07341*, 2025.

Rick Durrett. *Probability: theory and examples*. Cambridge University Press, 5th edition, 2019.

Alina Ene, Huy L Nguyen, and Adrian Vladu. Adaptive gradient methods for constrained convex optimization and variational inequalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7314–7321, 2021.

Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer, 2003.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Sergey Guminov, Alexander Gasnikov, Anton Anikin, and Alexander Gornov. A universal modification of the linear coupling method. *Optimization methods and software*, 34(3):560–577, 2019.

Masaru Ito, Zhaosong Lu, and Chuan He. A parameter-free conditional gradient method for composite minimization under hölder condition. *Journal of Machine Learning Research*, 24(166):1–34, 2023.

Maor Ivgi, Oliver Hinder, and Yair Carmon. Dog is SGD's best friend: A parameter-free dynamic step size schedule. In *International Conference on Machine Learning*, pp. 14465–14499. PMLR, 2023.

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

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

Dmitry Kamzolov, Pavel Dvurechensky, and Alexander V Gasnikov. Universal intermediate gradient method for convex problems with inexact oracle. *Optimization Methods and Software*, 36(6): 1289–1316, 2021.

Ahmed Khaled, Konstantin Mishchenko, and Chi Jin. DoWG unleashed: An efficient universal parameter-free gradient descent method. *Advances in Neural Information Processing Systems*, 36: 6748–6769, 2023.

David Kinderlehrer and Guido Stampacchia. *An introduction to variational inequalities and their applications*. SIAM, 2000.

Anton Klimza, Alexander Gasnikov, Fedor Stonyakin, and Mohammad Alkousa. Universal methods for variational inequalities: Deterministic and stochastic cases. *Chaos, Solitons & Fractals*, 187: 115418, 2024.

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

Marc Lassonde. Convex kkm maps, monotone operators and minty variational inequalities. *Journal of Fixed Point Theory and Applications*, 17(1):137–143, 2015.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Zijian Liu and Zhengyuan Zhou. Stochastic nonsmooth convex optimization with heavy-tailed noises: High-probability bound, in-expectation rate and initial distance adaptation. *arXiv preprint arXiv:2303.12277*, 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Yura Malitsky. Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184 (1):383–410, 2020.

Patrice Marcotte and Daoli Zhu. Weak sharp solutions of variational inequalities. *SIAM Journal on Optimization*, 9(1):179–189, 1998.

George J Minty. Monotone (nonlinear) operators in Hilbert space. 29(3):341–346, 1962.

Mohammad Moshtaghifar, Anton Rodomanov, Daniil Vankov, and U Sebastian Stich. DADA: Dual averaging with distance adaptation. In *International Conference on Learning Representations*, volume 2026, pp. 155565–155587, 2026.

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.

Arkadi Nemirovski, Shmuel Onn, and Uriel G Rothblum. Accuracy certificates for computational problems with convex structure. *Mathematics of Operations Research*, 35(1):52–78, 2010.

Yu Nesterov. Smooth minimization of non-smooth functions. *Mathematical programming*, 103(1): 127–152, 2005.

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

Yurii Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical programming*, 120(1):221–259, 2009.

Yurii Nesterov. Universal complexity bounds for universal gradient methods in nonlinear optimization. *arXiv preprint arXiv:2509.20902*, 2025.

Yurii Nesterov and Arkadi Nemirovski. On first-order algorithms for ℓ_1/nuclear norm minimization. *Acta Numerica*, 22:509–575, 2013. doi: 10.1017/S096249291300007X.

Anton Rodomanov and Yurii Nesterov. Subgradient ellipsoid method for nonsmooth convex problems. *Mathematical Programming*, 199(1):305–341, 2023.

Anton Rodomanov, Xiaowen Jiang, and Sebastian U Stich. Universality of adagrad stepsizes for stochastic optimization: Inexact oracle, acceleration and variance reduction. *Advances in Neural Information Processing Systems*, 37:26770–26813, 2024a.

Anton Rodomanov, Ali Kavis, Yongtao Wu, Kimon Antonakopoulos, and Volkan Cevher. Universal gradient methods for stochastic convex optimization. In *Forty-first International Conference on Machine Learning*, 2024b.

Oleg S Savchuk, Fedor Sergeevich Stonyakin, Aleksandr A Vyguzov, Mohammad Soud Alkousa, and Alexander Vladimirovich Gasnikov. Adaptive primal–dual methods with an inexact oracle for relatively smooth optimization problems and their applications to recovering low-rank matrices. *Computational Mathematics and Mathematical Physics*, 65(7):1605–1627, 2025.

Fedor Stonyakin, Alexander Tyurin, Alexander Gasnikov, Pavel Dvurechensky, Artem Agafonov, Darina Dvinskikh, Mohammad Alkousa, Dmitry Pasechnyuk, Sergei Artamonov, and Victorya Piskunova. Inexact model: A framework for optimization and variational inequalities. *Optimization Methods and Software*, 36(6):1155–1201, 2021.

Fedor Stonyakin, Alexander Gasnikov, Pavel Dvurechensky, Alexander Titov, and Mohammad Alkousa. Generalized mirror prox algorithm for monotone variational inequalities: Universality and inexact oracle. *Journal of Optimization Theory and Applications*, 194(3):988–1013, 2022.

Rachel Ward, Xiaoxia Wu, and Leon Bottou. Adagrad stepsizes: Sharp convergence over nonconvex landscapes. *Journal of Machine Learning Research*, 21(219):1–30, 2020.

Alp Yurtsever, Quoc Tran Dinh, and Volkan Cevher. A universal primal-dual convex optimization framework. *Advances in Neural Information Processing Systems*, 28, 2015.

A Comparison with (Klimza et al., 2024)

Klimza et al. (2024, Algorithm 1) consider Hölder-continuous F and the balance equation

$$\frac{H_{k+1}-H_k}{2}D^2 = \left[\langle \widehat{F}(x_{k+1},\omega_{k+1}), x_{k+1}-v_{k+1}\rangle - \frac{H_{k+1}}{2}\|v_k-v_{k+1}\|^2\right]_+,$$

which is equivalent to the explicit update

$$H_{k+1} = H_k + \frac{2\left[\langle \widehat{F}(x_{k+1},\omega_{k+1}), x_{k+1}-v_{k+1}\rangle - \frac{H_k}{2}\|v_k-v_{k+1}\|^2\right]_+}{D^2+\|v_k-v_{k+1}\|^2}.$$

We point out three steps of the analysis that we were unable to verify.

(i) Lower bound in Klimza et al. (2024, Equation 33). The lower bound $D^2 + \frac{1}{2}\|v_k - v_{k+1}\|^2 - \frac{1}{2}\|v_k + v_{k+1} - 2x_{k+1}\|^2 \geq \frac{D^2}{2}$ is claimed using compactness. However, this does not follow from boundedness alone. For instance, if $v_{k+1} = v_k$, the left-hand side equals $D^2 - 2\|x_{k+1} - v_k\|^2$, which is smaller than $D^2/2$ whenever $\|x_{k+1} - v_k\| > D/2$.

(ii) Order of maximum and expectation in Equations (35)–(37). The analysis exchanges $\max_{x \in X}$ and $\mathbb{E}$. Since $\mathbb{E}[\max_x \cdot] \geq \max_x \mathbb{E}[\cdot]$, the resulting bound controls only the weaker quantity and not the expected gap. We avoid this by using an auxiliary sequence $(h_k)_{k \geq 0}$ (Lemma D.2), which allows us to take $\max_{x \in X}$ first and then the total expectation.

(iii) Corner cases. As noted in the updated arXiv version of Klimza et al. (2024), some steps of the proof do not apply when the Hölder exponent is $\nu = 0$ or $\nu = 1$. The guarantees therefore do not cover the bounded-variation and Lipschitz regimes, both of which our analysis includes.

B Auxiliary Results

Lemma B.1. *(Rodomanov et al., 2024a, Lemma A.6) Let $(H_k)_{k=0}^{\infty}$ be a non-negative non-decreasing sequence and let $\bar{H} \geq 0$ be a constant. Then, for any $T \geq 1$, we have*

$$\sum_{k=0}^{T-1} \left[\min\{\bar{H}, H_{k+1}\} - H_k\right]_+ = \left[\min\{\bar{H}, H_T\} - H_0\right]_+ .$$

Lemma B.2. *Let $(\gamma_k)_{k \geq 0}$ be a nonnegative sequence and let $V > 0$. Suppose that*

$$\gamma_0 \leq V$$

and that, for some constant $a > 0$ and arbitrary numbers $e_k \geq 0$,

$$\gamma_{k+1} - \gamma_k \leq a\,[V - \gamma_{k+1}]_+ + e_k \qquad \forall k \geq 0. \tag{18}$$

Then, for any $k \geq 1$,

$$\gamma_k \leq V + \sum_{i=0}^{k-1} e_i.$$

Proof. We prove the claim by induction on t. For $t = 0$, the claim follows directly from $\gamma_0 \leq V$. Suppose that, for some $k \geq 0$,

$$\gamma_k \leq V + \sum_{i=0}^{k-1} e_i.$$

Now, we consider two cases. If $\gamma_{k+1} \leq V$, then, since $e_i \geq 0$,

$$\gamma_{k+1} \leq V \leq V + \sum_{i=0}^{k} e_i.$$

Else, if $\gamma_{k+1} > V$, then

$$[V - \gamma_{k+1}]_+ = 0.$$

Therefore, relation (18) yields

$$\gamma_{k+1} - \gamma_k \leq e_k.$$

Using the induction hypothesis, we obtain

$$\gamma_{k+1} \leq \gamma_k + e_k \leq V + \sum_{i=0}^{k-1} e_i + e_k = V + \sum_{i=0}^{k} e_i.$$

Thus, the induction is now complete. **Q.E.D.**

Lemma B.3. *(Liu & Zhou, 2023, Lemma 30), (Moshtaghifar et al., 2026, Lemma 2), (Chakraborty & Nedić, 2026, Lemma B.3) Let the sequence $(D_t)_{t \geq 0}$ be positive and non-decreasing. Then for any $T \geq 1$, we obtain*

$$\min_{1 \leq t \leq T} \frac{D_t}{\sum_{k=0}^{t-1} D_k} \leq \frac{\left(\frac{D_T}{D_0}\right)^{\frac{1}{T}} \log \frac{eD_T}{D_0}}{T}.$$

Lemma B.4. *(Durrett, 2019, Theorem 4.4.4) If the sequence $(B_t)_{t=1}^T$ is a martingale, then by Doob's L^2 inequality we obtain the following for any $T \geq 1$:*

$$\mathbb{E}\Big[\max_{1\leq t\leq T} |B_t|^2\Big] \leq 4\mathbb{E}\big[|B_T|^2\big].$$

C Proofs for Section 2

C.1 Proof of Lemma 2.3

Proof. Let us fix arbitrary $x, y \in X$ and denote $r = \|x - y\|$. We will consider three case depending on the range of the value of ν.

If $\nu = 0$, then $\|F(x) - F(y)\|_*^2 \leq H_0^2$ and $L(\delta) = \frac{H_0^2}{4\delta}$, so the claim is trivial.

If $\nu = 1$, then $\|F(x) - F(y)\|_*^2 \leq H_1^2 r^2$ and $L(\delta) = H_1$, and the claim is again trivial.

Now suppose that $0 < \nu < 1$. Then, for any $L > 0$, we can estimate

$$\|F(x) - F(y)\|_*^2 \leq H_\nu^2 r^{2\nu} = L^2 r^2 + (H_\nu^2 r^{2\nu} - L^2 r^2) \leq L^2 r^2 + \Delta(L),$$

where

$$\Delta(L) = \max_{r\geq 0}\{H_\nu^2 r^{2\nu} - L^2 r^2\} = \max_{\tau\geq 0}\left\{H_\nu^2 \tau - L^2 \tau^{\frac{1}{\nu}}\right\} = (1-\nu)\left(\frac{\nu}{L^2}\right)^{\frac{\nu}{1-\nu}} H_\nu^{\frac{2}{1-\nu}},$$

where the final identity follows from obtaining the maximum value of the concave function. Let us fix an arbitrary $\delta > 0$ and choose L such that $\Delta(L) = 4L\delta$. This is $L = L(\delta)$ as defined in the statement of Lemma 2.3. The claim now follows. **Q.E.D.**

C.2 Proof of the Claim in Remark 2.4

Lemma C.1. *For any $\nu \in [0, 1]$, $H_\nu > 0$, and $\delta > 0$, let $L(\delta)$ and $\tilde{L}(\delta)$ be defined as*

$$L(\delta) = \left(\frac{1-\nu}{4\delta}\right)^{\frac{1-\nu}{1+\nu}} \nu^{\frac{\nu}{1+\nu}} H_\nu^{\frac{2}{1+\nu}}, \quad \tilde{L}(\delta) = \left(\frac{1-\nu}{2(1+\nu)\delta}\right)^{\frac{1-\nu}{1+\nu}} H_\nu^{\frac{2}{1+\nu}}$$

with the convention $0^0 = 1$. Then,

$$L(\delta) \leq \tilde{L}(\delta).$$

Proof. Using the definitions of $L(\delta)$ and $\tilde{L}(\delta)$, we hsve

$$\frac{L(\delta)}{\tilde{L}(\delta)} = \left(\frac{1+\nu}{2}\right)^{\frac{1-\nu}{1+\nu}} \nu^{\frac{\nu}{1+\nu}}.$$

The maximum value on the right-hand side is attained at $\nu = 1$. In order to show that, we apply log on both sides (since log is a monotone function) to obtain

$$(1+\nu)\ln\left(\frac{L(\delta)}{\tilde{L}(\delta)}\right) = (1-\nu)\ln\left(\frac{1+\nu}{2}\right) + \nu\ln\nu := \phi(\nu). \tag{19}$$

Applying double derivative on $\phi(\nu)$ with respect to ν, we obtain

$$\nabla_\nu^2 \phi(\nu) = \frac{1-\nu}{\nu(1+\nu)^2} \geq 0 \quad \text{for } \nu \in (0, 1].$$

Hence, $\phi(\nu)$ is a convex function with respect to ν. Since the function $\phi(\nu)$ is convex, the maximum value will be attained at one of the end points. We see $\lim_{\nu\to 0+} \phi(\nu) = \ln(0.5) + \lim_{\nu\to 0+} \nu\ln\nu = \ln(0.5) + \lim_{\nu\to 0+} \frac{1/\nu}{-1/\nu^2} = \ln(0.5) < 0$ and $\phi(1) = 0$. Hence, we obtain $\phi(\nu) \leq 0$ for any $\nu \in [0, 1]$, which when used in relation (19) yields $L(\delta) \leq \tilde{L}(\delta)$. **Q.E.D.**

C.3 Calculus properties of (δ, L)-continuity for VI

Everywhere for the ease of notation, we refer $L \equiv L(\delta)$.

Lemma C.2. *The (δ, L)-continuity defined in Assumption 2.2 satisfies the following properties:*
(i) (Scalar multiplication) If F is (δ, L)-continuous and $c > 0$ is a scalar, then cF is $(c\delta, cL)$-continuous.
(ii) (Summation) If F_1 is (δ_1, L_1)-continuous and F_2 is (δ_2, L_2)-continuous, then $F := F_1 + F_2$ is $(\delta_1 + \delta_2, L_1 + L_2)$-continuous.
(iii) (Composition with an affine mapping) If F is (δ, L)-continuous, then for any matrix $A \in \mathbb{R}^{n\times n}$ and scalar $b \in \mathbb{R}$, $F(Ax + b)$ is $\left(\frac{\delta}{\|A\|}, L\|A\|\right)$-continuous, where $\|A\| := \sup_{\|x\|\leq 1} \|Ax\|$ is the mapping norm.

Proof. Part (i): The proof for part (i) is immediate. For any $x, y \in X$,

$$\|cF(x) - cF(y)\|_*^2 = c^2\|F(x) - F(y)\|_*^2 \leq (cL)^2\|x - y\|^2 + 4(c\delta)(cL).$$

Hence, the mapping $cF(\cdot)$ is $(c\delta, cL)$ continuous.

Part (ii): For $F(\cdot) = F_1(\cdot) + F_2(\cdot)$ and any $x, y \in X$, we have

$$\|F(x) - F(y)\|_*^2 = \|(F_1(x) - F_1(y)) + (F_2(x) - F_2(y))\|_*^2. \tag{20}$$

For simplicity, let us define

$$\Delta_1 = F_1(x) - F_1(y) \quad \text{and} \quad \Delta_2 = F_2(x) - F_2(y).$$

Then, relation (20) reduces to

$$\begin{aligned}
&\|F(x) - F(y)\|_*^2 = \|\Delta_1 + \Delta_2\|_*^2 \leq (\|\Delta_1\|_* + \|\Delta_2\|_*)^2 = \|\Delta_1\|_*^2 + \|\Delta_2\|_*^2 + 2\|\Delta_1\|_*\|\Delta_2\|_* \\
&\leq (1+\tau)\|\Delta_1\|_*^2 + \left(1 + \frac{1}{\tau}\right)\|\Delta_2\|_*^2 \\
&\leq \left((1+\tau)L_1^2 + \left(1 + \frac{1}{\tau}\right)L_2^2\right)\|x - y\|^2 + 4\left((1+\tau)L_1\delta_1 + \left(1 + \frac{1}{\tau}\right)L_2\delta_2\right),
\end{aligned} \tag{21}$$

where the last two inequalities are obtained from Young's inequality, where $\tau > 0$ is arbitrary, and Assumption 2.2. The optimal value $\tau^* > 0$ that minimizes the coefficient of the first term in (21) is given by

$$\tau^* = \frac{L_2}{L_1}.$$

Substituting the value of τ^* back in equation (21) yields

$$\|F(x) - F(y)\|_*^2 \leq (L_1 + L_2)^2\|x - y\|^2 + 4(L_1 + L_2)(\delta_1 + \delta_2),$$

which implies that F is $(\delta_1 + \delta_2, L_1 + L_2)$ continuous.

Part (iii): Since F is (δ, L)-continuous, hence for any $x, y \in X$,

$$\begin{aligned}
\|F(Ax + b) - F(Ay + b)\|_*^2 &\leq L^2\|A(x - y)\|^2 + 4L\delta \\
&\leq (L\|A\|)^2\|x - y\|^2 + 4\left(L\|A\|\frac{\delta}{\|A\|}\right).
\end{aligned}$$

Hence, the claim follows. **Q.E.D.**

Lemma C.3. *Let F_i is Hölder continuous with parameters (ν_i, H_{ν_i}) for all $i \in |\mathcal{I}_1|$ and let $\bar{F}_j$ is (δ_j, L_j)-continuous for all $j \in |\mathcal{I}_2|$, where $|\mathcal{I}_1| \geq 0$ and $|\mathcal{I}_2| \geq 0$, then $F = \sum_{i\in|\mathcal{I}_1|} F_i + \sum_{j\in|\mathcal{I}_2|} \bar{F}_j$ is $(\sum_{i\in|\mathcal{I}_1|}\delta_i + \sum_{j\in|\mathcal{I}_2|}\delta_j, \sum_{i\in|\mathcal{I}_1|} L_i + \sum_{j\in|\mathcal{I}_2|} L_j)$-continuous, where $\delta_i > 0$ is an arbitrary constant and*

$$L_i \equiv L_i(\delta_i) = \left(\frac{1-\nu_i}{4\delta_i}\right)^{\frac{1-\nu_i}{1+\nu_i}} \nu_i^{\frac{\nu_i}{1+\nu_i}} H_{\nu_i}^{\frac{2}{1+\nu_i}} \quad \forall i \in |\mathcal{I}_1|,$$

with the convention $0^0 = 1$.

Proof. If F_i is Hölder continuous with parameters (ν_i, H_{ν_i}) for any $i \in |\mathcal{I}_1|$, then by Lemma 2.3, we can fix any arbitrary $\delta_i > 0$ and obtain the expression for $L_i(\delta_i)$ as provided in the lemma. Now, by applying Lemma C.2(ii) recursively, we obtain the desired result. **Q.E.D.**

D PROOFS FOR SECTION 3

D.1 GENERIC LEMMA FOR EGM

Lemma D.1. *Consider any constant $\eta \geq 0$ and define the iterate sequence $(h_k)_{k\geq 0}$ as*

$$h_{k+1} = \arg\min_{x\in X} \left\{ \langle -Z_{k+1}, x - h_k\rangle + \frac{\eta}{2}\|x - h_k\|^2 \right\},$$

where the stochastic error Z_k is defined in relation (9)*. Then, for method* (8) *under Assumption 2.1, we have, for any $x \in X$ and any $k \geq 0$,*

$$\langle F(x_{k+1}), x_{k+1} - x\rangle \leq \Delta_k(H_{k+1}) + \frac{3}{2}(H_{k+1} - H_k)D^2 + \langle Z_{k+1}, h_k - x_{k+1}\rangle + \frac{\|Z_{k+1}\|_*^2}{2\eta} + \frac{\eta}{2}[\|x - h_k\|^2 - \|x - h_{k+1}\|^2] + \frac{H_k}{2}\|x - v_k\|^2 - \frac{H_{k+1}}{2}\|x - v_{k+1}\|^2, \quad (22)$$

where $\Delta_k(H) := \xi_k - \frac{H}{2}r_k^2$ with $r_k^2 = \|v_k - x_{k+1}\|^2 + \|v_{k+1} - x_{k+1}\|^2$ and

$$\begin{aligned}\xi_k &= \langle \widehat{F}(x_{k+1}, \omega_{k+1}), x_{k+1} - v_{k+1}\rangle - \frac{H_k}{2}\|v_k - v_{k+1}\|^2 + \frac{H_k}{2}r_k^2 \\ &\leq \langle \widehat{F}(x_{k+1}, \omega_{k+1}) - \widehat{F}(v_k, \omega_k), x_{k+1} - v_{k+1}\rangle \equiv \xi_k'. \end{aligned} \quad (23)$$

Moreover,

$$\Delta_k(H) \leq \xi_k' - \frac{H}{2}r_k^2 \leq \frac{1}{2H}\|\widehat{F}(x_{k+1}, \omega_{k+1}) - \widehat{F}(v_k, \omega_k)\|_*^2 - \frac{H}{2}\|x_{k+1} - v_k\|^2 \equiv \bar{\Delta}_k(H). \quad (24)$$

Proof. Let $x \in X$ and $k \geq 0$ be arbitrary. From the definition of v_{k+1} and using H_k-strong convexity of the update, it follows that

$$\begin{aligned}&\langle \widehat{F}(x_{k+1}, \omega_{k+1}),\, v_{k+1} - x_{k+1}\rangle + \frac{H_k}{2}\|v_k - v_{k+1}\|^2 + \frac{H_k}{2}\|x - v_{k+1}\|^2 \\ &\leq \langle \widehat{F}(x_{k+1}, \omega_{k+1}),\, x - x_{k+1}\rangle + \frac{H_k}{2}\|v_k - x\|^2.\end{aligned}$$

Rearranging terms along with algebraic manipulations, we obtain

$$\langle \widehat{F}(x_{k+1}, \omega_{k+1}), x_{k+1} - x\rangle \leq \xi_k - \frac{H_{k+1}}{2}r_k^2 + \frac{H_{k+1} - H_k}{2}[r_k^2 + \|x - v_{k+1}\|^2] + \frac{H_k}{2}\|x - v_k\|^2 - \frac{H_{k+1}}{2}\|x - v_{k+1}\|^2, \quad (25)$$

where the quantity $r_k^2 = \|v_k - x_{k+1}\|^2 + \|v_{k+1} - x_{k+1}\|^2$ and

$$\xi_k = \langle \widehat{F}(x_{k+1}, \omega_{k+1}), x_{k+1} - v_{k+1}\rangle - \frac{H_k}{2}\|v_k - v_{k+1}\|^2 + \frac{H_k}{2}r_k^2. \quad (26)$$

For notational simplicity, we define

$$\Delta_k(H) = \xi_k - \frac{H}{2}r_k^2.$$

Hence, applying the preceding relation and Assumption 2.1 to relation (25), we obtain

$$\langle \widehat{F}(x_{k+1}, \omega_{k+1}), x_{k+1} - x\rangle \leq \Delta_k(H_{k+1}) + \frac{3}{2}(H_{k+1} - H_k)D^2 + \frac{H_k}{2}\|x - v_k\|^2 - \frac{H_{k+1}}{2}\|x - v_{k+1}\|^2, \quad (27)$$

The left-hand side of the preceding relation can be written as

$$\langle \widehat{F}(x_{k+1}, \omega_{k+1}), x_{k+1} - x\rangle = \langle F(x_{k+1}), x_{k+1} - x\rangle + \langle Z_{k+1}, x_{k+1} - h_k\rangle + \langle Z_{k+1}, h_k - x\rangle, \quad (28)$$

where $Z_{k+1} = \widehat{F}(x_{k+1}, \omega_{k+1}) - F(x_{k+1})$ (cf. relation (9)) and with $h_0 \in X$, h_{k+1} for any $k \geq 0$ is defined as

$$h_{k+1} = \arg\min_{x \in X} \left\{ \langle -Z_{k+1}, x - h_k \rangle + \frac{\eta}{2} \|x - h_k\|^2 \right\},$$

where $\eta \geq 0$ is an arbitrary constant. Substituting relation (28) to relation (27) we obtain

$$\begin{aligned} \langle F(x_{k+1}), x_{k+1} - x \rangle \leq \Delta_k(H_{k+1}) + \frac{3}{2}(H_{k+1} - H_k)D^2 + \langle Z_{k+1}, h_k - x_{k+1} \rangle \\ + \langle Z_{k+1}, x - h_k \rangle + \frac{H_k}{2}\|x - v_k\|^2 - \frac{H_{k+1}}{2}\|x - v_{k+1}\|^2. \end{aligned}$$

Applying Lemma D.2 on the fourth term of the right-hand side of the preceding relation yields relation (22).

Next, we prove that $\xi_k \leq \xi'_k$. We have

$$\begin{aligned} \xi_k =& \langle \widehat{F}(x_{k+1}, \omega_{k+1}) - \widehat{F}(v_k, \omega_k), x_{k+1} - v_{k+1} \rangle - \frac{H_k}{2}\|v_k - v_{k+1}\|^2 \\ &+ \langle \widehat{F}(v_k, \omega_k), x_{k+1} - v_{k+1} \rangle + \frac{H_k}{2} r_k^2. \end{aligned} \tag{29}$$

Since the iterate x_{k+1} is the solution of H_k-strongly convex function minimization over X, it follows that for any $y \in X$,

$$\begin{aligned} &\langle \widehat{F}(v_k, \omega_k), x_{k+1} - v_k \rangle + \frac{H_k}{2}\|v_k - x_{k+1}\|^2 + \frac{H_k}{2}\|y - x_{k+1}\|^2 \\ &\leq \langle \widehat{F}(v_k, \omega_k), y - v_k \rangle + \frac{H_k}{2}\|y - v_k\|^2, \end{aligned}$$

Substituting $y = v_{k+1}$ in the preceding relation and rearranging terms yields the following upper bound

$$\langle \widehat{F}(v_k, \omega_k), x_{k+1} - v_{k+1} \rangle \leq \frac{H_k}{2}\|v_k - v_{k+1}\|^2 - \frac{H_k}{2} r_k^2,$$

which when substituted back in relation (29) yields relation (23).

Finally, the upper bound in relation (24) is obtained using Young's inequality. This concludes the proof. **Q.E.D.**

D.2 Lemma to control the stochastic error

We present a lemma which is a modification of (Juditsky et al., 2011, Lemma 3 and Corollary 2) and (Chakraborty & Nedić, 2025, Lemma 5) that controls the stochastic error quantity required for our analysis.

Lemma D.2. *Let $h_0 \in X$ and h_{k+1} for all $k \geq 0$ be updated as*

$$h_{k+1} = \arg\min_{x \in X} \left\{ \langle -Z_{k+1}, x - h_k \rangle + \frac{\eta}{2} \|x - h_k\|^2 \right\},$$

where $\eta \geq 0$ is a constant and $Z_{k+1} := \widehat{F}(x_{k+1}, \omega_{k+1}) - F(x_{k+1})$ for any $\omega_{k+1} \in \Xi$. Then, for any $k \geq 0$ and any $x \in X$, we have

$$\langle Z_{k+1}, x - h_k \rangle \leq \frac{\|Z_{k+1}\|_*^2}{2\eta} + \frac{\eta}{2}[\|x - h_k\|^2 - \|x - h_{k+1}\|^2].$$

Proof. Using the H_k-strong convexity of the function in the update rule for h_{k+1}, we obtain for any $k \geq 0$ and any $x \in X$,

$$\begin{aligned} &\langle -Z_{k+1}, h_{k+1} - h_k \rangle + \frac{\eta}{2}\|h_{k+1} - h_k\|^2 + \frac{\eta}{2}\|x - h_{k+1}\|^2 \\ &\leq \langle -Z_{k+1}, x - h_k \rangle + \frac{\eta}{2}\|x - h_k\|^2. \end{aligned}$$

Simplifying the preceding relation yields

$$\langle Z_{k+1}, x - h_k \rangle \leq \langle Z_{k+1}, h_{k+1} - h_k \rangle - \frac{\eta}{2}\|h_{k+1} - h_k\|^2 + \frac{\eta}{2}[\|x - h_k\|^2 - \|x - h_{k+1}\|^2].$$

Applying Young's inequality to the first quantity on the right-hand side, we obtain the desired result. **Q.E.D.**

D.3 ALGORITHMIC UPDATES FOR VARIOUS EXAMPLES

Here, we derive the explicit updates for H_{k+1} under different balance equation examples. We first present a lemma that facilitates these derivations.

Lemma D.3. *(Rodomanov et al., 2024b, Lemma E.1) Let $\Omega > 0$ be an arbitrary constant and let $(\zeta_k)_{k\geq 0}$ be a sequence of non-negative scalars. Let $(H_k)_{k\geq 0}$ be a sequence of non-decreasing scalars and $(\beta_k)_{k\geq 0}$ be a scalar sequence. Then, for all $k \geq 0$, the balance equation*

$$\frac{H_{k+1} - H_k}{2}\Omega = \left[\beta_k - \frac{H_{k+1}}{2}\zeta_k\right]_+$$

is equivalent to the balance equation

$$\frac{H_{k+1} - H_k}{2}(\Omega + \zeta_k) = \left[\beta_k - \frac{H_k}{2}\zeta_k\right]_+.$$

We now present the explicit update rules for the sequence H_k corresponding to each balance-equation example separately, followed by the complete algorithm.

Rule 1: By Lemma D.3(i), the balance equation (13) has a unique solution given by

$$\begin{aligned} H_{k+1} &= H_k + \frac{2[\Delta_k(H_k)]_+}{R_k^2} = H_k + \frac{2\left[\xi_k - \frac{H_k}{2}r_k^2\right]_+}{R_k^2} \\ &= H_k + \frac{2\left[\langle \widehat{F}(x_{k+1}, \omega_{k+1}), x_{k+1} - v_{k+1}\rangle - \frac{H_k}{2}\|v_k - v_{k+1}\|^2\right]_+}{R_k^2}, \end{aligned} \tag{30}$$

where the quantity R_k is defined as

$$R_k^2 = \widehat{D}^2 + r_k^2 = \widehat{D}^2 + \|v_k - x_{k+1}\|^2 + \|v_{k+1} - x_{k+1}\|^2. \tag{31}$$

Rule 2: Here, for the balance equation (14), we obtain a unique solution for H_{k+1} for all $k \geq 0$, using Lemma D.3(i) as follows:

$$H_{k+1} = H_k + \frac{2\left[\xi_k' - \frac{H_k}{2}r_k^2\right]_+}{R_k^2}, \tag{32}$$

where the quantity R_k is defined in relation (31).

We present both Rules 1 and 2 in a combined way in Algorithm 1. If one compares the H_{k+1} update

Algorithm 1 Line search-free EGM using ξ_k or ξ_k'

1: **Input:** $v_0 \in X$, $H_0 = 0$, $\widehat{D} > 0$
2: **for** $k = 0, 1, 2, \ldots, T-1$ **do**
3: Compute x_{k+1} and v_{k+1} according to EGM (relation (8))
4: Compute R_k from relation (31)
5: $H_{k+1} = H_k + \frac{2\left[\bar{\xi}_k - \frac{H_k}{2}r_k^2\right]_+}{R_k^2}$ where $\bar{\xi}_k = \xi_k$ or ξ_k' in Lemma D.1
6: **end for**

rules in Rules 1 (cf. relation (30)) and 2 (cf. relation (32)), we observe that since $\xi_k \leq \xi_k'$, it follows that

$$H_k + \frac{2\left[\xi_k - \frac{H_k}{2}r_k^2\right]_+}{R_k^2} \leq H_k + \frac{2\left[\xi_k' - \frac{H_k}{2}r_k^2\right]_+}{R_k^2}.$$

This indicates that the balance equation in Rule 1 may yield larger step sizes ($\propto 1/H_{k+1}$) compared to Rule 2, which can be beneficial in practice.

Algorithm 2 Line search-free EGM using classic AdaGrad-norm

1: **Input:** $v_0 \in X$, $H_0 = 0$, $\widehat{D} > 0$
2: **for** $k = 0, 1, 2, \ldots, T-1$ **do**
3: Compute x_{k+1} and v_{k+1} according to EGM (relation (8))
4: $H_{k+1} = \sqrt{H_k^2 + \frac{\|\widehat{F}(x_{k+1},\omega_{k+1}) - \widehat{F}(v_k,\omega_k)\|_*^2}{\widehat{D}^2}}$
5: **end for**

Rule 3: The errors in Lemma D.1 can also be controlled using a classical AdaGrad-norm-style balance equation. In particular, large variations in $\widehat{F}$ should naturally lead to more conservative steps, while stable regions permit more aggressive updates. The classical AdaGrad-norm mechanism captures this intuition, as reflected in the balance equation (15). The complete method is presented in Algorithm 2. In view of the balance equation for this example, we always have

$$(H_{k+1} - H_k)\widehat{D}^2 = \frac{\|\widehat{F}(x_{k+1},\omega_{k+1}) - \widehat{F}(v_k,\omega_k)\|_*^2}{H_{k+1} + H_k}.$$

Since $H_k \leq H_{k+1}$, the following relation holds for all $k \geq 0$:

$$\frac{\|\widehat{F}(x_{k+1},\omega_{k+1}) - \widehat{F}(v_k,\omega_k)\|_*^2}{2H_{k+1}} \leq (H_{k+1} - H_k)\widehat{D}^2 \leq \frac{\|\widehat{F}(x_{k+1},\omega_{k+1}) - \widehat{F}(v_k,\omega_k)\|_*^2}{H_{k+1}}. \quad (33)$$

Relation (33) is essential for controlling the error quantities arising in the analysis of the method.

D.4 UPPER BOUND ON Γ_k

For the convenience of notation, we state $L(\delta) = L$ and $\bar{L}(\delta) = \bar{L}$ everywhere in the analysis.

Lemma D.4. *Let Assumptions 2.1 and 2.2 hold. Let $\widehat{D} > 0$ be a chosen constant. Define $\bar{L} = L$ for Algorithm 1 and $\bar{L} = L\sqrt{1 + 3D^2/\widehat{D}^2}$ for Algorithm 2. Then the iterates of Rules 1 and 2 (a.k.a. Algorithm 1) and Rule 3 (a.k.a. Algorithm 2) satisfy the following for any constant $\bar{H} \geq \sqrt{2}\bar{L}$, any $x \in X$, and any $k \geq 0$:*

$$\Gamma_k \leq \frac{\left[\min\{H_{k+1}, \bar{H}\} - H_k\right]_+}{2}(c_1 D^2 + c_2 \widehat{D}^2) + \frac{4L\delta + \|Z_{k+1} - Y_k\|_*^2}{\bar{H}\widehat{D}^2}(3D^2 + \widehat{D}^2),$$

where $c_1 = 3$, $c_2 = 1$ for Algorithm 1 and $c_1 = 3$, $c_2 = 2$ for Algorithm 2.

Proof. We consider the quantity Γ_k (cf. relation (11)) defined as

$$\Gamma_k \equiv \Delta_k(H_{k+1}) + \frac{3}{2}(H_{k+1} - H_k)D^2, \quad (34)$$

where $\Delta_k(H)$ is defined in Lemma D.1. We choose a constant $\bar{H} > 0$ which can be chosen arbitrarily. We consider two cases. For each case, we analyze each example of Algorithm 1 and Algorithm 2 separately.

Case $H_{k+1} \leq \bar{H}$: For Rule 1 (cf. relation (13)), we obtain

$$\begin{aligned}\Gamma_k &\leq [\Delta_k(H_{k+1})]_+ + \frac{3}{2}(H_{k+1} - H_k)D^2 = \frac{H_{k+1} - H_k}{2}(3D^2 + \widehat{D}^2) \\ &= \frac{\left[\min\{H_{k+1}, \bar{H}\} - H_k\right]_+}{2}(3D^2 + \widehat{D}^2). \quad (35)\end{aligned}$$

For Rule 2 (cf. relation (14)), we apply the fact that $\Delta_k(H_{k+1}) \leq \xi_k' - \frac{H_{k+1}}{2}r_k^2$, where r_k^2 is defined in Lemma D.1, and obtain

$$\Gamma_k \leq \left[\xi_k' - \frac{H_{k+1}}{2}r_k^2\right]_+ + \frac{3}{2}(H_{k+1} - H_k)D^2 = \frac{H_{k+1} - H_k}{2}(3D^2 + \widehat{D}^2).$$

Following similar lines of argument as above, we obtain relation (35).

Next, we analyze Rule 3 (a.k.a. Algorithm 2). Application of the upper bounds of relation (24) in Lemma D.1 to relation (34) yields

$$\Gamma_k \leq \bar{\Delta}_k(H_{k+1}) + \frac{3}{2}(H_{k+1} - H_k)D^2, \tag{36}$$

which using the definition of $\bar{\Delta}_k(H_{k+1})$ can be algebraically simplified as

$$\begin{aligned}
\Gamma_k &\leq \frac{\|\widehat{F}(x_{k+1}, \omega_{k+1}) - \widehat{F}(v_k, \omega_k)\|_*^2}{2H_{k+1}} - \frac{H_{k+1}}{2}\|v_k - x_{k+1}\|^2 + \frac{3}{2}(H_{k+1} - H_k)D^2 \\
&\leq \frac{H_{k+1} - H_k}{2}(3D^2 + 2\widehat{D}^2) - \frac{H_{k+1}}{2}\|v_k - x_{k+1}\|^2 \qquad (37) \\
&\leq \frac{\left[\min\{H_{k+1}, \bar{H}\} - H_k\right]_+}{2}(3D^2 + 2\widehat{D}^2). \qquad (38)
\end{aligned}$$

where relation (37) is obtained by applying relation (33).

Case $H_{k+1} \geq \bar{H}$: For this case, relation (36) is satisfied. Since $H_{k+1} \geq \bar{H}$, hence $\bar{\Delta}_k(H_{k+1}) \leq \bar{\Delta}_k(\bar{H})$ and we obtain

$$\Gamma_k \leq \bar{\Delta}_k(\bar{H}) + \frac{3}{2}(H_{k+1} - H_k)D^2 = \bar{\Delta}_k(\bar{H}) + \left[\frac{H_{k+1} - H_k}{2}\widehat{D}^2\right]\frac{3D^2}{\widehat{D}^2}. \tag{39}$$

Applying the upper bound relation (24) of Lemma D.1 and the fact that $H_{k+1} \geq \bar{H}$, for both Rules 1 and 2 of Algorithm 1, we have

$$\frac{H_{k+1} - H_k}{2}\widehat{D}^2 \leq [\bar{\Delta}_k(H_{k+1})]_+ \leq [\bar{\Delta}_k(\bar{H})]_+.$$

Hence, substituting the preceding relation back to relation (39) yields

$$\Gamma_k \leq [\bar{\Delta}_k(\bar{H})]_+ \left[1 + \frac{3D^2}{\widehat{D}^2}\right]. \tag{40}$$

Next, we show the upper bound of $\bar{\Delta}_k(\bar{H})$. We have

$$\begin{aligned}
\bar{\Delta}_k(\bar{H}) &= \frac{1}{2\bar{H}}\|\widehat{F}(x_{k+1}, \omega_{k+1}) - \widehat{F}(v_k, \omega_k)\|_*^2 - \frac{\bar{H}}{2}\|x_{k+1} - v_k\|^2 \\
&= \frac{1}{2\bar{H}}\|[F(x_{k+1}) - F(v_k)] + [Z_{k+1} - Y_k]\|_*^2 - \frac{\bar{H}}{2}\|x_{k+1} - v_k\|^2 \\
&\leq \frac{1}{\bar{H}}\|F(x_{k+1}) - F(v_k)\|_*^2 - \frac{\bar{H}}{2}\|x_{k+1} - v_k\|^2 + \frac{1}{\bar{H}}\|Z_{k+1} - Y_k\|_*^2,
\end{aligned}$$

where the stochastic errors Y_k and Z_k are defined in relation (12). Applying Assumption 2.2 to the preceding relation, we obtain

$$\bar{\Delta}_k(\bar{H}) \leq \left[\frac{L^2}{\bar{H}} - \frac{\bar{H}}{2}\right]\|x_{k+1} - v_k\|^2 + \frac{4L\delta}{\bar{H}} + \frac{1}{\bar{H}}\|Z_{k+1} - Y_k\|_*^2.$$

With the choice of $\bar{H} \geq \sqrt{2}L$, we obtain

$$\bar{\Delta}_k(\bar{H}) \leq \frac{4L\delta}{\bar{H}} + \frac{1}{\bar{H}}\|Z_{k+1} - Y_k\|_*^2.$$

Substituting the preceding relation to relation (40), we obtain

$$\Gamma_k \leq \frac{4L\delta + \|Z_{k+1} - Y_k\|_*^2}{\bar{H}\widehat{D}^2}(3D^2 + \widehat{D}^2). \tag{41}$$

For Rule 3 (Algorithm 2), we apply relation (33) to relation (39) to obtain the upper bound

$$\begin{aligned}
\Gamma_k &\leq \bar{\Delta}_k(\bar{H}) + \frac{\|\widehat{F}(x_{k+1}, \omega_{k+1}) - \widehat{F}(v_k, \omega_k)\|_*^2}{2H_{k+1}}\frac{3D^2}{\widehat{D}^2} \\
&\leq \frac{1}{2\bar{H}}\left[1 + \frac{3D^2}{\widehat{D}^2}\right]\|\widehat{F}(x_{k+1}, \omega_{k+1}) - \widehat{F}(v_k, \omega_k)\|_*^2 - \frac{\bar{H}}{2}\|v_k - x_{k+1}\|^2 \\
&\leq \left[\frac{L^2}{\bar{H}}\left[1 + \frac{3D^2}{\widehat{D}^2}\right] - \frac{\bar{H}}{2}\right]\|x_{k+1} - v_k\|^2 + \frac{4L\delta + \|Z_{k+1} - Y_k\|_*^2}{\bar{H}\widehat{D}^2}(3D^2 + \widehat{D}^2).
\end{aligned}$$

We choose $\bar{H} \geq \sqrt{2}L\left[\frac{\sqrt{3D^2+\widehat{D}^2}}{\widehat{D}}\right]$, and see that for Rule 3 also relation (41) is satisfied. For notational simplicity, we define the constant $\bar{D}_2^2 = 3D^2 + \widehat{D}^2$. The constants D_1 and $\bar{L}$ are defined as

$$D_1^2 = \begin{cases} \bar{D}_2^2 & \text{for Algorithm 1,} \\ \bar{D}_2^2 + \widehat{D}^2 & \text{for Algorithm 2.} \end{cases} \qquad \text{and} \qquad \bar{L} = \begin{cases} L & \text{for Algorithm 1,} \\ L\bar{D}_2/\widehat{D} & \text{for Algorithm 2.} \end{cases}$$

Now, we combine both cases $H_{k+1} \leq \bar{H}$ and $H_{k+1} \geq \bar{H}$ for both Rules 1 and 2 of Algorithm 1 (cf. relations (35) and (41)) and Rule 3 of Algorithm 2 (cf. relations (38) and (41)) to obtain the following for any $\bar{H} \geq \sqrt{2}\bar{L}$, i.e.,

$$\begin{aligned} \Gamma_k &\leq \max\left\{\frac{\left[\min\{H_{k+1}, \bar{H}\} - H_k\right]_+}{2}\bar{D}_1^2, \frac{4L\delta + \|Z_{k+1} - Y_k\|_*^2}{\bar{H}\widehat{D}^2}\bar{D}_2^2\right\} \\ &\leq \frac{\left[\min\{H_{k+1}, \bar{H}\} - H_k\right]_+}{2}\bar{D}_1^2 + \frac{4L\delta + \|Z_{k+1} - Y_k\|_*^2}{\bar{H}\widehat{D}^2}\bar{D}_2^2. \end{aligned}$$

Finally, substituting the values of $\bar{D}_1$ and $\bar{D}_2$ concludes the proof. **Q.E.D.**

E Proofs for Section 4

Throughout this section, for any fixed $\delta > 0$, we write $L(\delta) = L$ and $\bar{L}(\delta) = \bar{L}$ for notational simplicity. For the algorithmic analysis, we encode the stochastic history generated by EGM (cf. relation (8)) up to iteration k in the sigma-algebra $\mathcal{F}_k$, defined as follows:

$$\mathcal{F}_k = \{v_0\} \cup \{\omega_t \mid 0 \leq t \leq k\} \qquad \forall k \geq 0.$$

E.1 Proof of Theorem 4.1

Proof. Let $x \in X$ and $T \geq 1$ be arbitrary. Let us denote

$$\bar{D}_1^2 = c_1 D^2 + c_2 \widehat{D}^2 \qquad \text{and} \qquad \bar{D}_2^2 = 3D^2 + \widehat{D}^2. \tag{42}$$

Applying relation (11) and (12) to relation (22) of Lemma D.1, we obtain

$$\begin{aligned} \langle F(x_{k+1}), x_{k+1} - x\rangle \leq& \frac{\left[\min\{H_{k+1}, \bar{H}\} - H_k\right]_+}{2}\bar{D}_1^2 + \langle Z_{k+1}, h_k - x_{k+1}\rangle \\ &+ \frac{4L\delta + \|Z_{k+1} - Y_k\|_*^2}{\bar{H}\widehat{D}^2}\bar{D}_2^2 + \frac{\|Z_{k+1}\|_*^2}{2\eta} \\ &+ \frac{\eta}{2}[\|x - h_k\|^2 - \|x - h_{k+1}\|^2] \\ &+ \frac{H_k}{2}\|x - v_k\|^2 - \frac{H_{k+1}}{2}\|x - v_{k+1}\|^2. \end{aligned}$$

We sum both sides of the preceding relation from $k = 0$ to $T - 1$ and divide both sides by T to obtain

$$\begin{aligned} \frac{1}{T}\sum_{k=0}^{T-1}\langle F(x_{k+1}), x_{k+1} - x\rangle \leq& \frac{\bar{D}_1^2}{2T}\left[\min\{H_T, \bar{H}\} - H_0\right]_+ + \frac{1}{T}\sum_{k=0}^{T-1}\langle Z_{k+1}, h_k - x_{k+1}\rangle \\ &+ \frac{1}{2\bar{H}}\left[\frac{8L\delta\bar{D}_2^2}{\widehat{D}^2} + \frac{1}{T}\sum_{k=0}^{T-1}\frac{2\bar{D}_2^2}{\widehat{D}^2}\|Z_{k+1} - Y_k\|_*^2\right] \\ &+ \frac{1}{2\eta T}\sum_{k=0}^{T-1}\|Z_{k+1}\|_*^2 + \frac{\eta}{2T}\left(\|x - h_0\|^2 - \|x - h_T\|^2\right) \\ &+ \frac{H_0}{2T}\|x - v_0\|^2 - \frac{H_T}{2T}\|x - v_T\|^2, \end{aligned} \tag{43}$$

where Lemma B.1 is applied to simplify the first quantity on the right-hand side. Recall that we choose $H_0 = 0$ for all the algorithms. Hence,

$$\big[\min\{H_T, \bar{H}\} - H_0\big]_+ = \min\{H_T, \bar{H}\} \leq \bar{H}.$$

We take maximum with respect to $x \in X$ on both sides of relation (43) and apply the preceding relation and Assumption 2.1 to the quantity $\max_{x\in X} \|x - h_0\|$ to obtain

$$\epsilon_T \leq \frac{\bar{H}\bar{D}_1^2}{2T} + \frac{1}{T}\sum_{k=0}^{T-1}\langle Z_{k+1}, h_k - x_{k+1}\rangle + \frac{1}{2\bar{H}}\left[\frac{8L\delta\bar{D}_2^2}{\widehat{D}^2} + \frac{1}{T}\sum_{k=0}^{T-1}\frac{2\bar{D}_2^2}{\widehat{D}^2}\|Z_{k+1} - Y_k\|_*^2\right] + \frac{1}{2\eta T}\sum_{k=0}^{T-1}\|Z_{k+1}\|_*^2 + \frac{\eta D^2}{2T},$$

where ϵ_T is the residual function given in relation (7). Next, we take total expectation on both sides of the preceding relation to obtain

$$\begin{aligned}\mathbb{E}[\epsilon_T] \leq \frac{\bar{H}\bar{D}_1^2}{2T} &+ \frac{1}{T}\mathbb{E}\left[\sum_{k=0}^{T-1}\langle Z_{k+1}, h_k - x_{k+1}\rangle\right] \\ &+ \frac{1}{2\bar{H}}\left[\frac{8L\delta\bar{D}_2^2}{\widehat{D}^2} + \frac{1}{T}\mathbb{E}\left[\sum_{k=0}^{T-1}\frac{2\bar{D}_2^2}{\widehat{D}^2}\|Z_{k+1} - Y_k\|_*^2\right]\right] \\ &+ \frac{1}{2\eta T}\mathbb{E}\left[\sum_{k=0}^{T-1}\|Z_{k+1}\|_*^2\right] + \frac{\eta D^2}{2T}. \end{aligned} \tag{44}$$

The second quantity on the right-hand side of relation (44) can be simplified as

$$\mathbb{E}\left[\sum_{k=0}^{T-1}\langle Z_{k+1}, h_k - x_{k+1}\rangle\right] = \mathbb{E}\left[\sum_{k=0}^{T-1}\langle \mathbb{E}[Z_{k+1} \mid \mathcal{F}_k], h_k - x_{k+1}\rangle\right] = 0,$$

where we used the unbiasedness property of Assumption 3.1. The quantity $\mathbb{E}\big[\sum_{k=0}^{T}\|Z_{k+1} - Y_k\|_*^2\big]$ can be simplified as

$$\begin{aligned}\mathbb{E}\left[\sum_{k=0}^{T-1}\|Z_{k+1} - Y_k\|_*^2\right] &= \mathbb{E}\left[\sum_{k=0}^{T-1}\left(\|Z_{k+1}\|_*^2 + \|Y_k\|_*^2 - 2\langle \mathbb{E}[Z_{k+1} \mid \mathcal{F}_k], Y_k\rangle\right)\right] \\ &= \mathbb{E}\left[\sum_{k=0}^{T-1}\left(\|Z_{k+1}\|_*^2 + \|Y_k\|_*^2\right)\right]. \end{aligned} \tag{45}$$

Hence applying the preceding two relations back to relation (44) and applying the variance bound in Assumption 3.1, we obtain

$$\mathbb{E}[\epsilon_T] \leq \frac{\bar{H}\bar{D}_1^2}{2T} + \frac{\theta}{2\bar{H}} + \frac{\eta D^2}{2T} + \frac{\sigma^2}{2\eta}, \tag{46}$$

where for the sake of simplicity, we denote

$$\theta := \frac{8L\delta\bar{D}_2^2}{\widehat{D}^2} + \frac{4\sigma^2\bar{D}_2^2}{\widehat{D}^2}. \tag{47}$$

Next, we minimize the right-hand side of the relation (46) with respect to both $\bar{H} \geq \sqrt{2}\bar{L}$ and $\eta > 0$ to obtain the respective optimal values as

$$\bar{H}^* = \max\left\{\sqrt{2}\,\bar{L}, \sqrt{\frac{T\theta}{\bar{D}_1^2}}\right\}, \qquad \eta^* = \frac{\sigma\sqrt{T}}{D}.$$

Substituting $\bar{H} = \bar{H}^*$ and $\eta = \eta^*$ back to relation (46), we obtain

$$\begin{aligned}\mathbb{E}[\epsilon_T] &\leq \max\left\{\frac{\bar{L}\bar{D}_1^2}{\sqrt{2}T}, \sqrt{\frac{\bar{D}_1^2\theta}{4T}}\right\} + \min\left\{\frac{\theta}{2\sqrt{2}\bar{L}}, \sqrt{\frac{\bar{D}_1^2\theta}{4T}}\right\} + \frac{D\sigma}{\sqrt{T}} \\ &\leq \frac{\bar{L}\bar{D}_1^2}{\sqrt{2}T} + 2\sqrt{\frac{\bar{D}_1^2\theta}{4T}} + \frac{D\sigma}{\sqrt{T}} \\ &= \frac{\bar{L}\bar{D}_1^2}{\sqrt{2}T} + \sqrt{\frac{\bar{D}_1^2\theta}{T}} + \frac{D\sigma}{\sqrt{T}}.\end{aligned}$$

Substituting the value of θ from relation (47), we obtain the final relation of the theorem. **Q.E.D.**

E.2 Proof of Corollary 4.2

Proof. We start from the relation of Theorem 4.1 and use $\bar{L} = L\left(\frac{\bar{D}_2}{\widehat{D}}\right)^a = L(\delta)\left(\frac{\bar{D}_2}{\widehat{D}}\right)^a$, where $a = 0$ for Algorithm 1 and $a = 1$ for Algorithm 2, and the quantity $L(\delta)$ is derived in Lemma 2.3, and obtain

$$\begin{aligned}\mathbb{E}[\epsilon_T] \leq{}& \frac{\bar{D}_1^2}{\sqrt{2}T}\left(\frac{\bar{D}_2}{\widehat{D}}\right)^a \left(\frac{1-\nu}{4}\right)^{\frac{1-\nu}{1+\nu}} \nu^{\frac{\nu}{1+\nu}} H_\nu^{\frac{2}{1+\nu}} \delta^{-\frac{1-\nu}{1+\nu}} \\ &+ \sqrt{\frac{\bar{D}_1^2}{T}}\left[\frac{2\sqrt{2}\,\bar{D}_2}{\widehat{D}}\left(\frac{1-\nu}{4}\right)^{\frac{1-\nu}{2(1+\nu)}} \nu^{\frac{\nu}{2(1+\nu)}} H_\nu^{\frac{1}{1+\nu}} \delta^{\frac{\nu}{1+\nu}} + \frac{2\bar{D}_2}{\widehat{D}}\sigma\right] + \frac{D\sigma}{\sqrt{T}} \qquad (48)\end{aligned}$$

for all $\nu \in [0, 1]$ with the convention that $0^0 = 1$, where $D_1^2 = c_1 D^2 + c_2\widehat{D}^2$ and $D_2^2 = 3D^2 + \widehat{D}^2$. Since $L(\delta)$ in Lemma 2.3 holds for any $\delta > 0$, we minimize relation (48) with respect to δ and obtain the optimal for any $\nu \in (0, 1)$ as

$$\delta^* = \left[\frac{1-\nu}{4\nu}\left(\frac{\widehat{D}}{\bar{D}_2}\right)^{1-a}\sqrt{\frac{\bar{D}_1^2}{T}}\right]^{1+\nu}\left(\frac{1-\nu}{4}\right)^{\frac{1-\nu}{2}}\nu^{\frac{\nu}{2}}H_\nu.$$

Substituting $\delta = \delta^*$ back to relation (48) yields the following for any $\nu \in (0, 1)$:

$$\begin{aligned}\mathbb{E}[\epsilon_T] &\leq \frac{2^{\frac{1}{2}-\nu}H_\nu}{(1-\nu)^{\frac{1-\nu}{2}}\nu^{\frac{\nu}{2}}}\left(\frac{\bar{D}_2}{\widehat{D}}\right)^{1+\nu(a-1)}\left(\frac{\bar{D}_1^2}{T}\right)^{\frac{1+\nu}{2}} + \sqrt{\frac{\bar{D}_1^2}{T}}\frac{2\bar{D}_2}{\widehat{D}}\sigma + \frac{D\sigma}{\sqrt{T}} \\ &= \frac{2^{\frac{1}{2}-\nu}H_\nu}{(1-\nu)^{\frac{1-\nu}{2}}\nu^{\frac{\nu}{2}}}\left(\frac{\bar{D}_2}{\widehat{D}}\right)^{1+\nu(a-1)}\left(\frac{\bar{D}_1^2}{T}\right)^{\frac{1+\nu}{2}} + \frac{\sigma}{\sqrt{T}}\left(\frac{2\bar{D}_1\bar{D}_2}{\widehat{D}} + D\right). \qquad (49)\end{aligned}$$

Next, we consider the corner cases. When $\nu = 0$, relation (48) yields

$$\begin{aligned}\mathbb{E}[\epsilon_T] &\leq \inf_{\delta>0}\left\{\frac{\bar{D}_1^2}{4\sqrt{2}T}\left(\frac{\bar{D}_2}{\widehat{D}}\right)^a H_0^2\delta^{-1} + \sqrt{\frac{\bar{D}_1^2}{T}}\left[\frac{\sqrt{2}\,\bar{D}_2}{\widehat{D}}H_0 + \frac{2\bar{D}_2}{\widehat{D}}\sigma\right]\right\} + \frac{D\sigma}{\sqrt{T}} \\ &= \lim_{\delta\to\infty}\left\{\frac{\bar{D}_1^2}{4\sqrt{2}T}\left(\frac{\bar{D}_2}{\widehat{D}}\right)^a H_0^2\delta^{-1}\right\} + \sqrt{\frac{\bar{D}_1^2}{T}}\left[\frac{\sqrt{2}\,\bar{D}_2}{\widehat{D}}H_0 + \frac{2\bar{D}_2}{\widehat{D}}\sigma\right] + \frac{D\sigma}{\sqrt{T}} \\ &= \sqrt{\frac{\bar{D}_1^2}{T}}\left[\frac{\sqrt{2}\,\bar{D}_2}{\widehat{D}}H_0 + \frac{2\bar{D}_2}{\widehat{D}}\sigma\right] + \frac{D\sigma}{\sqrt{T}} \\ &= \frac{\sqrt{2}\,H_0\bar{D}_1\bar{D}_2}{\widehat{D}\sqrt{T}} + \frac{\sigma}{\sqrt{T}}\left(\frac{2\bar{D}_1\bar{D}_2}{\widehat{D}} + D\right).\end{aligned}$$

On the other hand, when $\nu = 1$, relation (48) simplifies to

$$\begin{aligned}\mathbb{E}[\epsilon_T] &\le \inf_{\delta>0}\left\{\frac{\bar{D}_1^2}{\sqrt{2}T}\left(\frac{\bar{D}_2}{\widehat{D}}\right)^a H_1 + \sqrt{\frac{\bar{D}_1^2}{T}}\left[\frac{2\sqrt{2}\,\bar{D}_2}{\widehat{D}}\sqrt{\delta H_1} + \frac{2\bar{D}_2}{\widehat{D}}\sigma\right]\right\} + \frac{D\sigma}{\sqrt{T}} \\ &= \frac{\bar{D}_1^2}{\sqrt{2}T}\left(\frac{\bar{D}_2}{\widehat{D}}\right)^a H_1 + \sqrt{\frac{\bar{D}_1^2}{T}}\frac{2\bar{D}_2}{\widehat{D}}\sigma + \frac{D\sigma}{\sqrt{T}} + \lim_{\delta\to 0}\left\{\sqrt{\frac{\bar{D}_1^2}{T}}\frac{2\sqrt{2}\,\bar{D}_2}{\widehat{D}}\sqrt{\delta H_1}\right\} \\ &= \frac{\bar{D}_1^2}{\sqrt{2}T}\left(\frac{\bar{D}_2}{\widehat{D}}\right)^a H_1 + \frac{\sigma}{\sqrt{T}}\left(\frac{2\bar{D}_1\bar{D}_2}{\widehat{D}} + D\right).\end{aligned}$$

Hence, relation (49) holds for any $\nu \in [0,1]$, with the convention $0^0 = 1$. **Q.E.D.**

F Missing details of Section 5

F.1 Algorithmic Development for Distance Adaptation

We consider the modified ExtraGradient Method, aka. modified EGM, which we present again here. The method starts with the initialization $v_0 \in X$ and $H_0 = 0$, and update for all $k \ge 0$ as

$$\begin{aligned}x_{k+1} &= \arg\min_{x\in X}\left\{D_k\langle\widehat{F}(v_k,\omega_k),\, x - v_k\rangle + \frac{H_k}{2}\|x - v_k\|^2\right\}, \\ v_{k+1} &= \arg\min_{x\in X}\left\{D_k\langle\widehat{F}(x_{k+1},\omega_{k+1}),\, x - x_{k+1}\rangle + \frac{H_k}{2}\|x - v_k\|^2\right\},\end{aligned} \tag{50}$$

where D_k is the adaptive diameter and is updated for all $k \ge 0$ as

$$D_{k+1} = \max\{D_k, \|v_{k+1} - v_0\|, r_k\} \qquad \text{with} \quad D_0 = \widehat{D} > 0.$$

We present a lemma analogous to Lemma D.1 from where the error balance equations will be derived.

Lemma F.1. *For the modified EGM, the following holds for any $x \in X$ and any $k \ge 0$:*

$$\begin{aligned}D_k\langle F(x_{k+1}), x_{k+1} - x\rangle &\le \widehat{\Delta}_k(H_{k+1}) + \frac{H_{k+1} - H_k}{2}r_k^2 + D_k\langle Z_{k+1}, h_k - x_{k+1}\rangle \\ &+ \frac{D_k\|Z_{k+1}\|_*^2}{2\eta} + \frac{\eta D_k}{2}[\|x - h_k\|^2 - \|x - h_{k+1}\|^2] + \frac{H_k}{2}[\|x - v_k\|^2 - \|x - v_{k+1}\|^2],\end{aligned}$$

where the sequence $\{h_k\}$ is defined in Lemma D.1, $\eta > 0$ is any constant, Z_k is the stochastic error (cf. relation (9)), D_k is the adaptive diameter, and $\widehat{\Delta}_k(H) := \widehat{\xi}_k - \frac{H}{2}r_k^2$, where

$$\begin{aligned}r_k^2 &= \|v_k - x_{k+1}\|^2 + \|v_{k+1} - x_{k+1}\|^2 \\ \widehat{\xi}_k &= D_k\langle\widehat{F}(x_{k+1},\omega_{k+1}), x_{k+1} - v_{k+1}\rangle - \frac{H_k}{2}\|v_k - v_{k+1}\|^2 + \frac{H_k}{2}r_k^2 \\ &\le D_k\langle\widehat{F}(x_{k+1},\omega_{k+1}) - \widehat{F}(v_k,\omega_k), x_{k+1} - v_{k+1}\rangle \equiv \widehat{\xi}_k'.\end{aligned}$$

Moreover,

$$\widehat{\Delta}_k(H) \le \widehat{\xi}_k' - \frac{H}{2}r_k^2 \le \frac{D_k^2}{2H}\|\widehat{F}(x_{k+1},\omega_{k+1}) - \widehat{F}(v_k,\omega_k)\|_*^2 - \frac{H}{2}\|x_{k+1} - v_k\|^2 \equiv \widetilde{\Delta}_k(H). \tag{51}$$

Proof. Similar to the start of Lemma D.1, we apply H_k strong convexity of v_{k+1} update and simplify terms to obtain

$$\begin{aligned}D_k\langle\widehat{F}(x_{k+1},\omega_{k+1}), x_{k+1} - x\rangle &\le \widehat{\xi}_k - \frac{H_{k+1}}{2}r_k^2 + \frac{H_{k+1} - H_k}{2}r_k^2 \\ &+ \frac{H_k}{2}[\|x - v_k\|^2 - \|x - v_{k+1}\|^2],\end{aligned} \tag{52}$$

where $r_k^2 = \|v_k - x_{k+1}\|^2 + \|v_{k+1} - x_{k+1}\|^2$ and

$$\widehat{\xi}_k = D_k\langle \widehat{F}(x_{k+1}, \omega_{k+1}), x_{k+1} - v_{k+1}\rangle - \frac{H_k}{2}\|v_k - v_{k+1}\|^2 + \frac{H_k}{2} r_k^2.$$

We denote

$$\widehat{\Delta}_k(H_{k+1}) := \widehat{\xi}_k - \frac{H_{k+1}}{2} r_k^2.$$

The rest of the proof follows along lines similar to that of Lemma D.1, and applying Lemma D.2 to simplify the left-hand side of relation (52), we can finally obtain

$$\begin{aligned} D_k\langle F(x_{k+1}), x_{k+1} - x\rangle \leq \widehat{\Delta}_k(H_{k+1}) + \frac{H_{k+1} - H_k}{2} r_k^2 + D_k\langle Z_{k+1}, h_k - x_{k+1}\rangle \\ + \frac{D_k\|Z_{k+1}\|_*^2}{2\eta} + \frac{\eta D_k}{2}[\|x - h_k\|^2 - \|x - h_{k+1}\|^2] \\ + \frac{H_k}{2}[\|x - v_k\|^2 - \|x - v_{k+1}\|^2]. \end{aligned}$$

Moreover from H_k strong convexity of x_{k+1} update, we can obtain

$$D_k\left\langle \widehat{F}(v_k, \omega_k), x_{k+1} - v_{k+1}\right\rangle \leq \frac{H_k}{2}\|v_k - v_{k+1}\|^2 - \frac{H_k}{2} r_k^2.$$

Applying the preceding relation to upper bound $\widehat{\Delta}_k(H_{k+1})$, we obtain

$$\begin{aligned} \widehat{\Delta}_k(H_{k+1}) &\leq \underbrace{D_k\langle \widehat{F}(x_{k+1}, \omega_{k+1}) - \widehat{F}(v_k, \omega_k), x_{k+1} - v_{k+1}\rangle}_{\widehat{\xi}'_k} - \frac{H_{k+1}}{2} r_k^2 \\ &\leq \underbrace{\frac{D_k^2}{2H_{k+1}}\|\widehat{F}(x_{k+1}, \omega_{k+1}) - \widehat{F}(v_k, \omega_k)\|_*^2 - \frac{H_{k+1}}{2}\|x_{k+1} - v_k\|^2}_{\widetilde{\Delta}_k(H_{k+1})}. \end{aligned}$$

This concludes the proof. **Q.E.D.**

Similar to Section 3, here we also design balance equations to update the sequence H_k and control the error. We present two such examples below:

$$\frac{H_{k+1} - H_k}{2} D_{k+1}^2 = \left[\widehat{\Delta}_k(H_{k+1})\right]_+ \qquad \text{(Method 1)}, \tag{53}$$

$$\frac{H_{k+1} - H_k}{2} D_{k+1}^2 = \left[\widehat{\xi}'_k - \frac{H_{k+1}}{2} r_k^2\right]_+ \qquad \text{(Method 2)}. \tag{54}$$

Since the balance equations in Section 3 are labeled as "Rule", we use "Method" with the corresponding number here to distinguish. Applying Lemma D.3, a unique solution of H_{k+1} for both the methods can be obtained, i.e., for all $k \geq 0$,

$$\begin{aligned} H_{k+1} &= H_k + \frac{2\left[\widehat{\xi}_k - \frac{H_k}{2} r_k^2\right]_+}{\widehat{R}_k^2} \\ &= H_k + \frac{2\left[D_k\langle \widehat{F}(x_{k+1}, \omega_{k+1}), x_{k+1} - v_{k+1}\rangle - \frac{H_k}{2}\|v_k - v_{k+1}\|^2\right]_+}{\widehat{R}_k^2} \qquad \text{(Method 1)} \end{aligned}$$

$$\text{and } H_{k+1} = H_k + \frac{2\left[\widehat{\xi}'_k - \frac{H_k}{2} r_k^2\right]_+}{\widehat{R}_k^2} \qquad \text{(Method 2)},$$

where the quantity $\widehat{R}_k$ is defined as

$$\widehat{R}_k^2 = D_{k+1}^2 + r_k^2 = D_{k+1}^2 + \|v_k - x_{k+1}\|^2 + \|v_{k+1} - x_{k+1}\|^2. \tag{55}$$

Algorithm 3 Line search-free modified EGM using Distance Adaptation

1: **Input:** $v_0 \in X$, $H_0 = 0$, small estimate $D_0 = \widehat{D} > 0$
2: **for** $k = 0, 1, 2, \ldots, T-1$ **do**
3: Compute x_{k+1} and v_{k+1} according to modified EGM (relation (50))
4: Compute $D_{k+1} = \max\{D_k, \|v_{k+1} - v_0\|, r_k\}$
5: Compute $\widehat{R}_k$ as in relation (55)
6: $H_{k+1} = H_k + \frac{2\left[\bar{\xi}_k - \frac{H_k}{2} r_k^2\right]_+}{\widehat{R}_k^2}$ where $\bar{\xi}_k = \widehat{\xi}_k$ or $\widehat{\xi}'_k$ as in Lemma F.1
7: **end for**

We present the pseudocode of both methods in Algorithm 3. Both balance equations (53) and (54) yield an upper bound on H_k for all $k \geq 0$, which we present next.

Note that the effective step size of Algorithm 3 is D_k/H_k for all $k \geq 1$. There is a beautiful analogy between Rules 1 and 2 of Algorithm 1 and Methods 1 and 2 of Algorithm 3. Let us denote $M_k = H_k D_k$. Then, the balance equations (53) and (54) can be written as

$$\frac{M_{k+1}D_{k+1} - M_k D_k}{2} D_{k+1}^2 = \left[\widehat{\Delta}_k (M_{k+1} D_{k+1})\right]_+ \qquad \text{(Method 1)},$$

$$\frac{M_{k+1}D_{k+1} - M_k D_k}{2} D_{k+1}^2 = \left[\widehat{\xi}'_k - \frac{M_{k+1}D_{k+1}}{2} r_k^2\right]_+ \qquad \text{(Method 2)}.$$

In the special case where $D_k = \widehat{D}$ for all $k \geq 0$, Methods 1 and 2 reduce to Rules 1 and 2, respectively, of Algorithm 1, with M_k corresponding to H_k. Next, we present a lemma to obtain the upper bound on H_k for Algorithm 3.

Lemma F.2. *Under Assumptions 2.1 and 2.2, and with the initialization $H_0 = 0$, the balance equations of Methods 1 and 2 (cf. relations* (53) *and* (54)*) satisfy the following for any $k \geq 1$:*

$$H_k \leq \sqrt{2} L(\delta) D + 4\sqrt{L(\delta)\delta k} + 2\sqrt{\sum_{i=0}^{k-1} \|Z_{i+1} - Y_i\|_*^2}.$$

Proof. Applying relation (51) of Lemma F.1, for both Methods 1 and 2, we have

$$\begin{aligned}
\frac{H_{k+1} - H_k}{2} D_{k+1}^2 &\leq \left[\frac{D_k^2}{2H_{k+1}} \|\widehat{F}(x_{k+1}, \omega_{k+1}) - \widehat{F}(v_k, \omega_k)\|_*^2 - \frac{H_{k+1}}{2}\|x_{k+1} - v_k\|^2\right]_+ \\
&\leq \left[\frac{D_k^2}{H_{k+1}} \|F(x_{k+1}) - F(v_k)\|_*^2 - \frac{H_{k+1}}{2}\|x_{k+1} - v_k\|^2\right]_+ \\
&\qquad + \frac{D_k^2}{H_{k+1}} \|Z_{k+1} - Y_k\|_*^2,
\end{aligned}$$

where the quantity Z_{k+1} and Y_k are the stochastic errors and are defined in relation (9). Now application of Assumption 2.2 on the preceding relation yields

$$\frac{H_{k+1} - H_k}{2} D_{k+1}^2 \leq \left[\frac{D_k^2}{H_{k+1}} L^2 - \frac{H_{k+1}}{2}\right]_+ \|x_{k+1} - v_k\|^2 + \frac{4L\delta D_k^2}{H_{k+1}} + \frac{D_k^2}{H_{k+1}} \|Z_{k+1} - Y_k\|_*^2,$$

where for notational simplicity, we denoted $L(\delta) \equiv L$ for any $\delta > 0$. We multiply both sides of the preceding relation by $4H_{k+1}$ and obtain

$$\begin{aligned}
2H_{k+1}(H_{k+1} - H_k) D_{k+1}^2 &\leq 2\left[2D_k^2 L^2 - H_{k+1}^2\right]_+ \|x_{k+1} - v_k\|^2 + 16L\delta D_k^2 + 4D_k^2 \|Z_{k+1} - Y_k\|_*^2 \\
&\leq 2\left[2D^2 L^2 - H_{k+1}^2\right]_+ D_{k+1}^2 + 16L\delta D_k^2 + 4D_k^2 \|Z_{k+1} - Y_k\|_*^2, \quad (56)
\end{aligned}$$

where the last relation is obtained using Assumption 2.1. Note that for any $k \geq 0$,

$$2H_{k+1}(H_{k+1} - H_k) = (H_{k+1} + H_{k+1})(H_{k+1} - H_k) \geq (H_{k+1} + H_k)(H_{k+1} - H_k) = H_{k+1}^2 - H_k^2.$$

Applying the preceding relation to lower bound the left-hand side of relation (56), we obtain

$$(H_{k+1}^2 - H_k^2)D_{k+1}^2 \leq 2\left[2D^2L^2 - H_{k+1}^2\right]_+ D_{k+1}^2 + 16L\delta D_k^2 + 4D_k^2\|Z_{k+1} - Y_k\|_*^2.$$

We divide both sides of the relation by D_{k+1}^2 and using the fact that $D_k^2/D_{k+1}^2 \leq 1$, we obtain

$$H_{k+1}^2 - H_k^2 \leq 2\left[2D^2L^2 - H_{k+1}^2\right]_+ + 16L\delta + 4\|Z_{k+1} - Y_k\|_*^2.$$

Applying Lemma B.2 with $\gamma_k = H_k^2$, $a = 2$, $V = 2L^2D^2$, $e_k = 16L\delta + 4\|Z_{k+1} - Y_k\|_*^2$, and noting that $\gamma_0 = H_0^2 = 0$ (based on the initialization $H_0 = 0$), we have for all $k \geq 1$,

$$H_k^2 \leq 2L^2D^2 + 16L\delta k + 4\sum_{i=0}^{k-1}\|Z_{i+1} - Y_i\|_*^2.$$

Taking square root on both sides of the preceding relation and using sub-additivity of the square root function yield the final relation of the lemma. **Q.E.D.**

Lemma F.2 serves as a counterpart to Lemma D.4. However, their analyses differ. Lemma D.4 also accommodates the classical AdaGrad-norm update (cf. Algorithm 2) within the same universal analysis by introducing the proxy constant $\bar{H}$. In the distance-adaptive setting, however, we no longer consider the classical AdaGrad-norm update, which allows us to improve the constants and considerably simplify the analysis. A similar analysis can also be applied to Algorithm 1.

F.2 Convergence analysis for Distance Adaptive VI algorithm

In this subsection, we present theorems that establish the convergence rate of Algorithm 3 under Assumption 2.2 and also study the rate for Hölder continuous mapping as a special case.

Theorem F.3. *Let Assumptions 2.1, 2.2 and 3.1 hold, and let $\rho = \max\{1, D/\widehat{D}\}$. Then, for any $T \geq 1$, the iterates generated by Algorithm 3 converge as*

$$\mathbb{E}\left[\epsilon_{\tau(T)}\right] \leq (\rho)^{1/T}\log(e\rho)\left[\frac{3\sqrt{2}L(\delta)D^2}{T} + \frac{12D\sqrt{L(\delta)\delta} + (6\sqrt{2}+5)\sigma D}{\sqrt{T}}\right],$$

where $\tau(T) := \arg\min_{1\leq t\leq T}\frac{D_t}{\sum_{k=0}^{t-1}D_k}$ and $\epsilon_{\tau(T)}$ is the corresponding residual function (cf. relation (7)).

Proof. We start from the first relation of Lemma F.1, and see that since $\widehat{\Delta}_k(H) \leq \widehat{\xi}_k' - \frac{H}{2}r_k^2$ from relation (51), hence using the balance equations of both Methods 1 and 2 can yield

$$\begin{aligned}D_k\langle F(x_{k+1}), x_{k+1} - x\rangle \leq{}& \frac{H_{k+1} - H_k}{2}(r_k^2 + D_{k+1}^2) + D_k\langle Z_{k+1}, h_k - x_{k+1}\rangle \\ &+ \frac{D_k\|Z_{k+1}\|_*^2}{2\eta} + \frac{\eta D_k}{2}[\|x - h_k\|^2 - \|x - h_{k+1}\|^2] \\ &+ \frac{H_k}{2}[\|x - v_k\|^2 - \|x - v_{k+1}\|^2].\end{aligned}$$

We apply the fact that $r_k \leq D_{k+1}$ in order to simplify the first quantity on the right-hand side. Then we sum both sides of the preceding relation from $k = 0$ to $T - 1$ for any $T \geq 1$, yielding

$$\begin{aligned}\sum_{k=0}^{T-1}D_k\langle F(x_{k+1}), x_{k+1} - x\rangle \leq{}&(H_T - H_0)D_T^2 + \sum_{k=0}^{T-1}D_k\langle Z_{k+1}, h_k - x_{k+1}\rangle \\ &+ \frac{\sum_{k=0}^{T-1}D_k\|Z_{k+1}\|_*^2}{2\eta} + \frac{\eta D_T}{2}[\|x - h_0\|^2 - \|x - h_T\|^2] \\ &+ \sum_{k=0}^{T-1}\frac{H_k}{2}[\|x - v_k\|^2 - \|x - v_{k+1}\|^2], \qquad (57)\end{aligned}$$

where in some of the terms, we use the fact that $D_k \leq D_T$ for any $0 \leq k \leq T$ and any $T \geq 1$. We initialized $H_0 = 0$. We also denote $d_k = \|x - v_k\|$. Then, the last term on the right-hand side of relation (57) can be simplified as

$$\begin{aligned}\frac{1}{2}\sum_{k=0}^{T-1} H_k[d_k^2 - d_{k+1}^2] &= \frac{1}{2}\left[H_0 d_0^2 - H_{T-1}d_T^2 + \sum_{k=1}^{T-1} d_k^2(H_k - H_{k-1})\right] \\ &= \frac{1}{2}\left[H_0 d_0^2 - H_{T-1}d_T^2 + \bar{d}_T^2(H_{T-1} - H_0)\right] \\ &= \frac{H_{T-1}}{2}(\bar{d}_T^2 - d_T^2) \leq H_T \bar{d}_T(\bar{d}_T - d_T),\end{aligned}$$

where $\bar{d}_T = \max_{1\leq k\leq T-1} d_k$. Using Assumption 2.1, we obtain $\bar{d}_T \leq D$. Let us denote the index $\iota = \arg\max_{1\leq k\leq T-1} d_k$. Then,

$$\bar{d}_T - d_T = d_\iota - d_T \leq \|v_T - v_\iota\| \leq \|v_T - v_0\| + \|v_\iota - v_0\| \leq 2D_T.$$

Hence, we obtain

$$\sum_{k=0}^{T-1} \frac{H_k}{2}[\|x - v_k\|^2 - \|x - v_{k+1}\|^2] \leq 2DD_T H_T.$$

Moreover using Assumption 2.1, the first quantity on the right-hand side of relation (57) can be upper bounded as

$$(H_T - H_0)D_T^2 \leq DD_T H_T.$$

Making use of the preceding two relationships, relation (57) simplifies as

$$\begin{aligned}\sum_{k=0}^{T-1} D_k\langle F(x_{k+1}), x_{k+1} - x\rangle \leq 3DD_T H_T + \sum_{k=0}^{T-1} D_k\langle Z_{k+1}, h_k - x_{k+1}\rangle \\ + \frac{D_T\sum_{k=0}^{T-1}\|Z_{k+1}\|_*^2}{2\eta} + \frac{\eta D_T}{2}\|x - h_0\|^2.\end{aligned}$$

Taking maximum with respect to $x \in X$ and dividing by $\sum_{k=0}^{T-1} D_k$ on both sides and using Assumption 2.1, we obtain the upper bound on the resudial function $\epsilon_T = \epsilon_T((x_t, D_t)_{t=1}^T)$ as

$$\epsilon_T \leq \frac{D_T}{\sum_{k=0}^{T-1} D_k}\left[3DH_T + \frac{\sum_{k=0}^{T-1}\|Z_{k+1}\|_*^2}{2\eta} + \frac{\eta D^2}{2}\right] + \frac{\sum_{k=0}^{T-1} D_k\langle Z_{k+1}, h_k - x_{k+1}\rangle}{\sum_{k=0}^{T-1} D_k}.$$

Next, we substitute the value of H_T from Lemma F.2 in the preceding relation and change the variable name from T to t to obtain

$$\begin{aligned}\epsilon_t \leq \frac{D_t}{\sum_{k=0}^{t-1} D_k}\left[3\sqrt{2}LD^2 + 12D\sqrt{L\delta t} + 6D\sqrt{\sum_{i=0}^{t-1}\|Z_{i+1} - Y_i\|_*^2} + \frac{\sum_{k=0}^{t-1}\|Z_{k+1}\|_*^2}{2\eta} + \frac{\eta D^2}{2}\right] \\ + \frac{\sum_{k=0}^{t-1} D_k\langle Z_{k+1}, h_k - x_{k+1}\rangle}{\sum_{k=0}^{t-1} D_k}, \quad (58)\end{aligned}$$

where for simplicity, we considered $L(\delta) \equiv L$ for any $\delta > 0$. Let us denote

$$\tau \equiv \tau(T) := \arg\min_{1\leq t\leq T} \frac{D_t}{\sum_{k=0}^{t-1} D_k}.$$

For notational simplicity, we will write $\tau(T)$ as τ throughout the rest of the proof. Hence, we see

$$\min_{1\leq t\leq T} \frac{D_t}{\sum_{k=0}^{t-1} D_k} = \frac{D_\tau}{\sum_{k=0}^{\tau-1} D_k}.$$

Note that $1 \leq \tau \leq T$. Then, relation (58) can be simplified as

$$\epsilon_\tau \leq \left[\frac{D_\tau}{\sum_{k=0}^{\tau-1} D_k}\right]\left[3\sqrt{2}LD^2 + 12D\sqrt{L\delta T} + 6D\sqrt{\sum_{i=0}^{T-1}\|Z_{i+1}-Y_i\|_*^2} + \frac{\sum_{k=0}^{T-1}\|Z_{k+1}\|_*^2}{2\eta} + \frac{\eta D^2}{2}\right] + \left[\frac{D_\tau}{\sum_{k=0}^{\tau-1} D_k}\right]\left|\frac{\sum_{k=0}^{\tau-1} D_k\langle Z_{k+1}, h_k - x_{k+1}\rangle}{D_\tau}\right|. \quad (59)$$

For notational simplicity, let us denote

$$q_\tau = \frac{D_\tau}{\sum_{k=0}^{\tau-1} D_k}, \quad \rho = \max\left\{1, \frac{D}{\widehat{D}}\right\}, \quad b_k = \langle Z_{k+1}, h_k - x_{k+1}\rangle, \quad B_t = \sum_{i=0}^{t-1} b_i \text{ with } B_0 = 0, \quad (60)$$

where the definition of ρ captures both the cases $\widehat{D} < D$ and $\widehat{D} \geq D$, and we note that $D_0 = \widehat{D} \leq D_T$. By Lemma B.3 and Assumption 2.1, we obtain

$$q_\tau \leq \frac{\left(\frac{D_T}{D_0}\right)^{\frac{1}{T}} \log\frac{eD_T}{D_0}}{T} \leq \frac{(\rho)^{\frac{1}{T}}\log(e\rho)}{T}, \quad (61)$$

Moreover, using the notation in (60), for the last summation term on the right-hand side of relation (59) we have

$$\sum_{k=0}^{\tau-1} D_k b_k = \sum_{k=0}^{\tau-1} D_k\left(B_{k+1} - B_k\right) = D_{\tau-1}B_\tau - \sum_{k=0}^{\tau-2}\left(D_{k+1} - D_k\right)B_{k+1},$$

which holds for any $\tau \geq 2$, and $\sum_{k=0}^{\tau-1} D_k b_k = D_0 b_0$ for $\tau = 1$. Now, using the triangle inequality, and the fact that $(D_k)_{k\geq 0}$ is positive and non-decreasing, we obtain for any $\tau \geq 2$,

$$\begin{aligned}\left|\frac{\sum_{k=0}^{\tau-1} D_k b_k}{D_\tau}\right| &= \frac{D_{\tau-1}}{D_\tau}|B_\tau| + \frac{\sum_{k=0}^{\tau-2}(D_{k+1}-D_k)|B_{k+1}|}{D_\tau} \\ &\leq \frac{D_{\tau-1} + \sum_{k=0}^{\tau-2}(D_{k+1}-D_k)}{D_\tau}\max_{1\leq t\leq \tau}|B_t| \\ &= \frac{2D_{\tau-1} - D_0}{D_\tau}\max_{1\leq t\leq \tau}|B_t| \\ &\leq 2\max_{1\leq t\leq T}|B_t|. \end{aligned} \quad (62)$$

Note that the right-hand side of the preceding relation also gives an upper bound for $\tau = 1$. Hence, substituting relations (61) and (62) back to (59), we obtain

$$\epsilon_\tau \leq \frac{(\rho)^{\frac{1}{T}}\log(e\rho)}{T}\left[3\sqrt{2}LD^2 + 12D\sqrt{L\delta T} + 6D\sqrt{\sum_{i=0}^{T-1}\|Z_{i+1}-Y_i\|_*^2} + \frac{\sum_{k=0}^{T-1}\|Z_{k+1}\|_*^2}{2\eta} + \frac{\eta D^2}{2} + 2\max_{1\leq t\leq T}|B_t|\right].$$

Taking total expectation on both sides of the preceding relation, we obtain for any $T \geq 1$,

$$\mathbb{E}[\epsilon_\tau] \leq \frac{(\rho)^{\frac{1}{T}}\log(e\rho)}{T}\left[3\sqrt{2}LD^2 + 12D\sqrt{L\delta T} + 6D\mathbb{E}\left[\sqrt{\sum_{i=0}^{T-1}\|Z_{i+1}-Y_i\|_*^2}\right] + \frac{\mathbb{E}\left[\sum_{k=0}^{T-1}\|Z_{k+1}\|_*^2\right]}{2\eta} + \frac{\eta D^2}{2} + 2\mathbb{E}\left[\max_{1\leq t\leq T}|B_t|\right]\right]. \quad (63)$$

Note that the third term on the right-hand side can be upper bounded using Jensen's inequality, relation (45), and Assumption 3.1 to yield

$$6D\mathbb{E}\left[\sqrt{\sum_{i=0}^{T-1}\|Z_{i+1}-Y_i\|_*^2}\right] \leq 6D\sqrt{\mathbb{E}\left[\sum_{i=0}^{T-1}\|Z_{i+1}-Y_i\|_*^2\right]} \leq 6\sqrt{2}\sigma D\sqrt{T}. \tag{64}$$

The fourth term on the right-hand side of relation (63) can be upper bounded using Assumption 3.1 yielding

$$\frac{\mathbb{E}\left[\sum_{k=0}^{T-1}\|Z_{k+1}\|_*^2\right]}{2\eta} \leq \frac{\sigma^2 T}{2\eta}. \tag{65}$$

Next, we focus on the sixth term on the right-hand side of relation (63). We have

$$\mathbb{E}[b_k \mid \mathcal{F}_k] = \langle \mathbb{E}[Z_{k+1} \mid \mathcal{F}_k], h_k - x_{k+1}\rangle = 0.$$

Next, we show that the sequence $(B_t)_{t=1}^T$ is a martingale. Using the definition of B_{t+1} from relation (60), we have

$$\mathbb{E}[B_{t+1} \mid \mathcal{F}_t] = \mathbb{E}[B_t + b_t \mid \mathcal{F}_t] = B_t + \mathbb{E}[b_t \mid \mathcal{F}_t] = B_t.$$

Hence, the sequence $(B_t)_{t=1}^T$ is a martingale. Hence, by Lemma B.4, we obtain

$$\mathbb{E}\left[\left(\max_{1\leq t\leq T}|B_t|\right)^2\right] = \mathbb{E}\left[\max_{1\leq t\leq T}|B_t|^2\right] \leq 4\mathbb{E}\left[|B_T|^2\right].$$

Next, for $\mathbb{E}\left[|B_T|^2\right]$ we have

$$\begin{aligned}
\mathbb{E}\left[|B_T|^2\right] &= \mathbb{E}\left[\left|\sum_{k=0}^{T-1}\langle Z_{k+1}, h_k - x_{k+1}\rangle\right|^2\right] \\
&= \sum_{k=0}^{T-1}\mathbb{E}\left[(\langle Z_{k+1}, h_k - x_{k+1}\rangle)^2\right] \\
&\quad + 2\sum_{0\leq i<j\leq T-1}\mathbb{E}[[\langle Z_{i+1}, h_i - x_{i+1}\rangle\langle Z_{j+1}, h_j - x_{j+1}\rangle]] \\
&\leq \sum_{k=0}^{T-1}\mathbb{E}\left[\|Z_{k+1}\|^2\underbrace{\|h_k - x_{k+1}\|^2}_{\leq D^2}\right] \\
&\quad + 2\sum_{0\leq i<j\leq T-1}\mathbb{E}\left[\left[\langle Z_{i+1}, h_i - x_{i+1}\rangle\underbrace{\mathbb{E}[\langle Z_{j+1}, h_j - x_{j+1}\rangle \mid \mathcal{F}_j]}_{=0}\right]\right] \\
&\leq \sigma^2 D^2 T,
\end{aligned}$$

where we used Assumptions 2.1 and 3.1. Hence, using Jensen's inequality, we obtain

$$\mathbb{E}\left[\max_{1\leq t\leq T}|B_t|\right] = \mathbb{E}\left[\sqrt{\left(\max_{1\leq t\leq T}|B_t|\right)^2}\right] \leq \sqrt{\mathbb{E}\left[\left(\max_{1\leq t\leq T}|B_t|\right)^2\right]} \leq 2\sigma D\sqrt{T}. \tag{66}$$

Now substituting relations (64), (65), and (66) back in relation (63), we further obtain

$$\mathbb{E}[\epsilon_\tau] \leq \frac{(\rho)^{\frac{1}{T}}\log(e\rho)}{T}\left[3\sqrt{2}LD^2 + 12D\sqrt{L\delta T} + 6\sqrt{2}\sigma D\sqrt{T} + \frac{\sigma^2 T}{2\eta} + \frac{\eta D^2}{2} + 4\sigma D\sqrt{T}\right]. \tag{67}$$

We minimize the preceding expression with respect to η and obtain the optimal selection

$$\eta^* = \frac{\sigma\sqrt{T}}{D}.$$

Substituting $\eta = \eta^*$ in relation (67), we obtain the final relation of the theorem. **Q.E.D.**

It is evident from the results of Theorem F.3 that Algorithm 3 is universal. Moreover, compared to Theorem 4.1, we are able to remove the constant $\frac{D}{\widehat{D}}$ directly, which can get large when the estimate is poor $\widehat{D} << D$. We just have the term $(\rho)^{1/T}$ which decreases very fast to 1 as T increases, where $\rho = \max\{1, D/\widehat{D}\}$. Moreover, when $T \geq \log \rho$, then $(\rho)^{1/T} \leq e$ which is a modest constant. Next, we present a corollary to show the convergence rate for Hölder continuous mapping.

Corollary F.4. *Suppose that Assumptions 2.1 and 3.1 hold, and F is a Hölder continuous mapping for some $\nu \in [0,1]$ and $H_\nu > 0$. Then, for any $T \geq 1$, the iterates generated by Algorithm 3 satisfy*

$$\mathbb{E}\big[\epsilon_{\tau(T)}\big] \leq (\rho)^{1/T} \log(e\rho) \left[\frac{3 \times 2^{1-\nu/2} H_\nu D^{1+\nu}}{(1-\nu)^{(1-\nu)/2} T^{(1+\nu)/2}} + (6\sqrt{2}+5)\frac{\sigma D}{\sqrt{T}} \right],$$

where $\rho = \max\{1, D/\widehat{D}\}$ and we use the convention $0^0 = 1$.

Proof. The proof follows along the similar lines as Corollary 4.2. We start from the final relation of Theorem F.3 and substitute the value of $L(\delta)$ from Lemma 2.3 to the expression to obtain

$$\begin{aligned}\mathbb{E}\big[\epsilon_{\tau(T)}\big] \leq \rho^{1/T} \log(e\rho) \Bigg[&\frac{3\sqrt{2}D^2}{T} \left(\frac{1-\nu}{4}\right)^{\frac{1-\nu}{1+\nu}} \nu^{\frac{2\nu}{1+\nu}} H_\nu^{\frac{2}{1+\nu}} \delta^{-\frac{1-\nu}{1+\nu}} \\ &+ \frac{12D}{\sqrt{T}} \left(\frac{1-\nu}{4}\right)^{\frac{1-\nu}{2(1+\nu)}} \nu^{\frac{\nu}{1+\nu}} H_\nu^{\frac{1}{1+\nu}} \delta^{\frac{\nu}{1+\nu}} + (6\sqrt{2}+5)\frac{\sigma D}{\sqrt{T}} \Bigg],\end{aligned}$$

where $\rho = \max\left\{1, \frac{D}{\widehat{D}}\right\}$. For $\nu \in (0,1)$, minimizing the first two terms on the right-hand side over $\delta > 0$ yields

$$\delta^* = \frac{(1-\nu)^{(3+\nu)/2}}{2^{(5+\nu)/2}\nu} \frac{H_\nu D^{1+\nu}}{T^{(1+\nu)/2}},$$

which when substituted back yields

$$\begin{aligned}\mathbb{E}\big[\epsilon_{\tau(T)}\big] &\leq \rho^{1/T} \log(e\rho) \left[\frac{(3\sqrt{2})^\nu 6^{1-\nu}}{(1-\nu)^{(1-\nu)/2}} \frac{H_\nu D^{1+\nu}}{T^{(1+\nu)/2}} + (6\sqrt{2}+5)\frac{\sigma D}{\sqrt{T}} \right] \\ &= \rho^{1/T} \log(e\rho) \left[\frac{3 \times 2^{1-\nu/2} H_\nu D^{1+\nu}}{(1-\nu)^{(1-\nu)/2} T^{(1+\nu)/2}} + (6\sqrt{2}+5)\frac{\sigma D}{\sqrt{T}} \right].\end{aligned}$$

For the corner cases, i.e., when $\nu = 0$ and $\nu = 1$, the proof follows similarly to that of Corollary 4.2, and we obtain the desired relation for every $\nu \in [0,1]$. **Q.E.D.**

G Experimental Details and Additional Experiments

For all experiments, the diameter-related parameters used by our methods and DoG are reported in Table 1. For projected AdaGrad-Norm in the optimization experiments, we set $b_0 = 1$ and $\eta = 1$, following the notation of (Ward et al., 2020).

For the optimization problems, we report the best-iterate gap $\min_{1 \leq j \leq k}\{f(x_j) - f^*\} = f(x_k^*) - f^*$, as well as the gap at the corresponding averaged iterate, $f(\widehat{x}_k) - f^*$. We also report the step sizes generated by the different methods. For the stochastic min-max problem, we report the primal-dual gap at the averaged iterate, together with the corresponding step sizes. For both optimization and min-max experiments, we use uniform averaging, i.e., $\widehat{x}_k = \frac{1}{k}\sum_{t=1}^k x_t$, for our EGM Rules 1–3 (cf. Algorithms 1 and 2), UGM, UGM with AdaGrad, projected AdaGrad-Norm, grid-tuned EGM, and adaptive GEG. Whereas, for our DA EGM Methods 1 and 2 (cf. Algorithm 3), and DoG, we instead use D_t-weighted averaging, i.e., $\widehat{x}_k = \frac{\sum_{t=1}^k D_t x_t}{\sum_{t=1}^k D_t}$. For the stochastic min-max problem, the corresponding averaged dual iterate $\widehat{y}_k$ is computed analogously.

Next, we present additional experiments on Softmax problem and Logistic-regression problem. Towards the end, we also provide some of the additional experimental details on Hölder-smooth polyhedral feasibility problem and Min-Max Dual formulation of Constrained ℓ_2-regression problem which are also studied in Section 6.

Table 1: Distance parameters used in the numerical experiments. We set $\widehat{D}$ equal to the diameter D of the feasible set and initialize DoG and Algorithm 3 with $r_\epsilon = D_0 = 0.1D$. DoG is not included in the dual formulation of least-squares min-max experiment.

Problem	**Estimate** $\widehat{D} = D$	**DoG** $r_\epsilon = 0.1D$	**Algorithm 3** $D_0 = 0.1D$
Softmax	2	0.2	0.2
Logistic regression	2	0.2	0.2
Hölder-smooth	2×10^6	2×10^5	2×10^5
Min-max	$2\sqrt{2}$	Not used	$0.2\sqrt{2}$

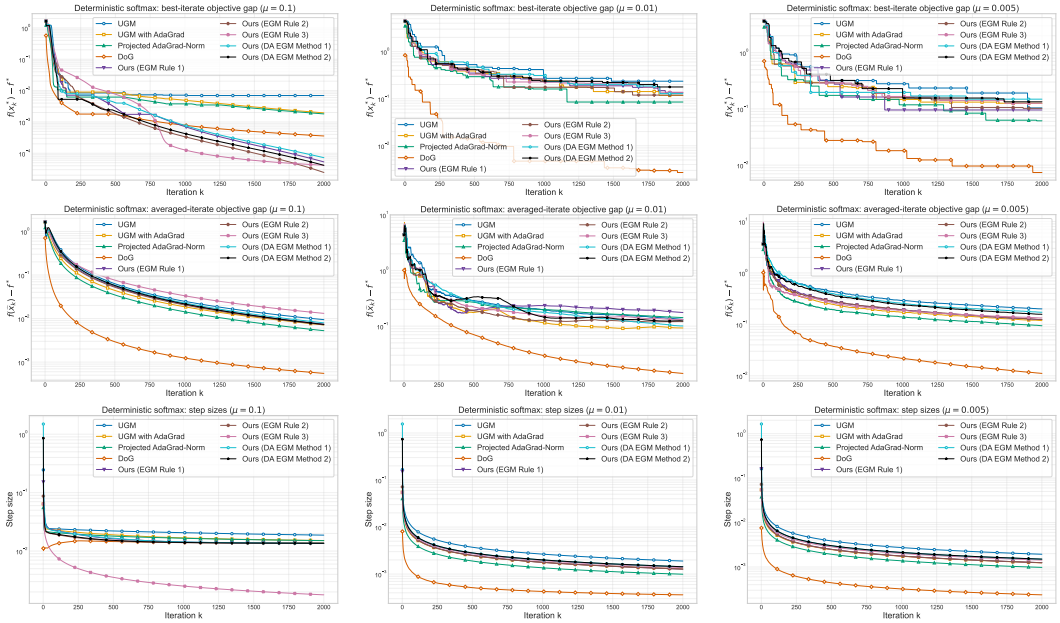


Figure 3: Results for the deterministic softmax problem. From top to bottom, the rows show the best-iterate objective function gaps, averaged-iterate objective function gaps, and step sizes.

Softmax problem. We consider the optimization problem

$$\min_{\|x\|_2 \le 1} f(x) \quad \text{where} \quad f(x) := \mu \log\left(\sum_{i=1}^{n} \exp\left(\frac{\langle a_i, x\rangle - b_i}{\mu}\right)\right),$$

where $(a_i, b_i)_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$ are the data points and $\mu > 0$ is a parameter. The problem is a smooth approximation of $\max_{1 \le i \le n}[\langle a_i, x\rangle - b_i]$ (Nesterov, 2005). The data are generated following the procedure in the paper (Moshtaghifar et al., 2026, Appendix E). The data generation technique ensures that the optimal function value $f^* = 0$. We use $n = 1000$, $d = 2000$, and $\mu \in \{0.1, 0.01, 0.005\}$. The experiment is deterministic, and the full batch is used at every iteration. As μ decreases, the softmax function becomes a sharper approximation of the non-smooth maximum function, making the optimization problem more difficult.

The convergence results in Figure 3 show that, in terms of the best-iterate objective function gap $f(x_k^*) - f^*$, our methods perform competitively on the smoother instance with $\mu = 0.1$. For the less smooth instances with $\mu = 0.01$ and $\mu = 0.005$, DoG attains smaller objective function gaps over the considered iteration horizon, while the other methods perform similarly. In terms of the averaged-iterate objective function gap $f(\widehat{x}_k) - f^*$, DoG performs better for all tested values of μ.

The purpose of these experiments is not to establish uniform empirical dominance over the existing methods. Rather, they demonstrate that our methods remain stable and consistently reduce the objective function gap across different smoothness regimes, in agreement with their theoretical

universality guarantees. Although their empirical performance may be improved by tuning $\widehat{D}$, we use the principled and uniform choice $\widehat{D} = D$ in our EGM Rules 1–3, without instance-specific tuning.

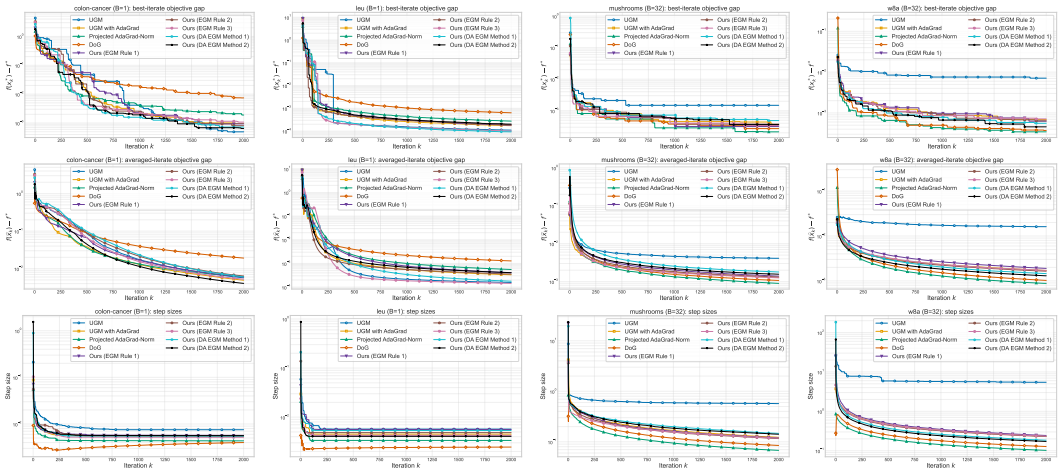


Figure 4: Results for logistic regression problem (stochastic). From top to bottom, the rows show the best-iterate objective function gaps, averaged-iterate objective function gaps, and step sizes.

Logistic-regression problem. We consider the stochastic constrained logistic-regression problem

$$\min_{\|x\|_2 \le 1} f(x) \quad \text{where} \quad f(x) := \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \exp\left(-b_i \langle a_i, x \rangle\right)\right),$$

where $a_i \in \mathbb{R}^d$ and $b_i \in \{-1, +1\}$. We use the following datasets: (i) colon-cancer, (ii) leu, (iii) mushrooms, and (iv) w8a. We select mini-batches of sizes 32 for mushrooms and w8a, and 1 for leu and colon-cancer. Since the exact optimal objective value $f^\star$ is unknown, for each dataset we use the objective value returned by L-BFGS (Liu & Nocedal, 1989) as an approximation to $f^\star$. We then use this approximation to compute the objective-function gap.

The objective function gap values evaluated at both the best-iterate and the averaged-iterate and also the step sizes are presented in Figure 4. For the averaged-iterate plot, our algorithms perform better on the colon-cancer and leukemia datasets, whereas Projected AdaGrad-Norm performs better on the mushrooms and w8a datasets.

Hölder-smooth polyhedral feasibility problem. In our experiments, we use $n = 1000$ inequality constraints, an optimization-variable dimension of $d = 1000$, and a Euclidean-ball radius of $R = 1$. We consider three values of q, namely, $q \in \{1, 1.5, 2\}$. The same realization of the generated data is used for all three values of q, so only the smoothness of the objective changes across the experiments.

Figure 5 shows the objective function gaps evaluated at the best and averaged iterates, along with the corresponding step sizes for all methods. Our algorithms perform comparably to the benchmark methods and, in most cases, achieve slightly smaller objective function gaps. The only exception is the averaged-iterate result for $q = 2$, for which UGM with AdaGrad and DoG perform slightly better.

Min-max dual formulation of a constrained ℓ_2-regression problem. In our experiments, we considered the matrix $A \in \mathbb{R}^{m \times n}$ with $m = 10000$ and $n = 8000$. The entries of the matrix A, denoted by A_{ij}, are sampled independently and identically from the uniform distribution $U[0, 1)$. We generate a target vector x_{tar} that satisfies $\|x_{\text{tar}}\|_2 = 1.5$ and we set

$$b = Ax_{\text{tar}} + \xi, \qquad \|\xi\|_2 = 0.05 \|Ax_{\text{tar}}\|_2.$$

Thus, the generating target lies outside the feasible ball, while the additive perturbation prevents a trivial exact fit.

Next, we discuss about the construction of the stochastic VI mapping. Following Nesterov & Nemirovski (2013, Section 3.6.6), we construct an unbiased stochastic estimate of the VI mapping

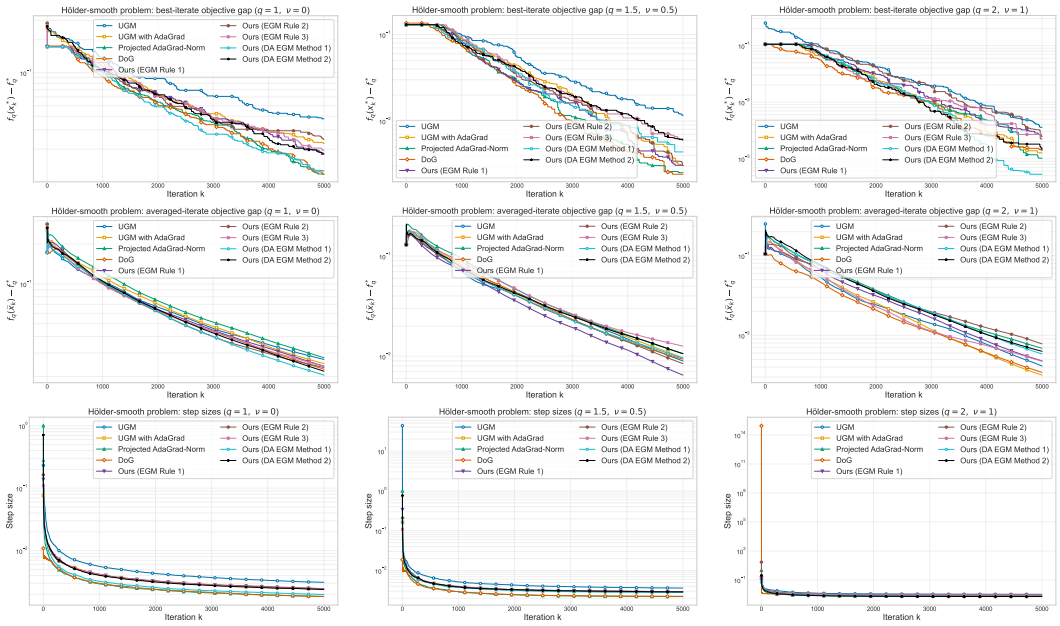


Figure 5: Results for the Hölder-smooth polyhedral feasibility problem (stochastic). From top to bottom, the rows show the best-iterate objective function gaps, averaged-iterate objective function gaps, and step sizes. The columns correspond to $q = 1$, $q = 1.5$, and $q = 2$, respectively.

using importance sampling. Let $\omega = (I, J)$ denote the sampled indices. For $x \neq 0$ and $y \neq 0$, we sample $I \in \{1, \dots, m\}$ and $J \in \{1, \dots, n\}$ independently according to

$$\mathbb{P}(I = i \mid y) = \frac{|y_i|}{\|y\|_1}, \qquad i \in \{1, \dots, m\},$$
$$\mathbb{P}(J = j \mid x) = \frac{|x_j|}{\|x\|_1}, \qquad j \in \{1, \dots, n\}.$$

We then define the stochastic matrix-vector product estimators

$$\widehat{A^\top y} = \|y\|_1 \operatorname{sign}(y_I) A^\top_{I,:},$$
$$\widehat{Ax} = \|x\|_1 \operatorname{sign}(x_J) A_{:,J}.$$

When $x = 0$ or $y = 0$, the corresponding estimator is defined to be zero. These estimators are conditionally unbiased because

$$\mathbb{E}\left[\widehat{A^\top y} \mid y\right] = \sum_{i=1}^{m} \frac{|y_i|}{\|y\|_1} \|y\|_1 \operatorname{sign}(y_i) A^\top_{i,:} = \sum_{i=1}^{m} y_i A^\top_{i,:} = A^\top y,$$
$$\mathbb{E}\left[\widehat{Ax} \mid x\right] = \sum_{j=1}^{n} \frac{|x_j|}{\|x\|_1} \|x\|_1 \operatorname{sign}(x_j) A_{:,j} = \sum_{j=1}^{n} x_j A_{:,j} = Ax.$$

Consequently, we define the stochastic VI mapping by

$$\widehat{F}([x, y], \omega) := \begin{bmatrix} \widehat{A^\top y} \\ b - \widehat{Ax} \end{bmatrix}.$$

It follows that

$$\mathbb{E}\left[\widehat{F}([x, y], \omega) \mid x, y\right] = \begin{bmatrix} A^\top y \\ b - Ax \end{bmatrix} = F([x, y]).$$

Hence, the stochastic mapping oracle is unbiased. Each oracle call accesses only one row and one column of A and has an arithmetic cost of $\mathcal{O}(m + n)$. By contrast, evaluating the exact mapping using the full matrix-vector products has a cost of $\mathcal{O}(mn)$ for a dense matrix.

For a feasible averaged point $(\widehat{x}_t, \widehat{y}_t)$ at iteration k for any $k \geq 0$, we evaluate the primal-dual gap function as

$$\mathrm{Gap}(\widehat{x}_t, \widehat{y}_t) := \max_{\|y\|_2 \leq 1} \langle A\widehat{x}_t - b, y \rangle - \min_{\|x\|_2 \leq 1} \langle Ax - b, \widehat{y}_t \rangle .$$

Both optimization problems admit closed-form solutions. In particular, we obtain

$$\max_{\|y\|_2 \leq 1} \langle A\widehat{x}_t - b, y \rangle = \|A\widehat{x}_t - b\|_2 ,$$
$$\min_{\|x\|_2 \leq 1} \langle Ax - b, \widehat{y}_t \rangle = - \left\|A^\top \widehat{y}_t\right\|_2 - \langle b, \widehat{y}_t \rangle .$$

Therefore, the primal-dual gap is evaluated exactly as

$$\mathrm{Gap}(\widehat{x}_t, \widehat{y}_t) = \|A\widehat{x}_t - b\|_2 + \left\|A^\top \widehat{y}_t\right\|_2 + \langle b, \widehat{y}_t \rangle .$$

The exact full matrix-vector products in this expression are used only to evaluate the reported primal-dual gap and are not used in the stochastic algorithmic updates.