When Extragradient Meets PAGE: Bridging Two Giants to Boost Variational Inequalities

Gleb Molodtsov[1,2] **Valery Parfenov**[1] **Egor Petrov**[1] **Evseev Grigoriy**[1] **Daniil Medyakov**[1,2] **Aleksandr Beznosikov**[2,1,3]

[1]Moscow Institute of Physics and Technology
[2]Ivannikov Institute for System Programming of the RAS
[3]Innopolis University

Abstract

Variational inequalities (VIs) have emerged as a universal framework for solving a wide range of problems. A broad spectrum of applications includes optimization, equilibrium analysis, reinforcement learning, and the rapidly evolving field of generative adversarial networks (GANs). Stochastic methods have proven to be powerful tools for addressing such problems, but they often suffer from irreducible variance, necessitating the development of variance reduction techniques. Among these, SARAH-based algorithms have demonstrated remarkable practical effectiveness. In this work, we propose a new stochastic variance reduced algorithm for solving stochastic variational inequalities. We push the boundaries of existing methodologies by leveraging PAGE method to solve VIs. Unlike prior studies that lacked theoretical guarantees under general assumptions, we establish rigorous convergence rates, thus closing a crucial gap in the literature. Our contributions extend both theoretical understanding and practical advancements in solving variational inequalities. To substantiate our claims, we conduct extensive experiments across diverse benchmarks, including a widely studied denoising task. The results consistently showcase the superior efficiency of our approach, underscoring its potential for real-world applications.

1 INTRODUCTION

Variational inequalities (VIs) have been a cornerstone of mathematical research for a long time, offering an approach to solving a wide range of problems. With the pioneering work [Browder, 1965], and since then, they have become an indispensable tool. We consider the VI problem by seeking a solution $z^* \in \mathcal{Z}$ that satisfies the following condition:

$$\forall z \in \mathcal{Z} \hookrightarrow \langle F(z^*), z - z^* \rangle \geqslant 0, \tag{1}$$

where F is a monotone operator. Variational inequalities offer a versatile framework for tackling various mathematical challenges, including minimization problems, saddle and fixed point problems [Stampacchia, 1964, Facchinei and Pang, 2003, Kinderlehrer and Stampacchia, 2000]. To build intuition, we present several illustrative examples.

Example 1 (Convex optimization)**.** *Consider the optimization problem:*

$$\min_{z \in \mathbb{R}^d} [f(z)]. \tag{2}$$

Here, f represents a smooth data fitting term. In this scenario, let $F(z) = \nabla f(z)$. Thus, if f is convex, the optimization problem (2) *can be reformulated within the variational inequality framework.*

One of the main reasons for the widespread use of VIs is that many non-smooth optimization problems can be reformulated as saddle point problems, significantly improving solution efficiency [Nesterov, 2005, Nemirovski, 2004, Chambolle and Pock, 2011, Esser et al., 2010].

Example 2 (Convex-concave saddle points)**.** *Now, consider the convex-concave saddle point problem:*

$$\min_{x \in \mathbb{R}^{d_x}} \max_{y \in \mathbb{R}^{d_y}} [f(x, y)]. \tag{3}$$

In this setting, f serves the same role as in Example 1. Define $F(z) = F(x, y) = [\nabla_x f(x, y), -\nabla_y f(x, y)]$. Thus, if f is smooth and convex-concave, this establishes the connection between the saddle point formulation (3) *and variational inequalities.*

The investigation of minimization problems is frequently conducted independently of VIs. However, the study of saddle point problems is closely intertwined with VIs, as these two areas share a strong theoretical and practical connection.

Example 3. *(Fixed points). Consider the fixed point problem:*

$$\text{Find } z^* \in \mathbb{R}^d \text{ such that } T(z^*) = z^*, \tag{4}$$

where $T : \mathbb{R}^d \to \mathbb{R}^d$ is an operator. With $F(z) = z - T(z)$, it can be proved that $z^* \in \mathbb{R}^d$ is a solution for (1) *if and only if $F(z^*) = 0$, i.e. $z^* \in \mathbb{R}^d$ is a solution for* (4)*.*

Additionally, recent research has established important connections between VIs and fields such as reinforcement learning [Omidshafiei et al., 2017, Jin and Sidford, 2020], adversarial training [Madry et al., 2017], and generative adversarial networks (GANs) [Goodfellow et al., 2014]. In particular, in-depth analysis of variational inequalities provides theoretical insights and practical guidance for improving GAN training methods [Daskalakis et al., 2017, Gidel et al., 2018, Mertikopoulos et al., 2018, Chavdarova et al., 2019, Liang and Stokes, 2019, Peng et al., 2020].

Beyond these modern applications, VIs also play a crucial role in classical problems such as clustering [Xu et al., 2004], matrix factorization [Bach et al., 2008], image denoising [Esser et al., 2010, Chambolle and Pock, 2011], robust optimization [Ben-Tal et al., 2009], economic modeling, game theory [Von Neumann and Morgenstern, 1953], and optimal control [Facchinei and Pang, 2003].

Despite their broad applicability, solving variational inequalities presents significant challenges. Traditional optimization techniques, such as the gradient method, often fail in this context, both in terms of efficiency and theoretical convergence guarantees [Harker and Pang, 1990, Beznosikov et al., 2023]. Among the many approaches developed for solving VIs, the EXTRAGRADIENT method [Korpelevich, 1976, Mokhtari et al., 2020] has proven to be one of the most effective.

Recent advances in machine learning and data science introduce additional complexities. The growing size of datasets and increasing model complexity demand computationally efficient algorithms [Bottou, 2010, Dean et al., 2012]. A fundamental optimization problem underlying many machine learning tasks is Empirical Risk Minimization (ERM). In the context of distributed systems, where data is spread across multiple devices, the ERM problem is commonly formulated as:

$$\min_{z \in \mathbb{R}^d} \left[f(z) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{\xi_i \sim \mathcal{D}_i} \left[f_{\xi_i}(z) \right] \right], \tag{5}$$

where $\mathcal{D}_i$ is an unknown distribution of the training sample data on the i-th device. A particularly important and widely studied case arises when all nodes share the same underlying data distribution, i.e., $\mathcal{D}_i = \mathcal{D}$ for all i. In this setting, the objective (5) reduces to a finite-sum optimization problem:

$$f(z) = \frac{1}{n} \sum_{i=1}^{n} f_i(z), \tag{6}$$

where each $f_i(z) := \mathbb{E}_{\xi \sim \mathcal{D}}[f_\xi(z)]$ corresponds to the expected loss over a local mini-batch or a single sample, assuming access to identical distributions across nodes.

Moreover, this finite-sum function can be reformulated as an adversarial training problem:

$$\min_{w \in \mathbb{R}^d} \max_{\|r_i\| \leqslant D} \left[\frac{1}{N} \sum\nolimits_{i=1}^{N} \left(w^T (x_i + r_i) - y_i \right)^2 + \frac{\lambda}{2} \|w\|^2 - \frac{\beta}{2} \|r\|^2 \right], \tag{7}$$

where the samples correspond to features x_i and targets y_i. This reformulation enables efficient large-scale problem solving.

Stochastic algorithms are particularly well-suited for handling such problems as (6). Instead of computing the full operator value at each iteration, stochastic methods randomly sample F_i. The stochastic EXTRAGRADIENT method [Juditsky et al., 2011] follows this principle by selecting independent random indices i_t, j_t at iteration t and performing the following updates:

$$\begin{aligned} z^{t+\frac{1}{2}} &= z^t - \gamma F_{i_t}(z^t), \\ z^{t+1} &= z^t - \gamma F_{j_t}(z^{t+\frac{1}{2}}). \end{aligned} \tag{8}$$

This method extends the classical SGD approach [Robbins and Monro, 1951] by incorporating an additional step to improve stability. However, it suffers from high variance in stochastic operator estimates, limiting its convergence to a neighborhood of the optimal solution rather than the exact solution itself [Juditsky et al., 2011, Mishchenko et al., 2020]. This issue also affects classical SGD [Bottou, 2009, Moulines and Bach, 2011, Gower et al., 2020]. The intuition behind this problem can be easily extracted from the example with a setup involving heterogeneous data, where near the optimal point $\nabla f(z^*) = \frac{1}{n} \sum_{i=1}^{n} \nabla f_i(z^*) \to 0$, while some $\nabla f_i(z^*)$ can retain finite values. As a result, in the optimum region, SGD will take large steps, preventing it from reaching the optimum.

A major breakthrough in addressing this limitation was the introduction of variance reduction (VR) techniques, originally developed for finite-sum minimization [Johnson and Zhang, 2013]. At each iteration, an index i_t is selected along with a reference point ω^t, which is periodically updated or chosen probabilistically [Kovalev et al., 2020]. In the context of convex optimization, the variance-reduced gradient at $z^{t+\frac{1}{2}}$ is given by:

$$g(z^t) = \nabla f_{i_t}(z^{t+\frac{1}{2}}) - \nabla f_{i_t}(\omega^t) + \nabla f(\omega^t). \tag{9}$$

Variance reduction techniques construct more accurate gradient estimators over time, enabling the use of larger step sizes and accelerating convergence.

In addition to the widely used SVRG [Johnson and Zhang, 2013], related methods include SAG [Roux et al., 2012, Schmidt et al., 2017], SAGA [Defazio et al., 2014a, Qian et al., 2019], and FINITO [Defazio et al., 2014b]. However, for both convex and non-convex smooth minimization problems, the best guarantees of convergence are given by other variance-reduced technique SARAH [Nguyen et al., 2017, Hu et al., 2019] (and its modifications: SPIDER [Fang et al., 2018], STORM [Cutkosky and Orabona, 2019]). Notably, loopless version PAGE [Li et al., 2021] has garnered significant interest due to its ability to provide improved convergence guarantees through probabilistic reference point updates.

SARAH technique rejects memorizing all components of the reference gradient and deals with the biased gradient estimator in the inner loop:

$$g^t = \nabla f_{i_t}(z^{t+\frac{1}{2}}) - \nabla f_{i_t}(z^{t-\frac{1}{2}}) + g^{t-1}. \qquad (10)$$

Biasedness complicates the theoretical analysis. At the same time, such an update rule leads to smoother changes in the gradient estimator g from iteration to iteration, lower memory costs, and demonstrates better practical performance. Returning to the example with heterogeneous data, this time the difference $\nabla f_{i_t}(z^t) - \nabla f_{i_t}(z^{t-1})$ is going to be small for small steps. This fact allows remain the scale of g^t after its initialization with original gradient in the outer loop. Therefore, the issue with large gradient estimators near the optimum is resolved. Additionally, provided demonstration outlines the practical difference of the (10) and (9) update rules. Indeed, (10) utilizes the gradient difference in consecutive points, while (9) considers the difference between the current and reference points. This provides an additional boost to (10).

The probabilistic approach simplifies the theoretical analysis achieving the best convergence guarantees. Particularly, we draw attention to the iteration of PAGE, which provides the intuition behind our algorithm:

$$g^t = \begin{cases} \nabla f(z^t) & , \textit{with prob. } p \\ \nabla f_{i_t}(z^t) - \nabla f_{i_t}(z^{t-1}) + g^{t-1} & , \textit{with prob. } 1-p. \end{cases}$$

Meanwhile, current research continues to explore the application of variance reduction techniques for solving variational inequalities. Although most methods in this area are based on SVRG [Alacaoglu and Malitsky, 2021, Medyakov et al., 2024], the more practically beneficial SARAH method has received limited attention, with only a few studies examining its application [Beznosikov and Gasnikov, 2023]. Our work bridges this gap by proposing the use of such variance reduction technique in the loopless version for variational inequalities under broader assumptions of Lipschitz continuity and monotonicity.

BRIEF LITERATURE REVIEW

• **Deterministic approaches for solving VIs.** As previously noted, the EXTRAGRADIENT method [Korpelevich, 1976] is a classical deterministic approach for solving the problem (1) in the Euclidean setting. Building on this, the MIRROR-PROX method [Nemirovski, 2004] was introduced, incorporating Bregman divergence to extend the framework to non-Euclidean geometries. In addition to these, several other deterministic methods have been proposed for solving VIs, including FORWARD-BACKWARD-FORWARD (FBF) [Tseng, 2000], DUAL EXTRAPOLATION [Nesterov, 2007], REFLECTED GRADIENT [Malitsky, 2015], and FORWARD-REFLECTED-BACKWARD (FoRB) [Malitsky and Tam, 2020].

• **Stochastic methods for VIs.** The application of various stochastic methods for solving variational inequalities and saddle point problems has been the subject of extensive research. The first stochastic versions of algorithms for solving variational inequalities were proposed by [Juditsky et al., 2011]. The idea was further developed in [Gidel et al., 2018, Hsieh et al., 2019, Mishchenko et al., 2020, Hsieh et al., 2020, Gorbunov et al., 2022, Beznosikov et al., 2023, 2024, Solodkin et al., 2024]. Subsequently, researchers employed variance reduction techniques to mitigate the inherent variance in these stochastic methods. Specifically, [Palaniappan and Bach, 2016] explored a stochastic GRADIENT METHOD with VR, combining SVRG with Catalyst acceleration.

The combination of these techniques with methods traditionally used for variational inequalities appeared in [Chavdarova et al., 2019] who integrated EXTRAGRADIENT with SVRG, leveraging variance reduction to achieve improved convergence rates. The aforementioned variance reduction technique has also been explored later [Alacaoglu et al., 2021, Alacaoglu and Malitsky, 2021, Kovalev et al., 2022, Beznosikov et al., 2022].

Although most of the methods were based on the SVRG approach, some studies focused on analyzing the SARAH method, which is more appealing from a practical standpoint for minimization problems. Thus, [Chen et al., 2022] proposed SPIDER-GDA, achieving a stochastic first-order oracle complexity of $\mathcal{O}\left(\left(n + \sqrt{n}\kappa_x\kappa_y^2\right)\log(1/\epsilon)\right)$ under two-sided conditions ($\kappa_x = L/\mu_x$, $\kappa_y = L/\mu_y$). The given estimate has a significant drawback: it depends cubically on L/μ. In reality, while batch size parameters n can be dynamically adjusted to influence convergence speed, the problem parameters remain fixed. Consequently, despite potential gains from adjusting n, the overall estimate typically presents a much worse scenario on average. Later, [Beznosikov and Gasnikov, 2023] presented results for the SARAH method with objective functions under a cocoercivity condition on the operator. However, the given assumption is a more stringent analogue of the Lipschitz continuity condition and does not hold even for training a neural net-

work with two convolutional layers [Cybenko, 1989]. A comparison of these assumptions for variational inequalities is provided in [Loizou et al., 2021]. In contrast, our study offers an analysis based on more general assumptions.

CONTRIBUTIONS

Our main contributions are highlighted here.

- *Adaptation of* PAGE *for Variational Inequalities.* We present an application of the PAGE method, leveraging its practically beneficial variance reduction technique for solving variational inequalities.
- *Convergence Estimates under General Assumptions.* We provide theoretical convergence estimates for our method under more general assumptions on the operator and problem conditions (Lipschitz constant), surpassing previous studies in this area.
- *Comprehensive Experimental Validation.* Extensive experiments demonstrate the superiority of applying PAGE to EXTRAGRADIENT methods over their vanilla versions or its previous combinations with variance reduction technique. To validate our approach, we conducted the following experiments:

1. Training ResNet-18 on CIFAR-10 for a multi-class classification task.
2. Image denoising as a practical application of saddle-point methods.
3. Solving toy bilinear tasks to analyze performance in controlled settings.
4. Adversarial training to highlight robustness and efficiency.

SETUP

Notation. In this paper, we use $\langle x, y\rangle := \sum_{i=1}^n x_i y_i$ to denote the standard inner product of $x, y \in \mathbb{R}^d$, where x_i corresponds to the i-th component of x in the standard basis of $\mathbb{R}^d$. It induces the ℓ_2-norm in $\mathbb{R}^d$ in the following way: $\|x\| := \|x\|_2 = \sqrt{\langle x, x\rangle}$.

Recall that we consider the problem (1), where the operator F has the form (6). Additionally, we present a list of assumptions within which we obtain the main statements.

Assumption 1. ***(Lipschitzness.)*** *The operator F has a stochastic oracle F_i that is unbiased $F(z_1) = \mathbb{E}[F_i(z_1)]$ and is L-Lipschitz in mean:*

$$\mathbb{E}_i\left[\|F_i(z_1) - F_i(z_2)\|^2\right] \leqslant L^2\|z_1 - z_2\|^2$$

for any $z_1, z_2 \in \mathcal{Z}$.

Note that F can be expressed as a finite sum, $F = \frac{1}{n}\sum_{i=1}^n F_i$, where each component F_i is L_i-Lipschitz continuous, and the full operator F is L_F-Lipschitz. By applying the triangle inequality, it naturally follows that $L_F \leqslant \frac{1}{n}\sum_{i=1}^n L_i$. On the one hand, the sum $\frac{1}{n}\sum_{i=1}^n L_i$ can be significantly larger than L_F. On the other hand, while computing each individual L_i may be straightforward, determining the exact value of L_F might not be feasible. In such cases, the inequality provides a practical upper bound for L_F.

Even in the general form, the problem demonstrates potential issues caused by suboptimal stochastic oracles. If the Lipschitz constant L of our stochastic oracle is significantly worse (i.e., larger) than L_F, it can negate the benefits of using inexpensive stochastic oracles. In what follows, for finite-sum problems, we assume that Lipschitz constants are similar for two arbitrary oracles from this sum.

Assumption 2. ***(Monotonicity conditions.)*** *We need two cases of monotonicity:*

(a) ***Strong monotonicity:*** *Operator F is μ-strongly monotone, i.e.,*

$$\langle F(z_1) - F(z_2), z_1 - z_2\rangle \geqslant \mu\|z_1 - z_2\|^2$$

for any $z_1, z_2 \in \mathcal{Z}$.

(b) ***Monotonicity:*** *Operator F is monotone, i.e.,*

$$\langle F(z_1) - F(z_2), z_1 - z_2\rangle \geqslant 0$$

for any $z_1, z_2 \in \mathcal{Z}$.

For minimization problems, Assumption 2(a) means strong convexity, and for saddle point problems, strong convexity–strong concavity. At the same time, variance reduction methods are usually considered under Assumption (1) or its analogues, such as L-smoothness in the worst-case scenario. In light of these facts, our assumptions are classic for such problems.

2 ALGORITHMS AND CONVERGENCE ANALYSIS

Having established the necessary background, we can now proceed to the main theoretical contribution of our paper. Let us start with our Algorithm 1 (EXTRAPAGE).

Line 5 demonstrates that EXTRAPAGE encapsulates PAGE update rule principle. In particular, the oracle uses information about the operator from previous iterations in order to reach a variance reduction effect. At the same time, it does not apply reference point concept, instead of this we use probabilistic approach, defining

$$G^t = F(z^{t+1/2})$$

Algorithm 1 EXTRAPAGE

1: **Input:** Initial points $z^{-1/2} = z^0 \in \mathbb{R}^d$; Initial gradient $G^{-1} = F(z^{-1/2})$
2: **Parameter:** Stepsize $\gamma > 0$, probability $p \in (0, 1]$
3: **for** $t = 0, 1, 2, \ldots, T-1$ **do**
4: $\quad z^{t+\frac{1}{2}} = z^t - \gamma G^{t-1}$
5: $\quad G^t = \begin{cases} F(z^{t+1/2}), & p \\ G^{t-1} + F_{i^t}(z^{t+1/2}) - F_{i^t}(z^{t-1/2}), & 1-p \end{cases}$
6: $\quad z^{t+1} = z^t - \gamma G^t$
7: **end for**
8: **Output:** z^T

with probability p and

$$G^t = G^{t-1} + F_{i^t}(z^{t+1/2}) - F_{i^t}(z^{t-1/2})$$

with probability $1-p$. It means, we compute full gradient every $\frac{1}{p}$ iterations in average. Therefore, using $p \sim \frac{1}{n}$ significantly reduces oracle complexity compared to classical GD, loosing to SGD in computational iteration cost only by a constant factor.

Let us now recall the classical EXTRAGRADIENT step (8) and pay special attention on how we adapt it to our case. As in the vanilla EXTRAGRADIENT method, we use $z^{t+1/2}$, computed in Line 4. One can note that we use a new computed reduced gradient to perform the main step of the method, and the previous one to find an extrapolation point. In this way, we accurately adapt the variance reduction idea to the EXTRAGRADIENT technique. However, in contrast to vanilla Stochastic EXTRAGRADIENT, our algorithm updates the gradient estimator only once per iteration. The theoretical analysis has revealed that the second gradient estimator update makes the recursion formulation more complex without yielding better convergence guarantees.

It is worth noting that SARAH lacks an essential feature of unbiasedness in stochastic operators compared to the SVRG algorithm:

$$\mathbb{E}_{i^t}\left[G_{i^t}(z^t)\right] \neq \frac{1}{n}\sum_{i=1}^{n} F_i(z^t) = F(z^t).$$

This limitation results in a more complex analysis and requires non-standard techniques to establish convergence. More particularly, the only remaining tool for theoretical analysis is evaluation of terms under the expectation, with explicit use of algorithm iterations.

Now we turn to the formal analysis. First, we would like to provide a brief discussion. In stochastic optimization, computational cheap gradient estimators are used instead of gradients. This way, stochastic algorithms reach acceleration in terms of iteration cost, but to remain its iterations effective, we strive to minimize the difference between the estimator and original gradient. As mentioned earlier, the concept of variance reduction methods is to collect information from previous iterations and use it to improve the quality of the gradient estimation at the current point. Subsequently, controlling the difference between the original gradient and its estimator is a key consideration in the development of an effective stochastic method. Our algorithm enables a recursive analysis of the squared norm of this difference, which is formalized in Lemma 1. This lemma is pivotal not only for deriving convergence guarantees but also for gaining fundamental insights into variance reduction techniques.

Lemma 1. *For iterations of Algorithm 1 the following inequality holds:*

$$\mathbb{E}_{i^t}\mathbb{E}_{G^{t-1}}\left\|F(z^{t+1/2}) - G^t\right\|^2 =$$
$$(1-p)\Bigg[\mathbb{E}_{i^t}\left\|F_{i^t}\left(z^{t+1/2}\right) - F_{i^t}\left(z^{t-1/2}\right)\right\|^2$$
$$+ \left\|F(z^{t-1/2}) - G^{t-1}\right\|^2 - \left\|F\left(z^{t+1/2}\right) - F\left(z^{t-1/2}\right)\right\|^2\Bigg].$$

If $F = \frac{1}{n}\sum\limits_{i=1}^{n} F_i$ represents the loss of the model on homogeneous data, the first term in brackets tends to be small, which means remaining the estimation quality during the executing of the algorithm. In our setting, Assumption 1 provides a further estimate:

$$\mathbb{E}\left\|F(z^{t+1/2}) - G^t\right\|^2 \leqslant (1-p)\mathbb{E}\left\|F(z^{t-1/2}) - G^{t-1}\right\|^2 + (1-p)L^2\mathbb{E}\left\|z^{t+1/2} - z^{t-1/2}\right\|^2.$$

At this point, we introduce our main theoretical result. Biasedness not only complicates the analysis but also affects the convergence criteria. Theorem 1 provides convergence guarantees for the EXTRAPAGE algorithm based on a specifically constructed function that ensures the stability of the method:

$$V^t = \mathbb{E}\Big[\left\|z^t - z^*\right\|^2 + \gamma^2 H\left\|F(z^{t-1/2}) - G^{t-1}\right\|^2 + 2\gamma M\left\langle F(z^{t-1/2}) - G^{t-1}, z^{t-1/2} - z^*\right\rangle + \gamma^2\left\|G^t - G^{t-1}\right\|^2\Big],$$

where $M = \frac{1-p}{p-\gamma\mu}$ and $H = 70n^3$. Taking into account the choice of the stepsize γ and probability p in Theorem 1, M can be estimated as $M \sim n$. We outline the third term in the Lyapunov function V^t, which represents a scalar product. Such a term can be negative and is unusual for Lyapunov functions in variational inequality problems.

Theorem 1. *Under Assumptions 1, 2(a), after T iterations of Algorithm 1 with $\gamma \leqslant \frac{1}{30Ln^{3/2}}, p = \frac{1}{n}$, the following holds:*

$$V^T \leqslant (1-\gamma\mu)^T\left\|z^0 - z^*\right\|^2.$$

Thus, we established the linear convergence of Algorithm 1 with respect to the function V. It is important to outline that V includes terms containing the difference $F(z^{t+1/2}) - G^t$, which imposes tight restrictions on $\left\|F(z^{t+1/2}) - G^t\right\|$.This suggests that G^t provides a sufficiently accurate approximation of $F(z^{t+1/2})$, thereby validating the effectiveness of our choice for the update rule of G^t.

Corollary 1 reflects the superiority of the obtained guarantees based on the function V^t over the usual criterion $\|z^t - z^*\|^2$, which is not obvious due to the possible negativity of the scalar product $\left\langle F(z^{t-1/2}) - G^{t-1}, z^{t-1/2} - z^* \right\rangle$.

Corollary 1. *In settings of Theorem 1, after T iterations of Algorithm 1 with $\gamma \leqslant \frac{1}{30Ln^{3/2}}$ and $p = \frac{1}{n}$, the following holds:*

$$\mathbb{E}\left[\tfrac{1}{2}\|z^T - z^*\|^2 + \tfrac{\gamma^2 H}{2}\left\|F(z^{T-1/2}) - G^{T-1}\right\|^2\right] \leqslant (1-\gamma\mu)^T \left\|z^0 - z^*\right\|^2.$$

As a final point of our theoretical analysis, we introduce Corollary 2.

Corollary 2. *Suppose Assumptions 1, 2(a) hold. Then Algorithm 1 with $\gamma = \frac{1}{30Ln^{3/2}}$ and $p = \frac{1}{n}$, to reach ε-accuracy, where $\varepsilon \sim V^T$, needs*

$$\mathcal{O}\left(\frac{Ln^{3/2}}{\mu}\log\frac{1}{\varepsilon}\right) \quad \textit{iterations and oracle calls.}$$

Let us briefly discuss the result. Comparing our result with other estimates in this class, under Assumptions 1 and 2(a), our algorithm has a worse dependence on n. Nevertheless, this phenomenon is explainable. The convergence analysis of variance reduction methods that are not unbiased, particularly in the extragradient setting, inherently introduces an inner product term of the form $\left\langle F(z^{t-1/2}) - G^{t-1}, z^{t-1/2} - z^* \right\rangle$ in the function governing the convergence rate. This inner product significantly complicates the analysis. Placing this term in the recursion leads to the emergence of an additional degree of $n^{\frac{1}{2}}$. However, this can be seen as a trade-off – a chance to obtain superior practical convergence at the cost of gradient bias.

Remark 1. *We can transform the obtained estimation for the case of monotone stochastic operators 2(b) acting on a bounded domain of diameter D. To do this, we use a regularization trick with $\mu \sim \frac{\varepsilon}{D^2}$. Thus, solving the problem with the operator $\hat{F}(z) = F(z) + \mu(z - z^0)$ with the accuracy $\frac{\varepsilon}{2}$, we solve the problem* (1) *with the accuracy ε and obtain $\widetilde{\mathcal{O}}\left(\frac{Ln^{3/2}}{\varepsilon}\right)$ iteration and oracle complexity. This is convergence in argument, it differs from the classical form.*

3 EXPERIMENTS

Our experimental evaluation spans a diverse set of tasks, illustrating the effectiveness of EXTRAPAGE in various practical settings. The structure of this section is as follows:

- **Analysis on Toy Bilinear Problems (Section 3.1):** We begin with an evaluation of EXTRAPAGE's performance on synthetic bilinear problems. These controlled experiments serve as a baseline for comparison with existing approaches.

- **Deep Learning Scalability (Section 3.2):** We assess the scalability and adaptability of EXTRAPAGE by training a ResNet-18 model on the CIFAR-10 image classification.

- **Practical Utility in Denoising (Section 3.3):** We then apply our method to image denoising — a canonical application of saddle-point optimization.

- **Performance on GAN Training (Section 3.4):** To further validate the robustness and convergence properties of EXTRAPAGE, we compare it against established baselines in the challenging setting of GAN training.

We compare Algorithm 1 EXTRAPAGE to those in the literature. Therefore, we take EXTRAGRADIENT [Juditsky et al., 2011], EXTRAGRADIENT WITH VARIANCE REDUCTION (EGVR) [Alacaoglu and Malitsky, 2022], Stochastic Gradient Descent Ascent (SGDA) [Nemirovski et al., 2009] with and without clipping and SPIDER-GDA [Chen et al., 2022] algorithms as a reference. Additional experiments, including adversarial training and extended formulations discussed above, are provided in Appendix A.

3.1 BILINEAR SADDLE POINT PROBLEM

We start our experiments with a distributed bilinear problem

$$\min_{x\in\mathbb{R}^{d_x}}\max_{y\in\mathbb{R}^{d_y}} x^T A y + a^T x + b^T y + \frac{\lambda^2}{2}\|x\|^2 - \frac{\lambda^2}{2}\|y\|^2, \quad (11)$$

where $A \in \mathbb{R}^{d\times d}, a, b \in \mathbb{R}^d$. This problem is λ-strongly convex–strongly-concave and, moreover, it is $\|A\|_2$-smooth. Therefore, this distributed problem is well suited for the primary comparison of our methods. We take $d = 100$, and in order to apply stochastic methods, we generate a set of positive definite matrices A_i and vectors a_i, b_i randomly. We represent matrix A as the sum of matrices A_i, that is, $A = \sum_{i=1}^n A_i$, where $n = 100$, the same operation is performed for vectors a and b.

The experiments are carried out for matrices with the ratio of eigenvalues $\frac{L}{\mu} = 10^4$ and $\frac{L}{\mu} = 10^2$, where L, μ are the maximum and minimum eigenvalues of the matrix A, respectively. The results are presented in Figures 1, 2.

The empirical findings reveal that EXTRAPAGE showcases enhanced convergence when contrasted with the aforementioned algorithms. Of particular note is the performance of our methodology at huge L/μ ratios. Although the convergence rate of L^3/μ^3 was formerly derived by [Chen et al., 2022], this bound fails to precisely depict the real-world behavior witnessed with substantial condition number. Despite the fast convergence of the algorithm under such conditions,

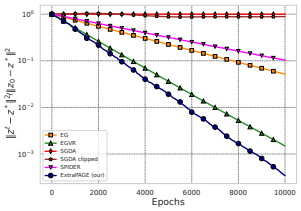


Figure 1: EXTRAPAGE compared to different baselines on the problem 11 with $\frac{L}{\mu} = 10^4$.

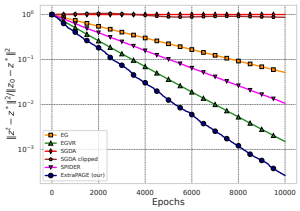


Figure 2: EXTRAPAGE compared to different baselines on the problem 11 with $\frac{L}{\mu} = 10^2$.

a considerable gap persisted between the theoretical prediction and the observed empirical outcomes. This paper bridges this shortcoming by establishing a refined, more accurate convergence rate. This rate faithfully reflects the actual performance, specifically with such large condition number, which are common in real-world scenarios.

3.2 IMAGE CLASSIFICATION

We investigate the performance of our method compared to the baselines on an image classification problem. We consider the ResNet-18 model [He et al., 2016] with the provided in this paper weight optimizers and the public CIFAR-10 [Krizhevsky et al., 2009] image dataset. To explore the robustness of the optimizers, we reformulate the standard minimization problem into the min-max optimization framework. Specifically, let $f(w, x, y)$ denote the loss function, where $w \in \mathbb{R}^{d_w}$ represents the model parameters, $x \in \mathbb{R}^{d_x}$ is the input, and $y \in \mathbb{R}$ is the corresponding label. We consider the following optimization problem:

$$\min_{w \in \mathbb{R}^{d_w}} \max_{\|r_i\| \leqslant D} \frac{1}{n} \sum_{i=1}^{n} f(w, x_i + r, y_i) + \frac{\lambda}{2}\|w\|^2 - \frac{\beta}{2}\|r\|^2,$$

where f is the cross-entropy loss function, r represents adversarial noise introduced to model data perturbations, and λ, β are regularization parameters. The formulation can be expressed as a variational inequality:

$$z = \begin{pmatrix} w \\ r \end{pmatrix}, \quad F_i(z) = \begin{pmatrix} \nabla_w f(w, x_i + r, y_i) + \lambda w \\ -\nabla_r f(w, x_i + r, y_i) + \beta r \end{pmatrix}.$$

The results are presented in Figure 3.

EXTRAPAGE exhibits stronger fluctuations in both accuracy and loss, yet this dynamic behavior enables it to achieve higher peak and average accuracy compared to other baselines. While the trajectory is more volatile, the algorithm consistently outperforms the alternatives, demonstrating its effectiveness for the applied image classification task. We also compare the running times for all methods in Table 1.

Despite being slower than the fastest baselines, EXTRAPAGE demonstrates a reasonable epoch time while delivering superior performance in terms of accuracy, as observed

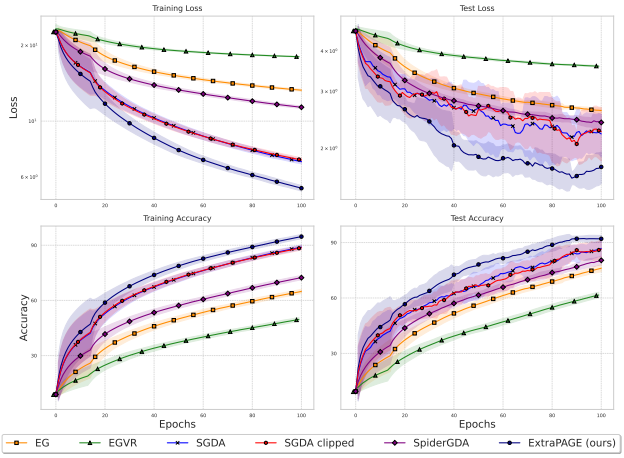


Figure 3: EXTRAPAGE compared to different baselines on CIFAR dataset. We choose $n = 100, p = \frac{1}{n}$.

Table 1: Runtime comparison of our algorithms with $n = 100, p = \frac{1}{n}$.

Algorithm	Total Time	Round Time
EG	467.447	4.674 ± 0.020
EGVR	618.560	6.186 ± 0.009
SGDA	433.672	4.337 ± 0.043
SGDA clipped	438.911	4.389 ± 0.252
SPIDER	841.750	8.417 ± 0.761
ExtraPAGE	634.692	6.347 ± 0.010

in the corresponding learning curves. This suggests that its computational overhead is justified by improved convergence behavior. In Appendix A.3 we further investigate the convergence and runtime values of ExtraPAGE and other baselines at different values of n and p.

3.3 IMAGE DENOISING

To formulate the image denoising problem [Chambolle and Pock, 2011], we consider the classic saddle point problem as we demonstrate in Example 2:

$$\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} \left[\langle Kx, y \rangle + G_1(x) - G_2(y) \right],$$

where regularizers G_1 and G_2 are proper convex lower semi-continuous functions, and K is a continuous linear operator. To proceed with image denoising, we consider g as a given noisy image and u as the solution we seek. We use the Cartesian grid with the step $h : \{(i \cdot h, j \cdot h)\}$. Thus, specifically for the image denoising, we consider:

$$\min_{u \in \mathcal{X}} \max_{p \in \mathcal{Y}} \left[\langle \nabla u, p \rangle_{\mathcal{Y}} + \lambda/2 \|u - g\|_2^2 - \delta_P(p) \right],$$

where p is a dual variable, $\delta_P(p)$ is the indicator function of the set P defined as: $P = \{p \in \mathcal{Y} : \|p(x)\| \leqslant 1\}$.

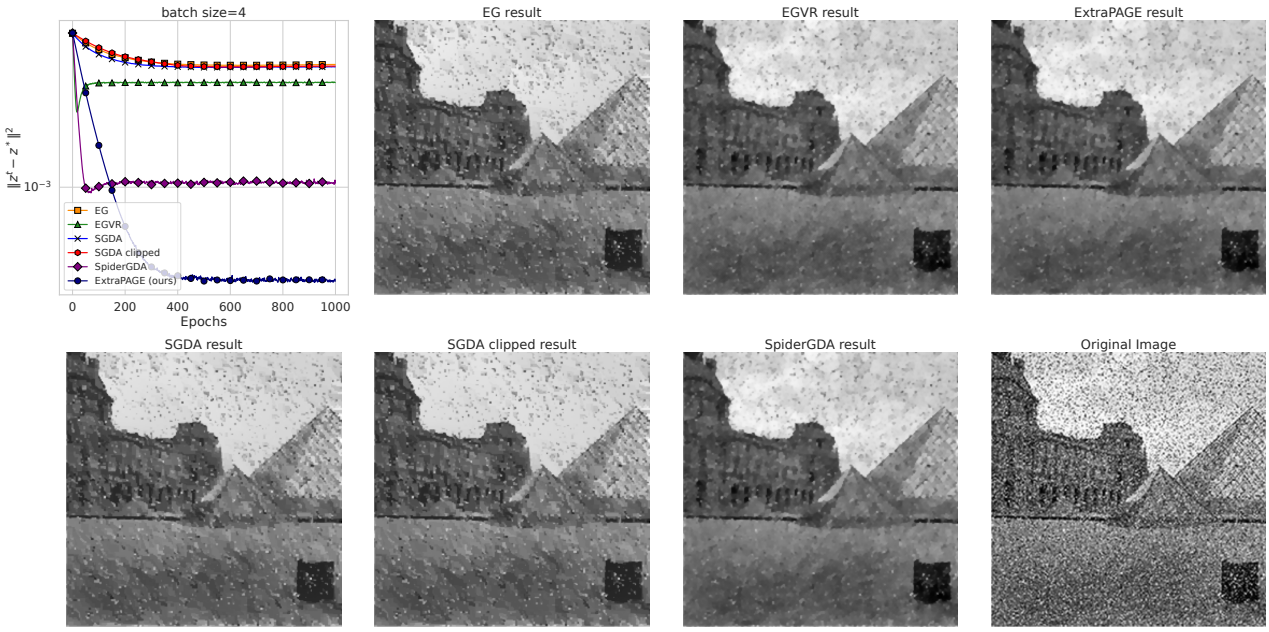


Figure 4: EXTRAPAGE and other baselines convergence on image with $\sigma = 0.1$ on the problem 12

The indicator function $\delta_P(p)$ is defined as zero if p belongs to the set P, and infinity otherwise. We define operator ∇u as the difference between neighboring pixels in the grid horizontally and vertically, normalizing by the step of the grid h. This formulation represents a saddle point problem, where we seek to minimize the first term with respect to u while simultaneously maximizing the second term with respect to p. Using duality, we can write the final formulation of considering problem as

$$\min_{u \in \mathcal{X}} \max_{p \in \mathcal{Y}} \left[-\langle u, \operatorname{div} p \rangle_{\mathcal{X}} + \lambda/2 \|u - g\|_2^2 - \delta_P(p) \right]. \quad (12)$$

We divide images into batches – equal squares. We consider two options: batches of size 4 and 8 according to the grid. Since the images are black and white, they are single-channel, which means that each batch is a square matrix with non-negative integers. It is also important to note that when calculating the gradient, the edges of the batch are processed according to the rule of adding a number equal to that of the nearest neighbor.

We select two images with different levels of additive zero-mean Gaussian noise: $\sigma = 0.05$ and $\sigma = 0.1$. Figure 4 provides a comparison of the proposed methods on the image with $\sigma = 0.1$. Additional results for all methods on another image are presented in Figure 7 in Appendix A.2.

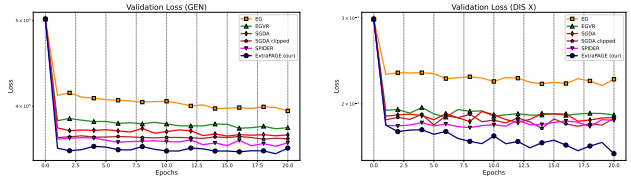
Comparing the images, it can be observed that EGVR demonstrates strong practical performance, with results that are nearly indistinguishable to the human eye from those of EXTRAPAGE. The slight difference lies in the loss behavior. We notice that EGVR performs slightly better than our algorithm during first epochs. Nonetheless, with continued training, the convergence rate of our algorithm surpasses superiority. Besides, while both EXTRAPAGE and EGVR converge well, EXTRAPAGE shows a smoother and more stable decline in error. In contrast, other methods struggle significantly with the problem, failing to reduce noise effectively. This can be attributed to its inherent limitations in handling variance-reduced stochastic updates, which are crucial for image denoising. Finally, compared to the original noisy image, all tested methods achieved significant noise reduction.

3.4 GAN TRAINING

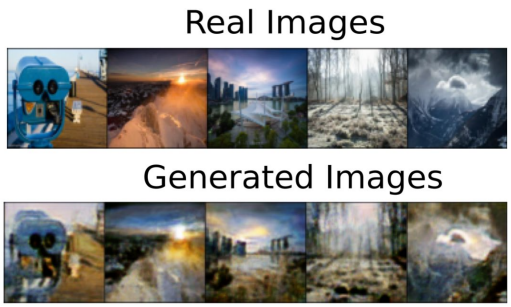
Generative Adversarial Networks (GANs) represent a powerful class of models widely applied in image generation tasks. StyleGAN [Karras, 2019] standing out for its ability to produce high-quality synthetic images. The adversarial nature of GAN training poses an $\min\max$ optimization problem, which can be effectively framed as a Variational Inequality. Thus, we explore the application of EXTRAPAGE in training a StyleGAN model for style translation.

We utilize the I'M SOMETHING OF A PAINTER MYSELF dataset, which consists of two distinct domains: a set of 300 Monet paintings and a set of 7028 photographs. Each image is resized to 256×256 pixels. We train the generator with an extrapolation step for both discriminators. We configure the training with a fixed learning rate $\gamma = 5 \times 10^{-5}$ and a batch size of 5, consistent across both domains. The probability parameter $p = \frac{1}{n}$ is set based on the effective dataset size, though for computational efficiency, we adapt it to the mini-batch context. Training is conducted for multiple random

initializations, specifically with random states 50 and 57. The results are presented in Figure 5.



Generator Loss Discriminator X Loss



Generated Images

Figure 5: All components use EXTRAPAGE with $\gamma = 5 \times 10^{-5}$ and batch size 5, random state 50

We provide additional results in Appendix A.4. As GANs represent one of the most prominent applications of VI algorithms in modern machine learning, EXTRAPAGE proves its applicability to a wide range of tasks.

4 DISCUSSION

In this paper, we present EXTRAPAGE, a novel algorithm for solving variational inequalities (VIs) and saddle point problems (SPPs). Our method is built upon variance-reduced algorithm PAGE and leads to superior theoretical convergence properties compared to baselines and slightly outperform them in practice. Additionally, our work closes an important gap in the theoretical understanding of SARAH-based methods applied to VIs and SPPs. Specifically, we derive a complexity bound with a linear dependence on the condition number of the problem under the assumption of Lipschitzness. Future research should refine theoretical bounds to establish the optimality of our method. What is more, further investigation into adaptive stepsize strategies could enhance the applicability of the method.

ACKNOWLEDGMENTS

The work was done in the Laboratory of Federated Learning Problems of the ISP RAS (Supported by Grant App. No. 2 to Agreement No. 075-03-2024-214).

REFERENCES

Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. *arXiv preprint arXiv:2102.08352*, 2021.

Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory*, pages 778–816. PMLR, 2022.

Ahmet Alacaoglu, Yura Malitsky, and Volkan Cevher. Forward-reflected-backward method with variance reduction. *Computational Optimization and Applications*, 80, 11 2021. doi:10.1007/s10589-021-00305-3.

Francis Bach, Julien Mairal, and Jean Ponce. Convex sparse matrix factorizations. *arXiv preprint arXiv:0812.1869*, 2008.

Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*, volume 28. Princeton university press, 2009.

A. Beznosikov, A. Gasnikov, K. Zainulina, A. Maslovskiy, and D. Pasechnyuk. A unified analysis of variational inequality methods: Variance reduction, sampling, quantization andcoordinate descent. *arXiv preprint arXiv:2201.12206*, 2022.

Aleksandr Beznosikov and Alexander Gasnikov. Sarah-based variance-reduced algorithm for stochastic finite-sum cocoercive variational inequalities. In *Data Analysis and Optimization: In Honor of Boris Mirkin's 80th Birthday*, pages 47–57. Springer, 2023.

Aleksandr Beznosikov, Boris Polyak, Eduard Gorbunov, Dmitry Kovalev, and Alexander Gasnikov. Smooth monotone stochastic variational inequalities and saddle point problems: A survey. *European Mathematical Society Magazine*, (127):15–28, 2023.

Aleksandr Beznosikov, Sergey Samsonov, Marina Sheshukova, Alexander Gasnikov, Alexey Naumov, and Eric Moulines. First order methods with markovian noise: from acceleration to variational inequalities. *Advances in Neural Information Processing Systems*, 36, 2024.

Léon Bottou. Curiously fast convergence of some stochastic gradient descent algorithms. In *Proceedings of the symposium on learning and data science, Paris*, volume 8, pages 2624–2633. Citeseer, 2009.

Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010: 19th International Conference on Computational StatisticsParis France, August 22-27, 2010 Keynote, Invited and Contributed Papers*, pages 177–186. Springer, 2010.

Felix E Browder. Nonexpansive nonlinear operators in a banach space. *Proceedings of the National Academy of Sciences*, 54(4):1041–1044, 1965.

Antonin Chambolle and Thomas Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of mathematical imaging and vision*, 40:120–145, 2011.

Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

Tatjana Chavdarova, Gauthier Gidel, François Fleuret, and Simon Lacoste-Julien. Reducing noise in gan training with variance reduced extragradient. *Advances in Neural Information Processing Systems*, 32, 2019.

Lesi Chen, Boyuan Yao, and Luo Luo. Faster stochastic algorithms for minimax optimization under polyak-{\L} ojasiewicz condition. *Advances in Neural Information Processing Systems*, 35:13921–13932, 2022.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex sgd. *arXiv preprint arXiv:1905.10018*, 2019.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Marc'aurelio Ranzato, Andrew Senior, Paul Tucker, Ke Yang, et al. Large scale distributed deep networks. *Advances in neural information processing systems*, 25, 2012.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014a.

Aaron Defazio, Justin Domke, et al. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning*, pages 1125–1133. PMLR, 2014b.

Ernie Esser, Xiaoqun Zhang, and Tony F Chan. A general framework for a class of first order primal-dual algorithms for convex optimization in imaging science. *SIAM Journal on Imaging Sciences*, 3(4):1015–1046, 2010.

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

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. *Advances in neural information processing systems*, 31, 2018.

Gauthier Gidel, Hugo Berard, Gaëtan Vignoud, Pascal Vincent, and Simon Lacoste-Julien. A variational inequality perspective on generative adversarial networks. *arXiv preprint arXiv:1802.10551*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Eduard Gorbunov, Hugo Berard, Gauthier Gidel, and Nicolas Loizou. Stochastic extragradient: General analysis and improved rates. In *International Conference on Artificial Intelligence and Statistics*, pages 7865–7901. PMLR, 2022.

Robert M Gower, Mark Schmidt, Francis Bach, and Peter Richtárik. Variance-reduced methods for machine learning. *Proceedings of the IEEE*, 108(11):1968–1983, 2020.

Patrick T Harker and Jong-Shi Pang. Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. *Mathematical programming*, 48(1):161–220, 1990.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. On the convergence of single-call stochastic extra-gradient methods. In *Advances in Neural Information Processing Systems*, volume 32, pages 6938–6948. Curran Associates, Inc., 2019.

Yu-Guan Hsieh, Franck Iutzeler, Jérôme Malick, and Panayotis Mertikopoulos. Explore aggressively, update conservatively: Stochastic extragradient methods with variable stepsize scaling. volume 33, pages 16223–16234, 2020.

Wenqing Hu, Chris Junchi Li, Xiangru Lian, Ji Liu, and Huizhuo Yuan. Efficient smooth non-convex stochastic compositional optimization via stochastic recursive gradient descent. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. *Advances in neural information processing systems*, 26, 2013.

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

Tero Karras. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2019.

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

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

Dmitry Kovalev, Samuel Horváth, and Peter Richtárik. Don't jump through hoops and remove those loops: Svrg and katyusha are better without the outer loop. In *Algorithmic Learning Theory*, pages 451–467. PMLR, 2020.

Dmitry Kovalev, Aleksandr Beznosikov, Abdurakhmon Sadiev, Michael Persiianov, Peter Richtárik, and Alexander Gasnikov. Optimal algorithms for decentralized stochastic variational inequalities. *Advances in Neural Information Processing Systems*, 35:31073–31088, 2022.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Zhize Li, Hongyan Bao, Xiangliang Zhang, and Peter Richtárik. Page: A simple and optimal probabilistic gradient estimator for nonconvex optimization. In *International Conference on Machine Learning*, pages 6286–6295. PMLR, 2021.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 907–915. PMLR, 2019.

Nicolas Loizou, Hugo Berard, Gauthier Gidel, Ioannis Mitliagkas, and Simon Lacoste-Julien. Stochastic gradient descent-ascent and consensus optimization for smooth games: Convergence analysis under expected co-coercivity. *Advances in Neural Information Processing Systems*, 34:19095–19108, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Yu Malitsky. Projected reflected gradient methods for monotone variational inequalities. *SIAM Journal on Optimization*, 25(1):502–520, 2015.

Yura Malitsky and Matthew K Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020.

Daniil Medyakov, Gleb Molodtsov, Evseev Grigoriy, Egor Petrov, and Aleksandr Beznosikov. Shuffling heuristic in variational inequalities: Establishing new convergence guarantees. In *International Conference on Computational Optimization*, 2024.

Panayotis Mertikopoulos, Bruno Lecouat, Houssam Zenati, Chuan-Sheng Foo, Vijay Chandrasekhar, and Georgios Piliouras. Optimistic mirror descent in saddle-point problems: Going the extra (gradient) mile. *arXiv preprint arXiv:1807.02629*, 2018.

Konstantin Mishchenko, Dmitry Kovalev, Egor Shulgin, Peter Richtárik, and Yura Malitsky. Revisiting stochastic extragradient. In *International Conference on Artificial Intelligence and Statistics*, pages 4573–4582. PMLR, 2020.

Aryan Mokhtari, Asuman Ozdaglar, and Sarath Pattathil. A unified analysis of extra-gradient and optimistic gradient methods for saddle point problems: Proximal point approach. In *International Conference on Artificial Intelligence and Statistics*, pages 1497–1507. PMLR, 2020.

Eric Moulines and Francis Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in neural information processing systems*, 24, 2011.

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, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

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

Yurii Nesterov. Dual extrapolation and its applications to solving variational inequalities and related problems. *Mathematical Programming*, 109(2):319–344, 2007.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In *International conference on machine learning*, pages 2613–2621. PMLR, 2017.

Shayegan Omidshafiei, Jason Pazis, Christopher Amato, Jonathan P How, and John Vian. Deep decentralized multi-task multi-agent reinforcement learning under partial observability. In *International Conference on Machine Learning*, pages 2681–2690. PMLR, 2017.

Balamurugan Palaniappan and Francis Bach. Stochastic variance reduction methods for saddle-point problems. *Advances in Neural Information Processing Systems*, 29, 2016.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Wei Peng, Yu-Hong Dai, Hui Zhang, and Lizhi Cheng. Training gans with centripetal acceleration. *Optimization Methods and Software*, 35(5):955–973, 2020.

Xun Qian, Zheng Qu, and Peter Richtárik. Saga with arbitrary sampling. In *International Conference on Machine Learning*, pages 5190–5199. PMLR, 2019.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Nicolas Roux, Mark Schmidt, and Francis Bach. A stochastic gradient method with an exponential convergence _rate for finite training sets. *Advances in neural information processing systems*, 25, 2012.

Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

Vladimir Solodkin, Andrew Veprikov, and Aleksandr Beznosikov. Methods for optimization problems with markovian stochasticity and non-euclidean geometry. *arXiv preprint arXiv:2408.01848*, 2024.

Guido Stampacchia. Formes bilineaires coercitives sur les ensembles convexes. *Comptes Rendus Hebdomadaires Des Seances De L Academie Des Sciences*, 258(18):4413, 1964.

Paul Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

John Von Neumann and Oskar Morgenstern. *Theory of games and economic behavior: by J. Von Neumann and O. Morgenstern*. Princeton university press, 1953.

Linli Xu, James Neufeld, Bryce Larson, and Dale Schuurmans. Maximum margin clustering. *Advances in neural information processing systems*, 17, 2004.

When Extragradient Meets PAGE: Bridging Two Giants to Boost Variational Inequalities (Supplementary Material)

Gleb Molodtsov[1,2] **Valery Parfenov**[1] **Egor Petrov**[1] **Evseev Grigoriy**[1] **Daniil Medyakov**[1,2] **Aleksandr Beznosikov**[2,1,3]

[1]Moscow Institute of Physics and Technology
[2]Ivannikov Institute for System Programming of the RAS
[3]Innopolis University

CONTENTS

A ADDITIONAL EXPERIMENTS

In this section, we present additional experiments that have been performed as well as the technical details for them.

A.1 ADVERSARIAL TRAINING

We address an adversarial training problem. We can formulate it the way as in (7).

We evaluate this issue across several datasets: `mushrooms`, `a9a`, `w8a`, and `ijcnn1`, sourced from the LIBSVM library [Chang and Lin, 2011]. A brief description of these datasets is provided in Table 2. The results are presented in Figure 6.

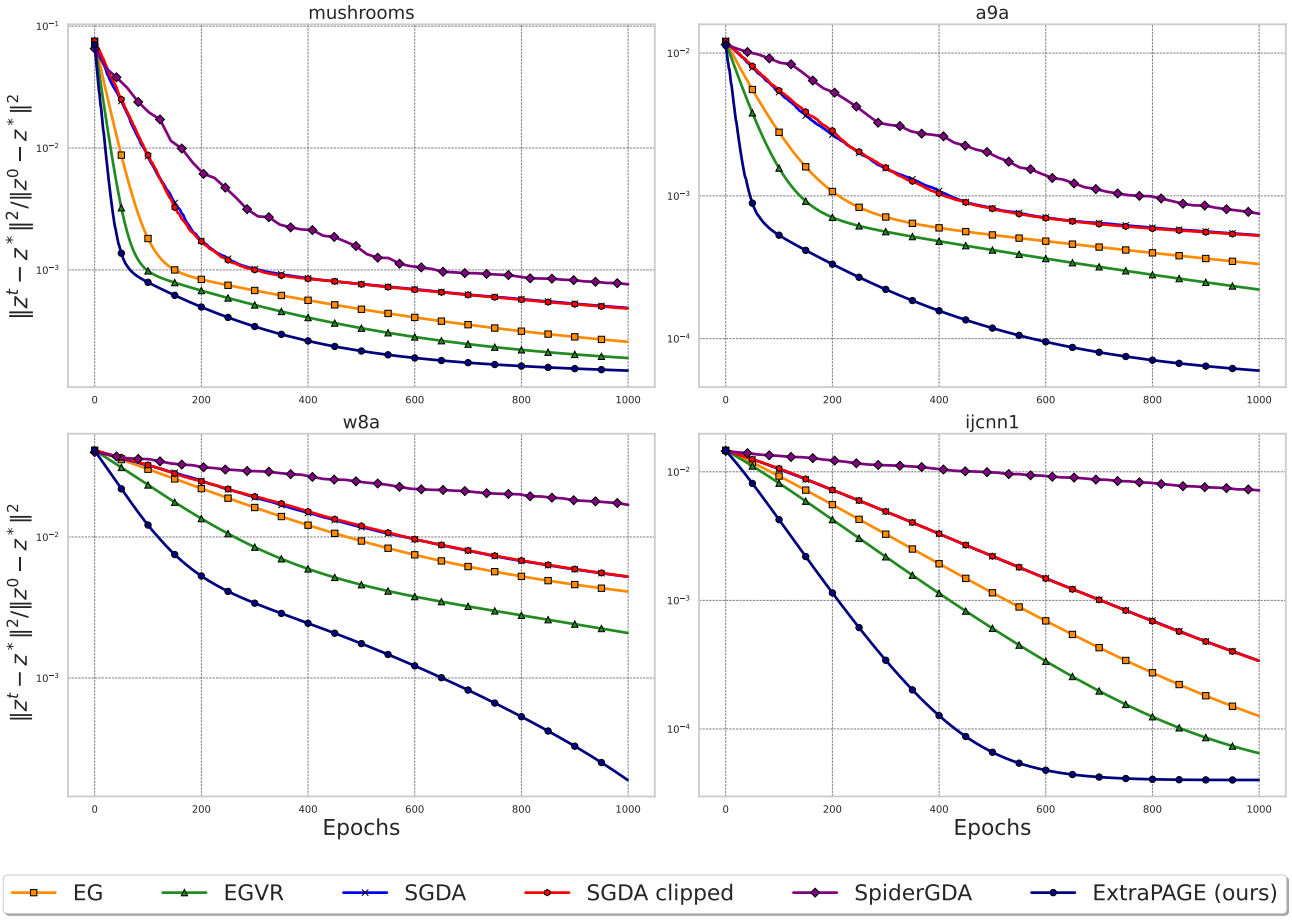


Figure 6: EXTRAPAGE compared to different baselines on `mushrooms`, `a9a`, `w8a`, and `ijcnn1` datasets on the problem (7).

As shown on plots, EXTRAPAGE consistently outperforms other methods across all datasets (`mushrooms`, `a9a`, `w8a`, `ijcnn1`). These datasets vary in size and complexity, providing a comprehensive evaluation of our proposed algorithms in the context of adversarial training.

Table 2: Summary of Datasets

Name	**Number of Instances**	**Number of Features**	**Number of Classes**
`mushrooms`	8,124	112	2
`a9a`	32,561	123	2
`w8a`	49,749	300	2
`ijcnn1`	49,990	22	2

A.2 IMAGE DENOISING

In this section, we present additional experiments conducted on image denoising. Consistent with our previous findings, a notable pattern emerges: although EGVR initially converges faster, EXTRAPAGE exhibits more stable convergence over time, ultimately reaching a more precise minimum. Both methods significantly outperform the vanilla EXTRAGRADIENT approach on this task. These results further reinforce the effectiveness of variance reduction techniques when applied to another image with noise level $\sigma = 0.05$, underscoring their utility in solving denoising problems.

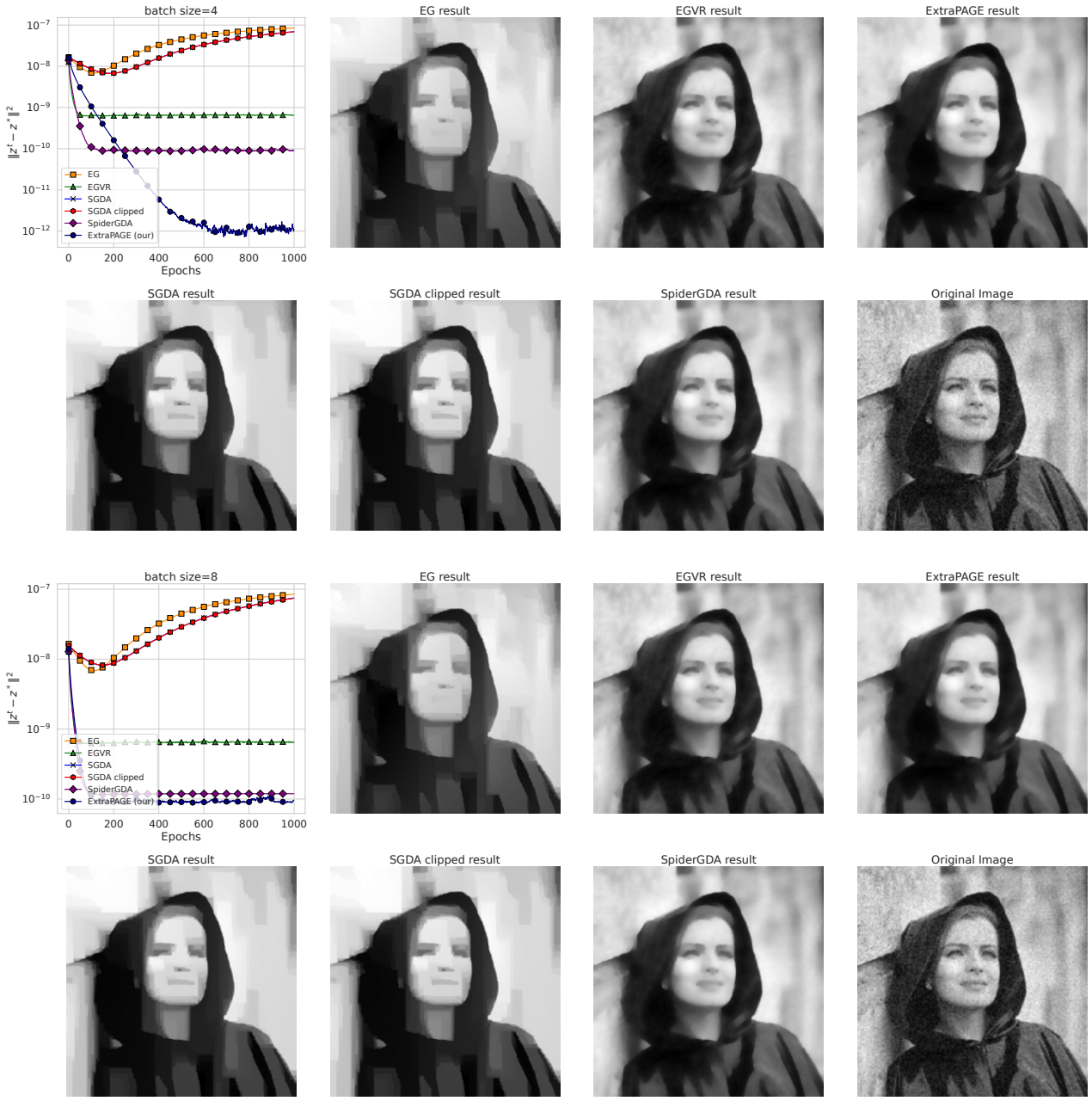


Figure 7: EXTRAPAGE and other baselines convergence on image with $\sigma = 0.05$ on the problem 12 with batch sizes $\in \{4, 8\}$

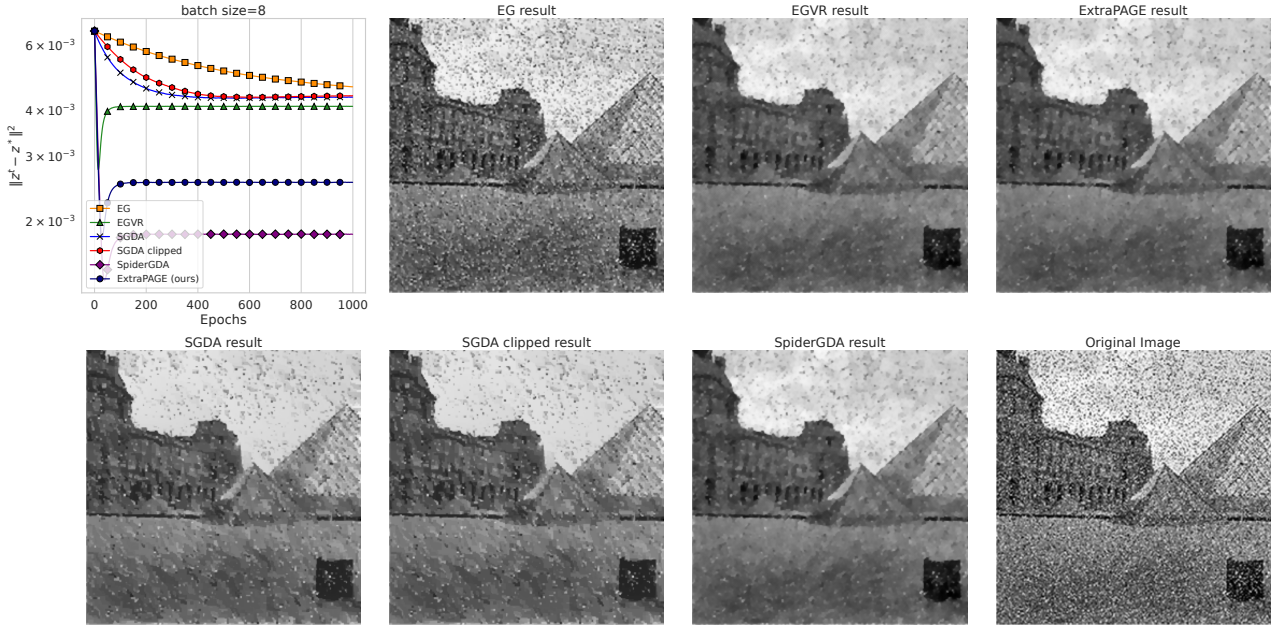


Figure 8: EXTRAPAGE and other baselines convergence on image with $\sigma = 0.1$ on the problem 12

A.3 IMAGE CLASSIFICATION

This experiment was conducted on the CIFAR-10 dataset [Krizhevsky et al., 2009], widely used as a benchmark in optimization community, consisting of 50,000 training and 10,000 test samples. Each sample is a 32×32 RGB image associated with one of ten class labels. The experiments were implemented in Python using the PyTorch library [Paszke et al., 2019], leveraging both a single CPU (Intel Xeon 2.20 GHz) and a single GPU (NVIDIA Tesla P100) for computation.

The experiments are conducted with the following setup:

- learning rate $\gamma = 0.01$ for all optimizers;
- regularization parameters $\lambda = \beta = 0.0005$.

Below, we present a series of convergence plots and corresponding runtime tables that illustrate the performance of EXTRAPAGE under different distributed settings. Each figure shows the optimization trajectory for EXTRAPAGE compared to baseline methods when varying the number of workers n and the update probability p. The tables summarize total training time and per-round timing statistics for each configuration, allowing a clear comparison of efficiency and stability across all tested scenarios.

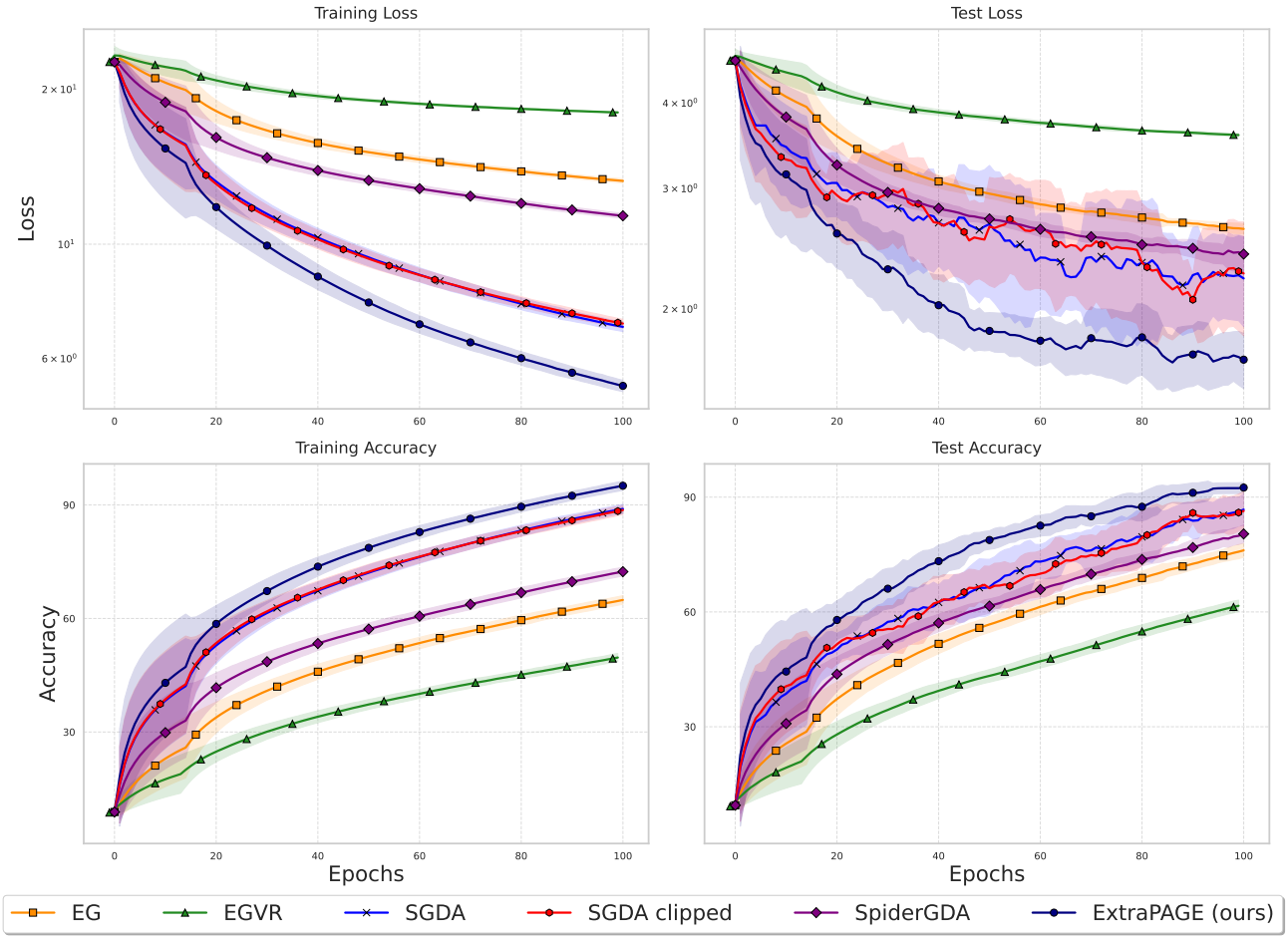


Figure 9: EXTRAPAGE compared to different baselines on `CIFAR` dataset. We choose $n = 100, p = \frac{1}{2n}$.

Algorithm	Total Time	Round Time
EG	467.447	4.674 ± 0.020
EGVR	618.560	6.186 ± 0.009
SGDA	433.672	4.337 ± 0.043
SGDA clipped	438.911	4.389 ± 0.252
SPIDER	841.750	8.417 ± 0.761
ExtraPAGE	633.366	6.334 ± 0.009

Table 3: Runtime comparison of our algorithms with $n = 100, p = \frac{1}{2n}$.

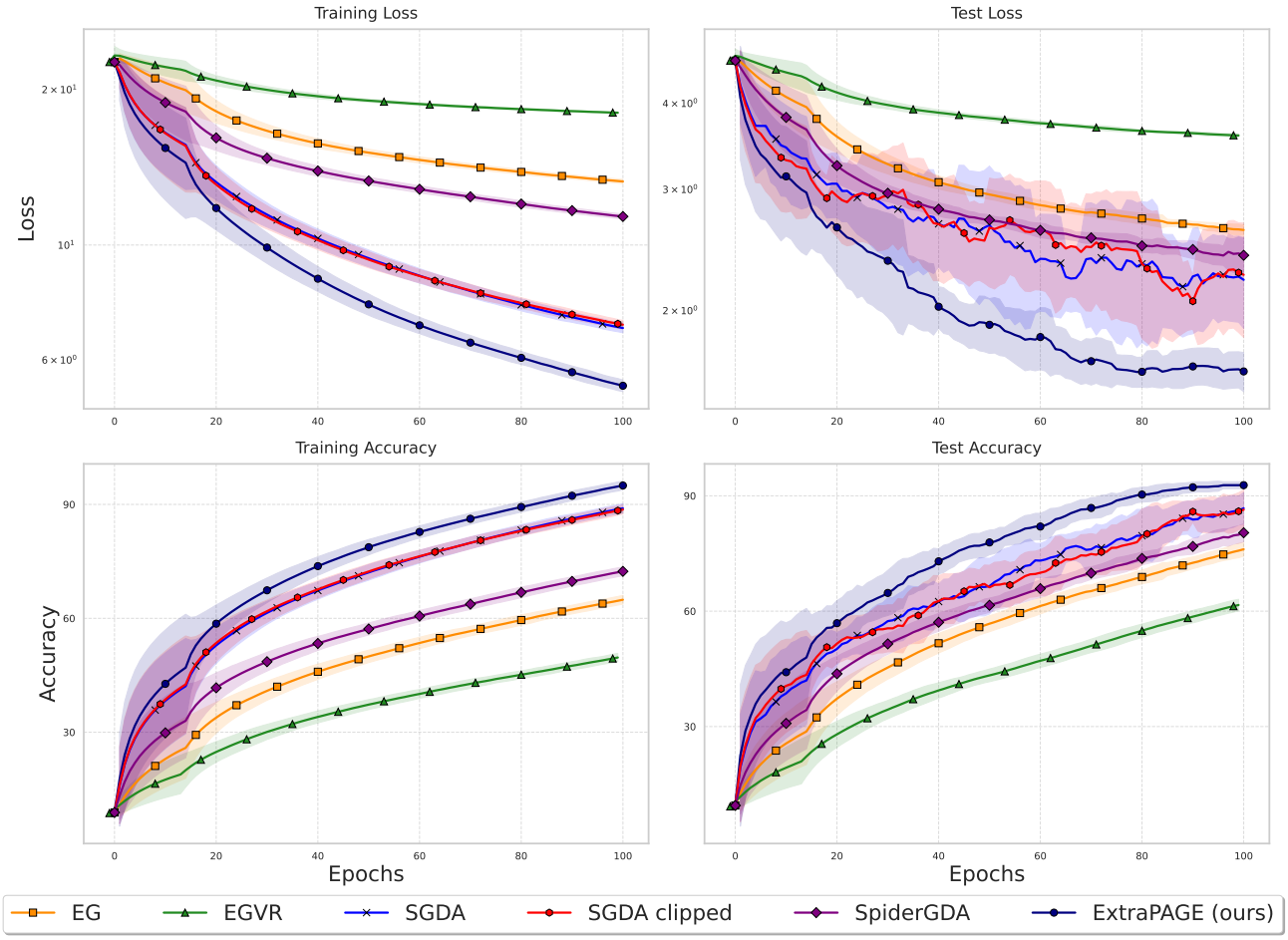


Figure 10: EXTRAPAGE compared to different baselines on `CIFAR` dataset. We choose $n = 100, p = \frac{2}{n}$.

Algorithm	Total Time	Round Time
EG	467.447	4.674 ± 0.020
EGVR	618.560	6.186 ± 0.009
SGDA	433.672	4.337 ± 0.043
SGDA clipped	438.911	4.389 ± 0.252
SPIDER	841.750	8.417 ± 0.761
ExtraPAGE	634.688	6.347 ± 0.011

Table 4: Runtime comparison of our algorithms with $n = 100, p = \frac{2}{n}$.

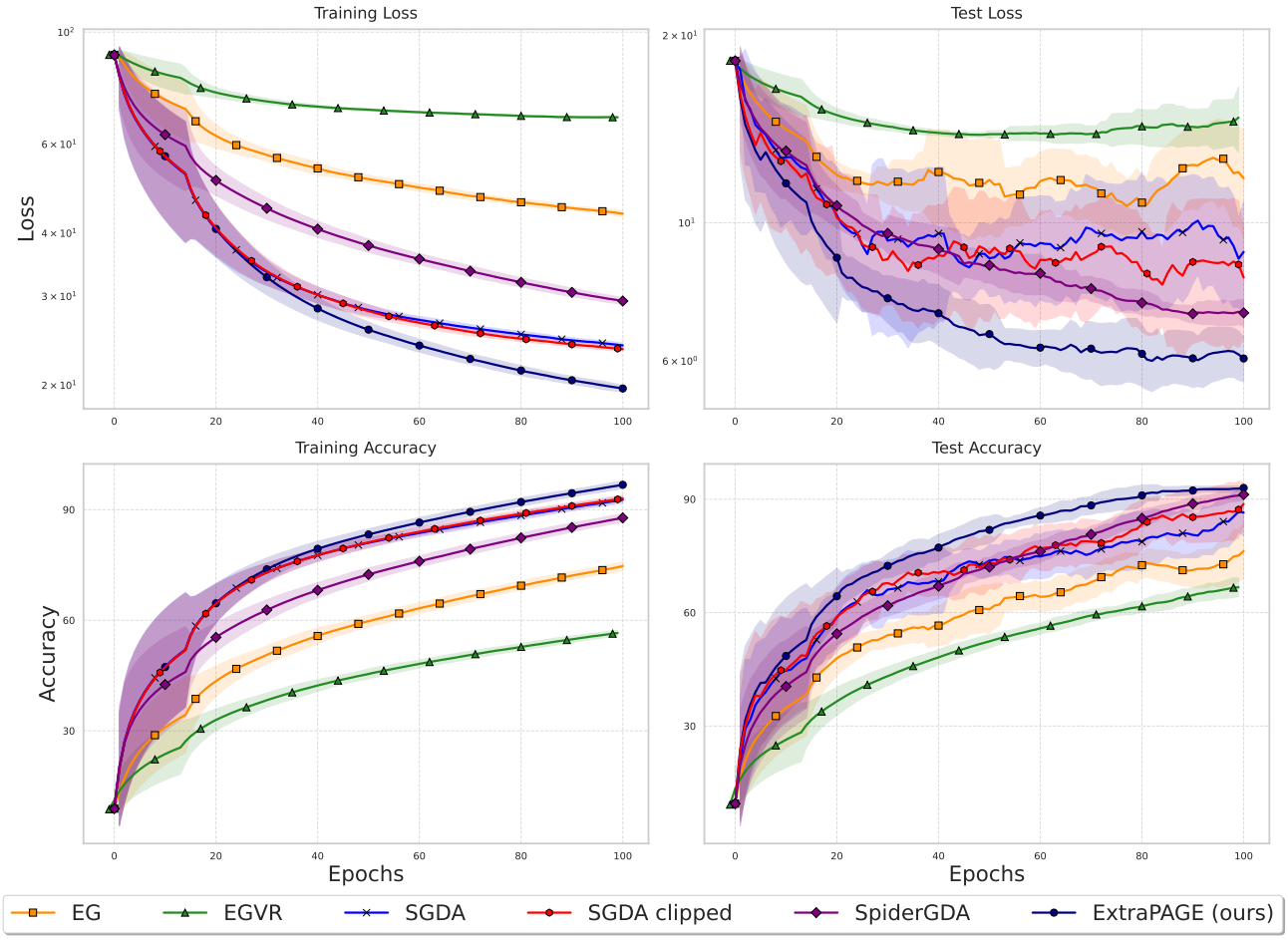


Figure 11: EXTRAPAGE compared to different baselines on CIFAR dataset. We choose $n = 390, p = \frac{1}{n}$.

Algorithm	Total Time	Round Time
EG	692.169	6.922 ± 0.292
EGVR	1129.090	11.291 ± 0.013
SGDA	526.992	5.270 ± 1.162
SGDA clipped	516.918	5.169 ± 1.154
SPIDER	1133.076	11.331 ± 4.043
ExtraPAGE	1178.660	11.787 ± 0.028

Table 5: Runtime comparison of our algorithms with $n = 390, p = \frac{1}{n}$.

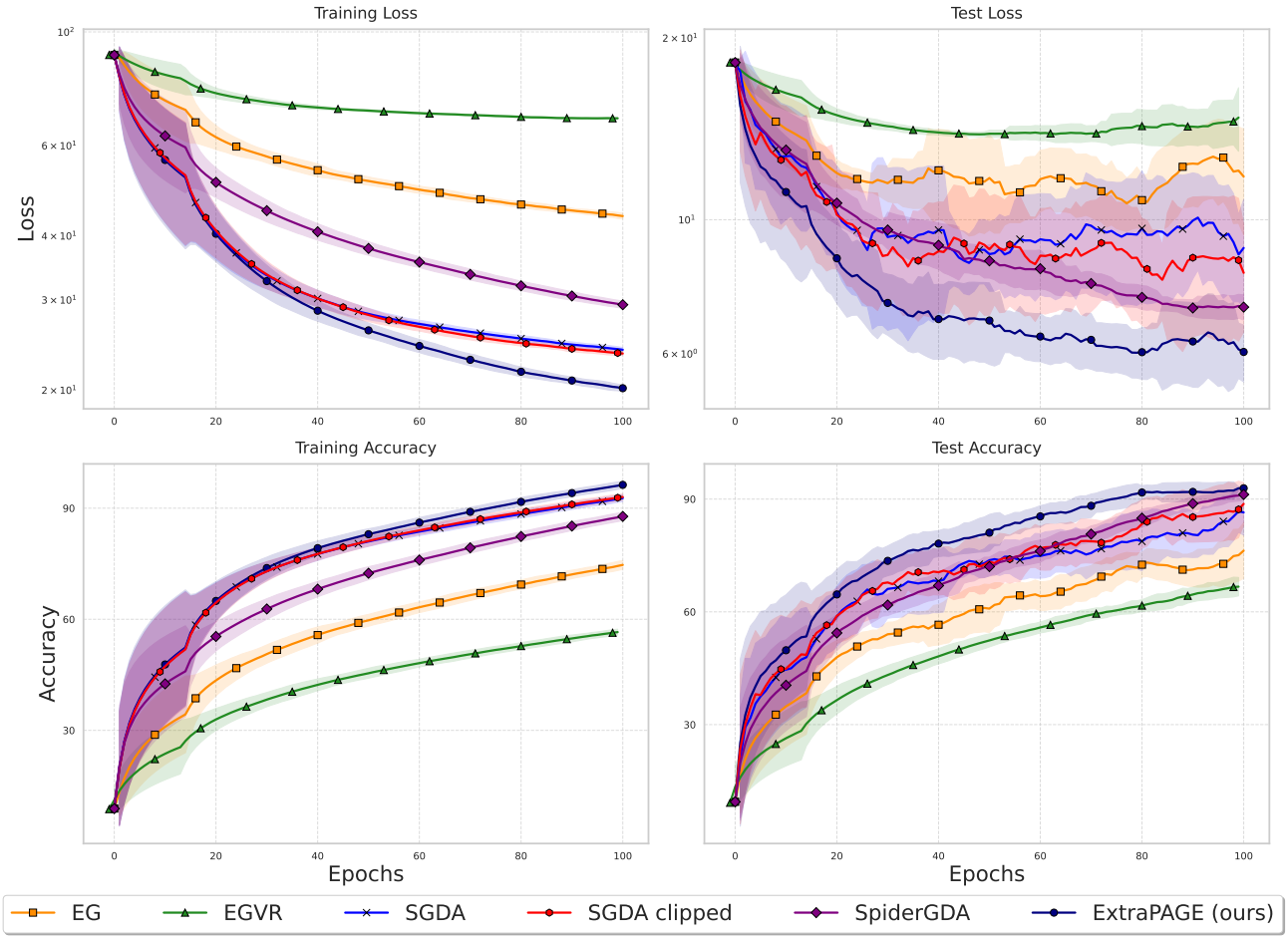


Figure 12: EXTRAPAGE compared to different baselines on CIFAR dataset. We choose $n = 390, p = \frac{1}{2n}$.

Algorithm	Total Time	Round Time
EG	692.169	6.922 ± 0.292
EGVR	1129.090	11.291 ± 0.013
SGDA	526.992	5.270 ± 1.162
SGDA clipped	516.918	5.169 ± 1.154
SPIDER	1133.076	11.331 ± 4.043
ExtraPAGE	1177.751	11.778 ± 0.152

Table 6: Runtime comparison of our algorithms with $n = 390, p = \frac{1}{2n}$.

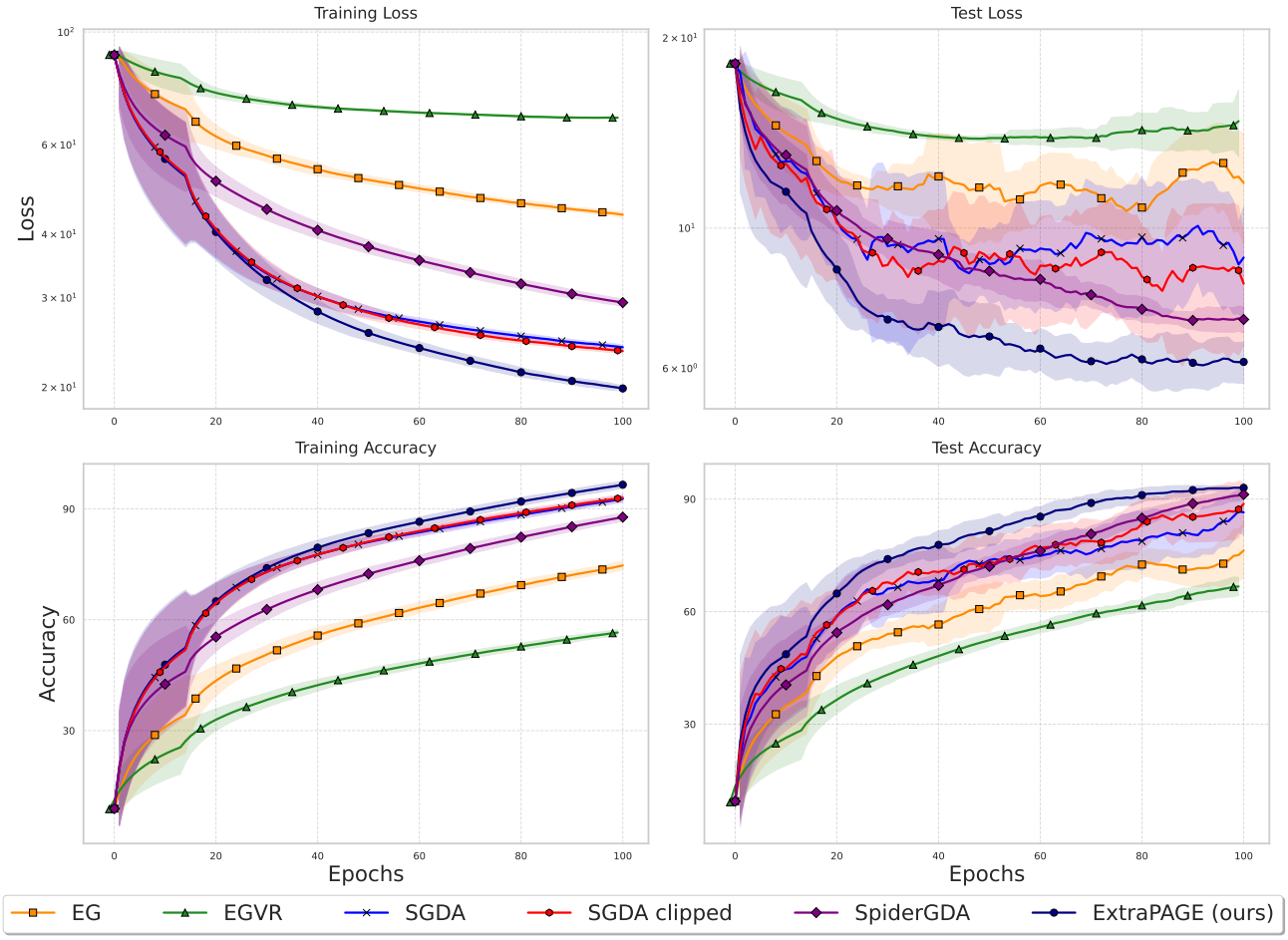


Figure 13: EXTRAPAGE compared to different baselines on CIFAR dataset. We choose $n = 390, p = \frac{2}{n}$.

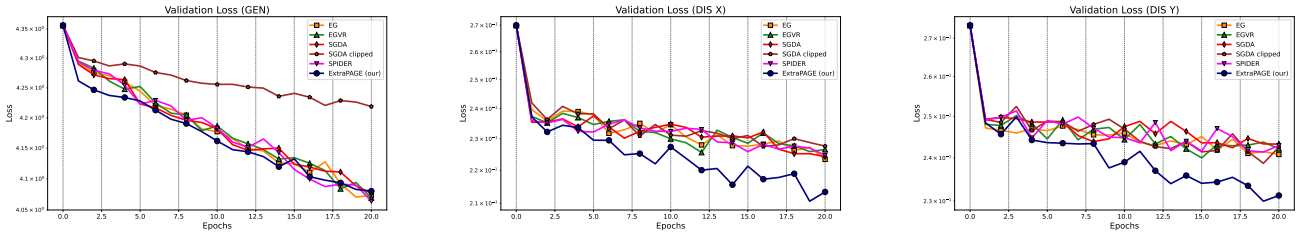
Algorithm	Total Time	Round Time
EG	692.169	6.922 ± 0.292
EGVR	1129.090	11.291 ± 0.013
SGDA	526.992	5.270 ± 1.162
SGDA clipped	516.918	5.169 ± 1.154
SPIDER	1133.076	11.331 ± 4.043
ExtraPAGE	1173.841	11.738 ± 0.023

Table 7: Runtime comparison of our algorithms with $n = 390, p = \frac{2}{n}$.

Based on the conducted experiments, EXTRAPAGE consistently demonstrates robust and stable convergence regardless of the number of workers n or the update probability p. Through experiments with varying n, we confirm that EXTRAPAGE outperforms existing benchmarks in ill-conditioned problems and high condition number scenarios relative to batch count, while spending comparable time for each iteration. Parameter analysis further shows minimal sensitivity to p, indicating that EXTRAPAGE effectively balances computation and communication overhead without degradation in efficiency. Overall, these results highlight EXTRAPAGE's stability and insensitivity to both n and p.

A.4 GAN TRAINING

In this section, we provide additional experiments with random state 57. The experiments were held on a single GPU (NVIDIA A100). The results are presented in Figure 14.



Generator Loss | Discriminator X Loss | Discriminator Y Loss

Figure 14: All components use EXTRAPAGE with $\gamma = 5 \times 10^{-5}$ and batch size 5, random state 57

B GENERAL INEQUALITIES

First, we mention important inequalities that are used in further proofs. Consider a function f satisfying Assumption 1. Then for any n in the real numbers and for all vectors x, y, x_i in $\mathbb{R}^n$ with a positive scalar c, the following inequalities hold.

$$|\langle x, y\rangle| \leqslant \frac{\|x\|^2}{2c} + \frac{c\|y\|^2}{2} \tag{Young}$$

$$\begin{aligned} -\langle x, y\rangle &= -\frac{\|x\|^2}{2} - \frac{\|y\|^2}{2} + \frac{\|x-y\|^2}{2} \\ \|x+y\|^2 &= \|x\|^2 + \|y\|^2 + 2\langle x, y\rangle \end{aligned} \tag{Norm}$$

$$\begin{aligned} \|\nabla f(x) - \nabla f(y)\|^2 &\leqslant L^2\|x-y\|^2 \\ f(x) &\leqslant f(y) + \langle \nabla f(y), x-y\rangle + \frac{L}{2}\|x-y\|^2 \\ f(x) &\leqslant f(y) - \langle \nabla f(x), y-x\rangle - \frac{1}{2L}\|\nabla f(x) - \nabla f(y)\|^2 \end{aligned} \tag{Lip}$$

$$\begin{aligned} \left\|\sum_{i=1}^n x_i\right\|^2 &\leqslant n\sum_{i=1}^n \|x_i\|^2 \\ \|x+y\|^2 &\leqslant (1+c)\|x\|^2 + \left(1+\frac{1}{c}\right)\|y\|^2 \end{aligned} \tag{CS}$$

C PROOF

In this section, we provide all the necessary proofs. First, we prove Lemmas 1, 2 and 3. Then, we use them to establish the main Theorem 1, which guarantees the convergence of our algorithm. Finally, we derive Corollaries 1 and 2.

We begin with Lemma 1, which reflects the change in the quality of the gradient estimation from one iteration to the next. This lemma is crucial not only for the subsequent analysis but also has independent significance in developing intuition about variance reduction methods.

Lemma 1. *For iterations of Algorithm 1 the following equation holds:*

$$\mathbb{E}_{i^t}\mathbb{E}_{G^{t-1}}\left\|F(z^{t+1/2})-G^t\right\|^2=(1-p)\left[\left\|F(z^{t-1/2})-G^{t-1}\right\|^2+\mathbb{E}_{i^t}\left\|F_{i^t}\left(z^{t+1/2}\right)-F_{i^t}\left(z^{t-1/2}\right)\right\|^2-\left\|F\left(z^{t+1/2}\right)-F\left(z^{t-1/2}\right)\right\|^2\right].$$

Proof. We examine the following term using the update rule (Line 5). We take the expectation over G^{t-1}:

$$\begin{aligned}\mathbb{E}_{G^{t-1}}\left\|F(z^{t+1/2})-G^t\right\|^2 &= (1-p)\left\|F(z^{t+1/2})-F_{i^t}(z^{t+1/2})+F_{i^t}(z^{t-1/2})-G^{t-1}\right\|^2\\ &= (1-p)\left\|F(z^{t+1/2})-F_{i^t}(z^{t+1/2})+F_{i^t}(z^{t-1/2})-F(z^{t-1/2})+F(z^{t-1/2})-G^{t-1}\right\|^2\\ &= (1-p)\left\|F(z^{t+1/2})-F_{i^t}(z^{t+1/2})+F_{i^t}(z^{t-1/2})-F(z^{t-1/2})\right\|^2\\ &\quad+(1-p)\left\|F(z^{t-1/2})-G^{t-1}\right\|^2\\ &\quad+2(1-p)\left\langle F(z^{t+1/2})-F_{i^t}(z^{t+1/2})+F_{i^t}(z^{t-1/2})-F(z^{t-1/2}),F(z^{t-1/2})-G^{t-1}\right\rangle.\end{aligned}$$

Taking the expectation over i^t and utilizing $\mathbb{E}_{i^t}\left[F_{i^t}(z)\right]=F(z)$, we derive that the scalar product is equal to zero. We continue the equation:

$$\begin{aligned}\mathbb{E}_{i^t}\mathbb{E}_{G^{t-1}}\left\|F(z^{t+1/2})-G^t\right\|^2 &= (1-p)\mathbb{E}_{i^t}\left\|F(z^{t+1/2})-F_{i^t}(z^{t+1/2})+F_{i^t}(z^{t-1/2})-F(z^{t-1/2})\right\|^2\\ &\quad+(1-p)\left\|F(z^{t-1/2})-G^{t-1}\right\|^2\\ &= (1-p)\left\|F(z^{t+1/2})-F(z^{t-1/2})\right\|^2+(1-p)\mathbb{E}_{i^t}\left\|F_{i^t}(z^{t+1/2})-F_{i^t}(z^{t-1/2})\right\|^2\\ &\quad-2(1-p)\mathbb{E}_{i^t}\left\langle F(z^{t+1/2})-F(z^{t-1/2}),F_{i^t}(z^{t+1/2})-F_{i^t}(z^{t-1/2})\right\rangle\\ &\quad+(1-p)\left\|F(z^{t-1/2})-G^{t-1}\right\|^2\\ &= -(1-p)\left\|F(z^{t+1/2})-F(z^{t-1/2})\right\|^2\\ &\quad+(1-p)\mathbb{E}_{i^t}\left\|F_{i^t}(z^{t+1/2})-F_{i^t}(z^{t-1/2})\right\|^2+(1-p)\left\|F(z^{t-1/2})-G^{t-1}\right\|^2.\end{aligned}$$

This finishes the proof of the lemma. □

We emphasize that Lemma 1 is formulated as an equation without any assumptions on the operator F, making it an exact and general estimate. Now, we proceed to the proof of Lemma 2 which is a general statement of Euclidean geometry and serves as the first step in deriving the recursion.

Lemma 2. *Let $z,y\in\mathbb{R}^d$, $z^+=z-y$. Then for all $u\in\mathbb{R}^d$ the following equation holds:*

$$\|z^+-u\|^2=\|z-u\|^2-2\langle y,z^+-u\rangle-\|z^+-z\|^2.$$

Proof. We transform the left part as follows:

$$\begin{aligned}\|z^+-u\|^2 &= \|z^+-z+z-u\|^2\\ &= \|z-u\|^2+2\langle z^+-z,z-u\rangle+\|z^+-z\|^2\\ &= \|z-u\|^2+2\langle z^+-z,z^+-u\rangle-\|z^+-z\|^2\\ &= \|z-u\|^2-2\langle y,z^+-u\rangle-\|z^+-z\|^2.\end{aligned}$$

This finishes the proof of the lemma. □

Lemma 3 serves as the final prerequisite for the main analysis in Theorem 1 and relies on the assumption of the strong monotonicity of the operator F to derive a recursion for the term $\|z^t - z^*\|^2$.

Lemma 3 (Descent lemma)**.** *Under Assumption 2(a), after T iterations of Algorithm 1 the following equation holds:*

$$\begin{aligned}\left\|z^{t+1} - z^*\right\|^2 &= (1-\gamma\mu)\left\|z^t - z^*\right\|^2 - (1-2\gamma\mu)\left\|z^{t+1/2} - z^t\right\|^2 \\ &\quad + 2\gamma\left\langle F(z^{t+1/2}) - G^t, z^{t+1/2} - z^*\right\rangle + \gamma^2\left\|G^t - G^{t-1}\right\|^2.\end{aligned}$$

Proof. We substitute $z = z^t, y = \gamma G^t, z^+ = z^{t+1}, u = z^*$, and $z = z^t, y = \gamma G^{t-1}, z^+ = z^{t+1/2}, u = z^{t+1}$ into Lemma 2 and summing the obtained equations. It yields

$$\begin{aligned}\left\|z^{t+1} - z^*\right\|^2 + \left\|z^{t+1/2} - z^{t+1}\right\|^2 &= \left\|z^t - z^*\right\|^2 - \left\|z^{t+1/2} - z^t\right\|^2 \\ &\quad - 2\gamma\left\langle G^t, z^{t+1} - z^*\right\rangle - 2\gamma\left\langle G^{t-1}, z^{t+1/2} - z^{t+1}\right\rangle \\ &= \left\|z^t - z^*\right\|^2 - \left\|z^{t+1/2} - z^t\right\|^2 \\ &\quad - 2\gamma\left\langle G^t, z^{t+1/2} - z^*\right\rangle - 2\gamma\left\langle G^{t-1} - G^t, z^{t+1/2} - z^{t+1}\right\rangle. \qquad (13)\end{aligned}$$

Now we examine the first scalar product:

$$-2\gamma\left\langle G^t, z^{t+1/2} - z^*\right\rangle = 2\gamma\left\langle F(z^{t+1/2}) - G^t, z^{t+1/2} - z^*\right\rangle - 2\gamma\left\langle F(z^{t+1/2}), z^{t+1/2} - z^*\right\rangle.$$

Under the setting (1) and the strong monotonicity (Assumption 2(a)), the last term transforms into

$$\begin{aligned}-2\gamma\left\langle F(z^{t+1/2}), z^{t+1/2} - z^*\right\rangle &= -2\gamma\left\langle F(z^{t+1/2}) - F(z^*), z^{t+1/2} - z^*\right\rangle - 2\gamma\left\langle F(z^*), z^{t+\frac{1}{2}} - z^*\right\rangle \\ &\leqslant -2\gamma\mu\left\|z^{t+1/2} - z^*\right\|^2 \leqslant -\gamma\mu\left\|z^t - z^*\right\|^2 + 2\gamma\mu\left\|z^{t+1/2} - z^t\right\|^2,\end{aligned}$$

where the last inequality utilizes (CS). Finally, we obtain

$$-2\gamma\left\langle G^t, z^{t+1/2} - z^*\right\rangle \leqslant 2\gamma\left\langle F(z^{t+1/2}) - G^t, z^{t+1/2} - z^*\right\rangle - \gamma\mu\left\|z^t - z^*\right\|^2 + 2\gamma\mu\left\|z^{t+1/2} - z^t\right\|^2. \qquad (14)$$

Now it is sufficient to note that according to Lines 6 and 4, $\gamma(G^{t-1} - G^t) = z^{t+1} - z^{t+1/2}$, which transforms the second scalar product in (13) into

$$-2\gamma\left\langle G^{t-1} - G^t, z^{t+1/2} - z^{t+1}\right\rangle = 2\gamma^2\left\|G^t - G^{t-1}\right\|^2. \qquad (15)$$

Substituting (15) and (14) into (13) finishes the proof of the lemma. □

Now, we are fully prepared to present the main analysis in Theorem 1. We begin with Lemma 3, deriving recursive relations for its terms, including those representing scalar products. Such terms can be negative and impose significant constraints.

Theorem 1. *Under Assumptions 1, 2(a), after T iterations of Algorithm 1 with $\gamma \leqslant \frac{1}{30Ln^{3/2}}, p = \frac{1}{n}$, the following inequality holds:*

$$\begin{aligned}\mathbb{E}\Big[\left\|z^T - z^*\right\|^2 &+ \gamma^2\left\|G^{T-1} - G^{T-2}\right\|^2 + 2\gamma M\left\langle F(z^{T-1/2}) - G^{T-1}, z^{T-1/2} - z^*\right\rangle \\ &+ \gamma^2 H\left\|F(z^{T-1/2}) - G^{T-1}\right\|^2\Big] \leqslant (1-\gamma\mu)^T\left\|z^0 - z^*\right\|^2,\end{aligned}$$

where $M = \frac{1-p}{p-\gamma\mu}$ and $H = 70n^3$.

Proof. We start with the result provided by Lemma 3:

$$\left\|z^{t+1} - z^*\right\|^2 = (1-\gamma\mu)\left\|z^t - z^*\right\|^2 - (1-2\gamma\mu)\left\|z^{t+1/2} - z^t\right\|^2$$

$$+2\gamma\left\langle F(z^{t+1/2})-G^t, z^{t+1/2}-z^*\right\rangle + 2\gamma^2\left\|G^t-G^{t-1}\right\|^2.$$

Taking the expectation over G^{t-1} and i^t from the both sides of the inequality,

$$\begin{aligned}\mathbb{E}_{i^t}\mathbb{E}_{G^{t-1}}\left\|z^{t+1}-z^*\right\|^2 &= (1-\gamma\mu)\left\|z^t-z^*\right\|^2-(1-2\gamma\mu)\left\|z^{t+1/2}-z^t\right\|^2 \\ &\quad +2\gamma\mathbb{E}_{i^t}\mathbb{E}_{G^{t-1}}\left\langle F(z^{t+1/2})-G^t, z^{t+1/2}-z^*\right\rangle + 2\gamma^2\mathbb{E}_{i^t}\mathbb{E}_{G^{t-1}}\left\|G^t-G^{t-1}\right\|^2.\end{aligned} \tag{16}$$

There let us consider the obtained in (16) terms separately. First,

$$\begin{aligned}\mathbb{E}_{i^t}\mathbb{E}_{G^{t-1}}\left\langle F(z^{t+1/2})-G^t, z^{t+1/2}-z^*\right\rangle &= p\left\langle F(z^{t+1/2})-F(z^{t+1/2}), z^{t+1/2}-z^*\right\rangle \\ &\quad +(1-p)\left\langle F(z^{t+1/2})-G^{t-1}, z^{t+1/2}-z^*\right\rangle \\ &\quad +(1-p)\mathbb{E}_{i^t}\left\langle -F_{i^t}(z^{t+1/2})+F_{i^t}(z^{t-1/2}), z^{t+1/2}-z^*\right\rangle \\ &= (1-p)\left\langle F(z^{t-1/2})-G^{t-1}, z^{t+1/2}-z^*\right\rangle \\ &= (1-p)\left\langle F(z^{t-1/2})-G^{t-1}, z^{t-1/2}-z^*\right\rangle \\ &\quad +(1-p)\left\langle F(z^{t-1/2})-G^{t-1}, z^{t+1/2}-z^{t-1/2}\right\rangle.\end{aligned}$$

Using the (Young)'s inequality, we get

$$\begin{aligned}2\gamma\mathbb{E}_{i^t}\mathbb{E}_{G^{t-1}}\left\langle F(z^{t+1/2})-G^t, z^{t+1/2}-z^*\right\rangle \leqslant& 2\gamma(1-p)\mathbb{E}\left\langle F(z^{t-1/2})-G^{t-1}, z^{t-1/2}-z^*\right\rangle + \frac{1-p}{c}\left\|z^{t+1/2}-z^{t-1/2}\right\|^2 \\ &+ c(1-p)\gamma^2\left\|F(z^{t-1/2})-G^{t-1}\right\|^2,\end{aligned} \tag{17}$$

where c we define later. Now we focuse on the last term of (17). Lemma 1 provides:

$$\mathbb{E}_{i^t}\mathbb{E}_{G^{t-1}}\left\|F(z^{t+1/2})-G^t\right\|^2 \leqslant (1-p)\mathbb{E}_{i^t}\left\|F_{i^t}(z^{t+1/2})-F_{i^t}(z^{t-1/2})\right\|^2 + (1-p)\left\|F(z^{t-1/2})-G^{t-1}\right\|^2.$$

Using Assumption 1,

$$\mathbb{E}_{i^t}\mathbb{E}_{G^{t-1}}\left\|F(z^{t+1/2})-G^t\right\|^2 \leqslant (1-p)L^2\left\|z^{t+1/2}-z^{t-1/2}\right\|^2 + (1-p)\left\|F(z^{t-1/2})-G^{t-1}\right\|^2. \tag{18}$$

Then we reflect on the last term in (16) and, making similar transformations, obtain

$$\begin{aligned}\mathbb{E}_{i^t}\mathbb{E}_{G^{t-1}}\left\|G^t-G^{t-1}\right\|^2 &= (1-p)\mathbb{E}_{i^t}\left\|F_{i^t}(z^{t+1/2})-F_{i^t}(z^{t-1/2})\right\|^2 + p\left\|F(z^{t+1/2})-G^{t-1}\right\|^2 \\ &\overset{\text{Ass. 1}}{\leqslant} (1-p)L^2\left\|z^{t+1/2}-z^{t-1/2}\right\|^2 + p\left\|F(z^{t+1/2})-F(z^{t-1/2})+F(z^{t-1/2})-G^{t-1}\right\|^2 \\ &\overset{\text{(CS),Ass. 1}}{\leqslant} (1-p)L^2\left\|z^{t+1/2}-z^{t-1/2}\right\|^2 + 2pL^2\left\|z^{t+1/2}-z^{t-1/2}\right\|^2 \\ &\quad +2p\mathbb{E}\left\|F(z^{t-1/2})-G^{t-1}\right\|^2 \\ &= (1+p)L^2\left\|z^{t+1/2}-z^{t-1/2}\right\|^2 + 2p\left\|F(z^{t-1/2})-G^{t-1}\right\|^2.\end{aligned} \tag{19}$$

Note that the first term frequently appears in convergence analysis. Let us expand it in detail:

$$\begin{aligned}\left\|z^{t+1/2}-z^{t-1/2}\right\|^2 &\overset{\text{(CS)}}{\leqslant} \left(1+\frac{1}{a}\right)\left\|z^{t+1/2}-z^t\right\|^2 + (1+a)\left\|z^t-z^{t-1/2}\right\|^2 \\ &= \left(1+\frac{1}{a}\right)\left\|z^{t+1/2}-z^t\right\|^2 + (1+a)\,\gamma^2\left\|G^{t-1}-G^{t-2}\right\|^2.\end{aligned}$$

Taking expectations and applying (19),

$$\mathbb{E}_{i^{t-1}}\mathbb{E}_{G^{t-2}}\left\|z^{t+1/2}-z^{t-1/2}\right\|^2 \leqslant \left(1+\frac{1}{a}\right)\mathbb{E}_{i^{t-1}}\mathbb{E}_{G^{t-2}}\left\|z^{t+1/2}-z^t\right\|^2 + 2p\,(1+a)\,\gamma^2\left\|F(z^{t-3/2})-G^{t-2}\right\|^2$$

$$+ (1+p)(1+a) L^2\gamma^2 \left\| z^{t-1/2} - z^{t-3/2} \right\|^2 .$$

We now enter the recursion, considering that

$$\|z^{1/2} - z^{-1/2}\|^2 \leqslant \left(1 + \frac{1}{a}\right) \|z^{1/2} - z^0\|^2 + (1+a)\|z^0 - z^{-1/2}\|^2.$$

Putting $z^{-1} = z^{-1/2} = z^0$, as well as $G^{-1} = F(z^{-1/2})$, we derive the estimate:

$$\begin{aligned}
&\mathbb{E}_{i^0}\mathbb{E}_{G^{-1}} \ldots \mathbb{E}_{i^{t-1}}\mathbb{E}_{G^{t-2}} \left\| z^{t+1/2} - z^{t-1/2} \right\|^2 \\
&\leqslant \left(1 + \frac{1}{a}\right) \sum_{k=0}^{t} \left((1+p)(1+a) L^2\gamma^2\right)^{t-k} \mathbb{E}_{i^k}\mathbb{E}_{G^{k-1}} \ldots \mathbb{E}_{i^{t-1}}\mathbb{E}_{G^{t-2}} \left\| z^{k+1/2} - z^k \right\|^2 \\
&\quad + 2p(1+a)\gamma^2 \sum_{k=0}^{t} \left((1+p)(1+a) L^2\gamma^2\right)^{t-k} \mathbb{E}_{i^k}\mathbb{E}_{G^{k-1}} \ldots \mathbb{E}_{i^{t-1}}\mathbb{E}_{G^{t-2}} \left\| F(z^{k-1/2}) - G^{k-1} \right\|^2 .
\end{aligned}$$

If we choose $\gamma \leqslant \frac{1}{L\sqrt{2(1+p)(1+a)}}$, we get that

$$\begin{aligned}
&\mathbb{E}_{i^0}\mathbb{E}_{G^{-1}} \ldots \mathbb{E}_{i^{t-1}}\mathbb{E}_{G^{t-2}} \left\| z^{t+1/2} - z^{t-1/2} \right\|^2 \\
&\leqslant \left(1 + \frac{1}{a}\right) \sum_{k=0}^{t} \left(\frac{1}{2}\right)^{t-k} \mathbb{E}_{i^k}\mathbb{E}_{G^{k-1}} \ldots \mathbb{E}_{i^{t-1}}\mathbb{E}_{G^{t-2}} \left\| z^{k+1/2} - z^k \right\|^2 \\
&\quad + 2p(1+a)\gamma^2 \sum_{k=0}^{t} \left(\frac{1}{2}\right)^{t-k} \mathbb{E}_{i^k}\mathbb{E}_{G^{k-1}} \ldots \mathbb{E}_{i^{t-1}}\mathbb{E}_{G^{t-2}} \left\| F(z^{k-1/2}) - G^{k-1} \right\|^2 .
\end{aligned} \tag{20}$$

For the sake of clarity, let us redefine the terms important for analysis:

$$\begin{cases}
\delta^t &= \left\| z^t - z^* \right\|^2 , \\
S^t &= \left\langle F(z^{t-1/2}) - G^{t-1}, z^{t-1/2} - z^* \right\rangle , \\
g^t &= \left\| F(z^{t-1/2}) - G^{t-1} \right\|^2 .
\end{cases}$$

We sum (16), (17) and $2\gamma^2\cdot$(19), and take additional expectations. We get

$$\begin{aligned}
&\mathbb{E}_{i^0}\mathbb{E}_{G^{-1}} \ldots \mathbb{E}_{i^t}\mathbb{E}_{G^{t-1}} \delta^{t+1} + 2\gamma^2 \mathbb{E}_{i^0}\mathbb{E}_{G^{-1}} \ldots \mathbb{E}_{i^t}\mathbb{E}_{G^{t-1}} \left\| G^t - G^{t-1} \right\|^2 \\
&\leqslant \mathbb{E}_{i^0}\mathbb{E}_{G^{-1}} \ldots \mathbb{E}_{i^{t-1}}\mathbb{E}_{G^{t-2}} \left[(1 - \gamma\mu)\delta^t - (1 - 2\gamma\mu)) \left\| z^{t+1/2} - z^t \right\|^2 + (1-p) 2\gamma S^t \right. \\
&\quad \left. + \left(\frac{1-p}{c} + 2(1+p)L^2\gamma^2 \right) \left\| z^{t+1/2} - z^{t-1/2} \right\|^2 + (c(1-p) + 4p)\gamma^2 g^t \right] .
\end{aligned} \tag{21}$$

We proceed to analyze the convergence of the sum $V^t = \mathbb{E}\left[\delta^t + 2\gamma M S^t + \gamma^2 H g^t + 2\gamma^2 \left\| G^t - G^{t-1} \right\|^2\right]$. At this stage, (21) $+ M \cdot$ (17) $+ \gamma^2 H \cdot$ (18) reveals:

$$\begin{aligned}
&\mathbb{E}_{i^0}\mathbb{E}_{G^{-1}} \ldots \mathbb{E}_{i^{t-1}}\mathbb{E}_{G^{t-2}} \left[\delta^{t+1} + 2\gamma M S^{t+1} + \gamma^2 H g^{t+1} + 2\gamma^2 \left\| G^{t+1} - G^t \right\|^2 \right] \\
&\leqslant \mathbb{E}_{i^0}\mathbb{E}_{G^{-1}} \ldots \mathbb{E}_{i^{t-1}}\mathbb{E}_{G^{t-2}} \left[(1 - \gamma\mu)\delta^t - (1 - 2\gamma\mu)\mathbb{E}\left\| z^{t+1/2} - z^t \right\|^2 + (1-p)(M+1) 2\gamma S^t \right. \\
&\quad + \left(\frac{(1-p)(M+1)}{c} + 2(1+p)L^2\gamma^2 + (1-p)L^2\gamma^2 H \right) \mathbb{E}\left\| z^{t+1/2} - z^{t-1/2} \right\|^2 \\
&\quad \left. + ((1-p)(H + cM) + c(1-p) + 4p)\gamma^2 g^t \right] .
\end{aligned}$$

Using (20) and taking the full expectation,

$$\begin{aligned} V^{t+1} &\leqslant (1-\gamma\mu)\mathbb{E}\delta^t - (1-2\gamma\mu)\mathbb{E}\left\|z^{t+1/2}-z^t\right\|^2 + (1-p)(M+1)2\gamma\mathbb{E}S^t \\ &+ \left(\frac{(1-p)(M+1)}{c} + 2(1+p)L^2\gamma^2 + (1-p)L^2\gamma^2 H\right) \\ &\cdot\left[\left(1+\frac{1}{a}\right)\sum_{k=0}^{t}\left(\frac{1}{2}\right)^{t-k}\mathbb{E}\left\|z^{k+1/2}-z^k\right\|^2\right. \\ &\left. + 2p\left(1+a\right)\gamma^2\sum_{k=0}^{t}\left(\frac{1}{2}\right)^{t-k}\mathbb{E}g^k\right] \\ &+ \left((1-p)(H+cM)+c(1-p)+4p\right)\gamma^2\mathbb{E}g^t. \end{aligned}$$

Now we sum this over all iterations with positive coefficients q^t:

$$\begin{aligned} \sum_{t=0}^{T-1} q^t V^{t+1} &\leqslant (1-\gamma\mu)\sum_{t=0}^{T-1} q^t\mathbb{E}\delta^t - (1-2\gamma\mu)\sum_{t=0}^{T-1} q^t\mathbb{E}\left\|z^{t+1/2}-z^t\right\|^2 \\ &+ (1-p)(M+1)2\gamma\sum_{t=0}^{T-1} q^t\mathbb{E}S^t \\ &+ \left(\frac{(1-p)(M+1)}{c} + 2(1+p)L^2\gamma^2 + (1-p)L^2\gamma^2 H\right) \\ &\cdot\left[\left(1+\frac{1}{a}\right)\sum_{t=0}^{T-1} q^t\sum_{k=0}^{t}\left(\frac{1}{2}\right)^{t-k}\mathbb{E}\left\|z^{k+1/2}-z^k\right\|^2\right. \\ &\left. + 2p\left(1+a\right)\gamma^2\sum_{t=0}^{T-1} q^t\sum_{k=0}^{t}\left(\frac{1}{2}\right)^{t-k}\mathbb{E}g^k\right] \\ &+ \left((1-p)(H+cM)+c(1-p)+4p\right)\gamma^2\sum_{t=0}^{T-1} q^t\mathbb{E}g^t. \end{aligned} \tag{22}$$

At this stage, our task is to choose the constants q^t, M, H such that the factors of identical terms in (22) cancel out. We commence with choosing $q^t = (1-\gamma\mu)^{-t}$.

- We proceed with taking a look at S^t. Our objective is to reduce terms involving $S^t, t \in \{1,2,...,T-1\}$ by equating their coefficients on the RHS and LHS:

$$\begin{aligned} 2\gamma M q^{t-1} &= (1-p)(M+1)2\gamma q^t, \\ M &= \frac{1-p}{p-\gamma\mu}. \end{aligned} \tag{23}$$

  For clarity, we define

$$D = \frac{(1-p)(M+1)}{c} + 2(1+p)L^2\gamma^2 + (1-p)L^2\gamma^2 H. \tag{24}$$

- Then we consider the coefficient at g^t.

$$LHS : \gamma^2 H\sum_{t=0}^{T-1} q^t g^{t+1}.$$

$$RHS : 2p\left(1+a\right)D\gamma^2\sum_{t=0}^{T-1} q^t\sum_{k=0}^{t}\left(\frac{1}{2}\right)^{t-k}g^k + \left((1-p)(H+cM)+c(1-p)+4p\right)\gamma^2\sum_{t=0}^{T-1} q^t g^t.$$

Equating the coefficients of g^t, yields

$$q^t\gamma^2 H \quad \geqslant \quad q^{t+1}\gamma^2\left((1-p)c+4p+(1-p)(H+cM)\right)+2p(1+a)D\gamma^2\underbrace{\left(\sum_{k=t}^{T-1}\left(\frac{1}{2}\right)^{k-t}q^k\right)}_{q^t\left(\sum_{k=0}^{T-1-t}\left(\frac{1}{2}\right)^k q^k\right)}.$$

After that, we can divide the left and right sides by $q^t\gamma^2$:

$$H \quad \geqslant \quad q\left((1-p)c+4p+(1-p)(H+cM)\right)+2p(1+a)D\cdot\frac{1}{1-\frac{q}{2}}.$$

Choosing $q \leqslant \frac{4}{3}$, we obtain $\frac{1}{1-\frac{q}{2}} \leqslant 3$. In that way, recalling the definition of D (24), we want

$$\begin{aligned} H\left(1-q(1-p)-6p(1+a)(1-p)L^2\gamma^2\right) \quad \geqslant \quad & q\left(4p+(1-p)\left(M+1\right)c\right) \\ & +6p(1+a)\left(\frac{(1-p)(1+M)}{c}+2(1+p)L^2\gamma^2\right). \end{aligned} \tag{25}$$

Which enable us to omit the terms containing $g^t, t \in \{1,2,...,T-1\}$ in inequality (22).

- Let us lastly dissect the coefficient with $\|z^{t+1/2}-z^t\|^2$ to eliminate this terms from (22). Here we use $\left(\sum\limits_{k=t}^{T-1}\left(\frac{1}{2}\right)^{k-t}q^k\right) = q^t\left(\sum\limits_{k=0}^{T-1-t}\left(\frac{1}{2}\right)^k q^k\right) \leqslant 3q^t$ again:

$$D\left(1+\frac{1}{a}\right)\sum_{k=t}^{T-1}q^k\left(\frac{1}{2}\right)^{k-t}-q^t(1-2\gamma\mu)\leqslant 0,$$

$$3D\left(1+\frac{1}{a}\right)\leqslant 1-2\gamma\mu.$$

Substituting D from the definition (24) we obtain:

$$3\left(\frac{(1-p)(M+1)}{c}+2(1+p)L^2\gamma^2+(1-p)L^2\gamma^2 H\right)\left(1+\frac{1}{a}\right) \quad \leqslant \quad 1-2\gamma\mu.$$

Which pose further restriction on H:

$$3H(1-p)L^2\gamma^2\left(1+\frac{1}{a}\right)\leqslant 1-2\gamma\mu-3\left(\frac{(1-p)(M+1)}{c}+2(1+p)L^2\gamma^2\right)\left(1+\frac{1}{a}\right). \tag{26}$$

For the proper H to exist, that satisfy both (26) and (25), the following inequality should hold:

$$\begin{aligned} &\frac{1}{1-q(1-p)-6p(1+a)(1-p)L^2\gamma^2}\left[q\left(4p+(1-p)(M+1)c\right)+6p(1+a)\left(\frac{(1-p)(1+M)}{c}+2(1+p)L^2\gamma^2\right)\right] \\ &\leqslant \frac{1}{3(1-p)\left(1+\frac{1}{a}\right)L^2\gamma^2}\left[1-2\gamma\mu-3\left(\frac{(1-p)(1+M)}{c}+2(1+p)L^2\gamma^2\right)\left(1+\frac{1}{a}\right)\right]. \end{aligned} \tag{27}$$

We evaluate (27) assuming $a=1, \gamma=\frac{1}{bLn^{3/2}}, b\geqslant 1, n\geqslant 4, p=\frac{1}{n}$. First of all, we recall (23) and derive a useful upper bound for $M+1$:

$$M+1=\frac{1-\gamma\mu}{p-\gamma\mu}=\frac{1}{p-p\frac{\mu}{Lb\sqrt{n}}}\leqslant\frac{2}{p}=2n. \tag{28}$$

Now our strategy is to obtain the upper bound for the LHS and lower for the RHS of (27). Let us start with the RHS:

$$RHS\geqslant\frac{b^2n^3}{6}\left[1-\frac{2\mu}{bn^{3/2}L}-6\left(\frac{2n}{c}+\frac{4}{b^2n^3}\right)\right].$$

Here we encountered with the necessity to set $c > 12n$. For further analysis we will use $c = 24n$:

$$RHS \geqslant \frac{b^2n^3}{6}\left[\frac{1}{2} - \frac{2}{bn^{3/2}} - \frac{24}{b^2n^3}\right] \geqslant \frac{b^2n^3}{6}\left[\frac{1}{2} - \frac{1}{2b} - \frac{3}{8b^2}\right]. \tag{29}$$

Then, we start evaluation of the LHS, using our choice of q, γ, c, a and derived upper bound for $M + 1$:

$$\begin{aligned} LHS &\leqslant \frac{1-\gamma\mu}{(1-\gamma\mu)-(1-p)-(1-\gamma\mu)2p(1-p)\frac{1}{b^2n^3}} \\ &\quad \cdot \left[\frac{4}{3}\left(4p + (1-p)c\frac{2}{p}\right) + 12p\left(\frac{(1-p)}{c}\frac{2}{p} + 2(1+p)\frac{1}{b^2n^3})\right)\right] \qquad (30) \\ &\leqslant \frac{1}{p\left(1 - \frac{1}{b\sqrt{n}} - \frac{12}{b^2n^2}\right)}\left[\frac{4}{3}\left(4p + 48\frac{1}{p^2}\right) + p + \frac{48p^4}{b^2}\right] \\ &\leqslant \frac{n^3}{1 - \frac{1}{2b} - \frac{3}{4b^2}}\left[64 + \frac{1}{12} + \frac{1}{64} + \frac{3}{256b^2}\right]. \qquad (31) \end{aligned}$$

Combining (29) and (30) we reach a sufficient condition for the existence of a solution to inequality (27):

$$\frac{1}{1 - \frac{1}{2b} - \frac{3}{4b^2}}\left[64 + \frac{1}{12} + \frac{1}{64} + \frac{3}{256b^2}\right] \leqslant \frac{b^2}{6}\left[\frac{1}{2} - \frac{1}{2b} - \frac{3}{8b^2}\right].$$

Straight forward evaluation shows $b \geqslant 28.6$ is sufficient. Which means that under assumption $b \geqslant 30$ lower bound of H less then upper one. To select the appropriate value for H with a compact notation, we observe that $b = 30$ implies:

$$\begin{aligned} RHS &\geqslant \frac{n^3b^2}{6}\left[\frac{1}{2} - \frac{1}{2b} - \frac{3}{8b^2}\right] = \frac{1159}{16}n^3 \geqslant 72n^3, \\ LHS &\leqslant \frac{n^3}{1 - \frac{1}{2b} - \frac{3}{4b^2}}\left[64 + \frac{1}{12} + \frac{1}{64} + \frac{3}{256b^2}\right] = \frac{4922801}{75456}n^3 \leqslant 66n^3. \end{aligned}$$

And an increase in b expands these boundaries. Therefore for all $\gamma \leqslant \frac{1}{30Ln^{3/2}}$ we can set $H = 70n^3$. Finally, we claim that $\gamma \leqslant \frac{1}{30Ln^{3/2}}$ satisfies all the previous constrains. After these preparations we can derive the necessary result from (22):

$$\begin{aligned} &\mathbb{E}\left[\left\|z^T - z^*\right\|^2 + \gamma^2\left\|G^{T-1} - G^{T-2}\right\|^2 + 2\gamma M\left\langle F(z^{T-1/2}) - G^{T-1}, z^{T-1/2} - z^*\right\rangle\right. \\ &\quad \left.+\gamma^2 H\left\|F(z^{T-1/2}) - G^{T-1}\right\|^2\right] \\ &\leqslant \left\|z^0 - z^*\right\|^2 + 2\gamma M\left\langle F(z^{-1/2}) - G^{-1}, z^{-1/2} - z^*\right\rangle + \gamma^2 H\left\|F(z^{-1/2}) - G^{-1}\right\|^2 = \left\|z^0 - z^*\right\|^2. \end{aligned}$$

This finishes the proof of the theorem. □

Corollary 1 highlights the advantage of the obtained guarantees based on the function V^t over the conventional criterion $\|z^t - z^*\|^2$. This superiority is not immediately apparent due to the potential negativity of the scalar product $\langle F(z^{t-1/2}) - G^{t-1}, z^{t-1/2} - z^*\rangle$.

Corollary 1. *In settings of Theorem 1, after T iterations of Algorithm 1 with $\gamma \leqslant \frac{1}{30Ln^{3/2}}$ and $p = \frac{1}{n}$, the following inequality holds:*

$$\mathbb{E}\left[\frac{1}{2}\left\|z^T - z^*\right\|^2 + \frac{\gamma^2 H}{2}\left\|F(z^{T-1/2}) - G^{T-1}\right\|^2\right] \leqslant (1-\gamma\mu)^T\left\|z^0 - z^*\right\|^2.$$

Proof. First of all, due to the (Young) inequality, we get

$$2\gamma M\left\langle F(z^{T-1/2}) - G^{T-1}, z^{T-1/2} - z^*\right\rangle \geqslant -\frac{\gamma^2 H}{2}\left\|F(z^{T-1/2}) - G^{T-1}\right\|^2 - \frac{2M^2}{H}\left\|z^{T-1/2} - z^*\right\|^2. \tag{32}$$

Then, we apply (CS) to derive

$$-\left\|z^{T-1/2}-z^*\right\|^2 \geqslant -2\left\|z^{T-1/2}-z^T\right\|^2 - 2\left\|z^T-z^*\right\|^2. \tag{33}$$

Recalling estimate (28) we conclude $M \leqslant 2n$. Combining it with (32), (33) and chosen $H = 70n^3$ we obtain

$$\begin{aligned} 2\gamma M\left\langle F(z^{T-1/2})-G^{T-1}, z^{T-1/2}-z^*\right\rangle &\geqslant -\frac{\gamma^2 H}{2}\left\|F(z^{T-1/2})-G^{T-1}\right\|^2 - \frac{4}{35n}\left\|z^T-z^*\right\|^2 \\ &\quad -\frac{4}{35n}\left\|z^T-z^{T-1/2}\right\|^2. \end{aligned} \tag{34}$$

After that, we examine $\left\|G^{T-1}-G^{T-2}\right\|^2$ using Lines 4 and 6 of Algorithm 1:

$$\gamma^2\left\|G^{T-1}-G^{T-2}\right\|^2 = \left\|(z^{T-1}-z^T)-(z^{T-1}-z^{T-1/2})\right\|^2 = \left\|z^T-z^{T-1/2}\right\|^2. \tag{35}$$

Finally, we plug (35) and (34) into the Lyapunov function V^t in the result of Theorem 1:

$$\begin{aligned} (1-\gamma\mu)^T\left\|z^0-z^*\right\|^2 &\geqslant \mathbb{E}\Big[\left\|z^T-z^*\right\|^2 + \gamma^2\left\|G^{T-1}-G^{T-2}\right\|^2 + 2\gamma M\left\langle F(z^{T-1/2})-G^{T-1}, z^{T-1/2}-z^*\right\rangle \\ &\quad +\gamma^2 H\left\|F(z^{T-1/2})-G^{T-1}\right\|^2\Big] \\ &\geqslant \mathbb{E}\left[\left(1-\frac{4}{35n}\right)\left(\left\|z^T-z^*\right\|^2+\gamma^2\left\|G^{T-1}-G^{T-2}\right\|^2\right) + \frac{\gamma^2 H}{2}\left\|F(z^{T-1/2})-G^{T-1}\right\|^2\right]. \end{aligned}$$

This finishes the proof of the corollary. □

Now we are ready to present the final convergence estimate for our method.

Corollary 2. *Suppose Assumptions 1, 2(a) hold. Then Algorithm 1 with $\gamma = \frac{1}{30Ln^{3/2}}$ and $p = \frac{1}{n}$, to reach ε-accuracy, where $\varepsilon \sim V^T$, needs*

$$\widetilde{\mathcal{O}}\left(\frac{Ln^{3/2}}{\mu}\log\frac{1}{\varepsilon}\right) \quad \textit{iterations and oracle calls.}$$

Proof. Theorem 1 guarantees that Algorithm 1 converges to ε-accuracy within $\widetilde{\mathcal{O}}\left(\frac{\log\frac{1}{\varepsilon}}{\gamma\mu}\right)$ iterations. Setting γ with the upper bound $\frac{1}{30Ln^{3/2}}$ we reach $\widetilde{\mathcal{O}}\left(\frac{Ln^{3/2}}{\mu}\log\frac{1}{\varepsilon}\right)$ iteration complexity. Then, we note that average iteration cost is $2(1-p)+pn = 3-2p$, which implies the same bound $\widetilde{\mathcal{O}}\left(\frac{Ln^{3/2}}{\mu}\log\frac{1}{\varepsilon}\right)$ for the oracle complexity. □