Shuffling Heuristic in Variational Inequalities: Establishing New Convergence Guarantees

Daniil Medyakov
MIPT[*], ISP RAS[†]
mediakov.do@phystech.edu

Gleb Molodtsov
MIPT, ISP RAS
molodtsov.gl@phystech.edu

Grigoriy Evseev
MIPT
evseev.gv@phystech.edu

Egor Petrov
MIPT
petrov.egor.d@phystech.edu

Aleksandr Beznosikov
MIPT, ISP RAS, Innopolis University
anbeznosikov@gmail.com

Abstract

Variational inequalities have gained significant attention in machine learning and optimization research. While stochastic methods for solving these problems typically assume independent data sampling, we investigate an alternative approach - the shuffling heuristic which involves permuting the dataset before sequential processing and ensures equal consideration of all data points. Despite its practical utility, theoretical guarantees for shuffling in variational inequalities remain unexplored. We address this gap by providing the first theoretical convergence estimates for shuffling methods in this context. Our analysis establishes rigorous bounds and convergence rates, extending the theoretical framework for this important class of algorithms. We validate our findings through extensive experiments on diverse benchmark variational inequality problems, demonstrating faster convergence of shuffling methods compared to independent sampling approaches.

1 Introduction

Variational inequalities (VIs) have been attracting researchers' attention in various fields for more than half a century (Browder, 1965). In this work, we investigate the variational inequality problem in the following form:

$$\text{find} \;\; z^* \in Z \;\; \text{such that} \;\; \forall z \in Z \hookrightarrow \langle F(z^*), z - z^* \rangle + g(z) - g(z^*) \geqslant 0, \tag{1}$$

where F is a monotone operator and g is a proper convex lower semicontinuous function, which plays the role of regularizer. Variational inequalities serve as a universal tool for addressing particular problems, such as minimization, saddle point problems, fixed point problems, and others (Facchinei and Pang, 2003; Kinderlehrer and Stampacchia, 2000). We give some examples to provide intuition about VIs.

Example 1 (Convex optimization)**.** *We consider the following convex regularized optimization problem:*

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

In this example, f is a smooth data representative term, and g is probably a non-smooth regularizer. In this setting, we define $F(z) = \nabla f(z)$. Then $z^* \in dom\ g$ is the solution of (1) if and only if $z^* \in dom\ g$ is the solution of (2). In this way, the problem (2) can be considered as a variational inequality.

[*]Moscow Institute of Physics and Technology
[†]Institute for System Programming RAS

Example 2 (Convex-concave saddles)**.** *We consider the following convex-concave saddle point problem:*

$$\min_{x\in\mathbb{R}^{d_x}} \max_{y\in\mathbb{R}^{d_y}} \left[f(x,y) + g_1(x) - g_2(y)\right]. \tag{3}$$

There, f has the same interpretation as in Example 1, and g_1, g_2 can also be perceived as regularizers. In this setting, we define $F(z) = F(x,y) = [\nabla_x f(x,y), -\nabla_y f(x,y)]$. Then $z^* \in \operatorname{dom} g_1 \times \operatorname{dom} g_2$ is the solution of (1) if and only if $z^* \in \operatorname{dom} g_1 \times \operatorname{dom} g_2$ is the solution of (3). In this way, the problem (3) can be considered as a variational inequality.

There are multiple practical reasons to focus on this formulation. Firstly, for numerous non-smooth problems, solutions are often more efficiently obtained when the former are formulated as saddle point problems (Nesterov, 2005; Nemirovski, 2004; Chambolle and Pock, 2011; Esser et al., 2010). Secondly, recent studies have found new links between VIs and reinforcement learning (Omidshafiei et al., 2017; Jin and Sidford, 2020), adversarial training (Madry et al., 2017), and GANs (Goodfellow et al., 2014). In particular, consideration of monotone and strongly monotone inequalities provides useful methods and recommendations for the GAN community (Daskalakis et al., 2017; Gidel et al., 2018; Mertikopoulos et al., 2018; Chavdarova et al., 2019; Liang and Stokes, 2019; Peng et al., 2020). VIs also have extensive applications in various classical problems, including discriminative 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), economics, game theory (Von Neumann and Morgenstern, 1953), and optimal control (Facchinei and Pang, 2003).

Solving the problem (1) requires specialized methods, as traditional optimization techniques, e.g. the gradient method, often fall short when applied to VIs and saddle point problems (Harker and Pang, 1990). These classical methods not only struggle with efficiency but also offer weak theoretical convergence guarantees in the VI context (Beznosikov et al., 2023). Among the various approaches developed for VIs, the Extragradient method (Korpelevich, 1976; Mokhtari et al., 2020) stands out as one of the most fundamental and effective techniques.

While variational inequalities provide a powerful framework for addressing a wide range of problems, recent trends in machine learning and data science present new challenges. The exponential growth in dataset sizes and increasing complexity of models have created a pressing need for more efficient computational approaches (Bottou, 2010; Dean et al., 2012; Medyakov et al., 2023). To address these challenges in the context of VIs, we reformulate the problem by considering the operator F as the finite sum of operators F_i:

$$F(z) = \frac{1}{n}\sum\nolimits_{i=1}^{n} F_i(z), \tag{4}$$

where each F_i corresponds to an individual data point. This decomposition allows us to tackle large-scale problems more effectively.

In this paper, we explore stochastic algorithms which are particularly suitable for practical extensive applications. As mentioned before, in such cases, the number of operators n is typically large, making the computation of the full operator value at each iteration computationally expensive. Instead, stochastic algorithms randomly select F_i at each iteration. The stochastic version of the Extragradient method (Juditsky et al., 2011) select random independent indexes i_t, j_t at iteration t and performs the following updates:

$$\begin{aligned} z^{t+\frac{1}{2}} &= z^t - \gamma F_{i_t}(z^t), \\ z^{t+1} &= z^t - \gamma F_{j_t}(z^{t+\frac{1}{2}}). \end{aligned} \tag{5}$$

Just as deterministic Extragradient is the modification of the classical gradient method with an additional step, similarly the stochastic Extragradient is the same modification of SGD (Robbins and Monro, 1951). Although this method performs well on the variational inequalities, it encounters a significant issue with its properties and performance thoroughly studied: the variance of its inherent stochastic estimators of operators remains high throughout the learning process. Hence, Extragradient with a constant learning rate converges linearly only to a neighborhood of the optimal solution, the size of which is proportional

to the step size and variance (Juditsky et al., 2011). This problem is also characteristic of classical SGD (Bottou, 2009; Moulines and Bach, 2011; Gower et al., 2020).

To address this limitation, the variance reduction (VR) technique was developed for a classical finite-sum minimization task (Johnson and Zhang, 2013). The method involves the following steps: at the t-th iteration, an index i_t is selected along with a reference point ω^t, which is updated once per epoch or selected probabilistically (as in loopless versions, e.g., (Kovalev et al., 2020)). Considering convex optimization problem (see Example 1), we can formally write the stochastic reduced gradient at the point $z^{t+\frac{1}{2}}$ as

$$\nabla \hat{f}_{i_t}(z^{t+\frac{1}{2}}) = \nabla f_{i_t}(z^{t+\frac{1}{2}}) - \nabla f_{i_t}(\omega^t) + \nabla f(\omega^t). \tag{6}$$

The objective of the variance reduction mechanisms is to overcome the limitations of naive gradient estimators. The former employ an iterative process to construct and apply a gradient estimator with progressively reduced variance. This approach allows for the safe use of larger learning rates, thereby accelerating the training process.

Besides, along with aforementioned SVRG, some of the most popular methods for solving the classical finite-sum problem based on this technique are SAG (Roux et al., 2012), SAGA (Defazio et al., 2014a), Finito (Defazio et al., 2014b), SARAH (Nguyen et al., 2017; Hu et al., 2019), and SPIDER (Fang et al., 2018). The technique of variance reduction is used not only in methods that solve the minimization problem. It is also applicable in the methods for the problem (1). Examples are variance reduced versions of Extragradient, Mirror-prox (Alacaoglu and Malitsky, 2022), gradient method (Palaniappan and Bach, 2016), and forward-reflected-backward (FoRB) (Alacaoglu et al., 2021).

In addition to stochastic methods, various heuristics exist for selecting the i_t-th index at each iteration of algorithms. Thoroughly examining these heuristics could lead to the development of more stable and efficient algorithms in the future. In this paper, we explore the shuffling heuristic (Mishchenko et al., 2020a; Safran and Shamir, 2020; Koloskova et al., 2024; Malinovsky et al., 2023). Unlike the random and independent selection of the index i_t at each iteration, which is common in classical stochastic methods, this heuristic adopts a more practical approach. Specifically, it involves permuting the sequence of indexes $\{1, \ldots, n\}$, where n is the number of data samples (4), and then selecting the index corresponding to the iteration number during the algorithm's execution. It guarantees us that at one epoch of training we make a step for each operator, and once. This property seems important and finds its application in many practical tasks (Chambolle and Pock, 2011; Xu et al., 2004; Bach et al., 2008). There are several shuffling techniques available. Among the most popular are Random Reshuffling (RR) (Gürbüzbalaban et al., 2021; Haochen and Sra, 2019; Nagaraj et al., 2019), where data is shuffled before each epoch; Shuffle Once (SO) (Safran and Shamir, 2020; Rajput et al., 2020), where shuffling occurs once before the start of training; and Cyclic permutation (Mangasarian and Solodov, 1993; Bertsekas and Tsitsiklis, 2000; Nedic and Bertsekas, 2001; Li et al., 2019), where data is accessed deterministically in a cyclic order.

Related works. There are many methods available to solve the problem of variational inequalities. As we mentioned above, the standard deterministic choice for solving the problem (1) is Extragradient (Korpelevich, 1976). This method deals with variational inequalities in the Euclidean setup. Later, Mirror-prox (Nemirovski, 2004), which exploits the Bregman divergence, was proposed. This approach allowed to take into account generalized geometry that could be non-Euclidean. Besides, there are a set of deterministic methods for solving VIs: forward-backward-forward (FBF) (Tseng, 2000), Dual extrapolation (Nesterov, 2007), reflected gradient (Malitsky, 2015), forward-reflected-backward (FoRB) (Malitsky and Tam, 2020).

For the first time, the stochastic version of algorithms for solving VIs was proposed in the work (Juditsky et al., 2011). Later, to reduce the variance inherent in these stochastic methods, researchers turned to the variance reduction technique. The first works in this field are (Palaniappan and Bach, 2016; Chavdarova et al., 2019). In particular, the stochastic gradient method with variance reduction was studied in (Palaniappan and Bach, 2016). The method was based on SVRG (Johnson and Zhang, 2013) and added the Catalyst envelope acceleration. The combination of Extragradient and SVRG was considered in (Chavdarova et al., 2019), which also utilizes the VR technique and achieves better

Table 1: Comparison of the convergence results for the methods for solving VI.

Algorithm	**Sampling**	**VR?**	**Strongly Monotone Complexity**	**Monotone Complexity**
Extragradient (Korpelevich, 1976; Mokhtari et al., 2020)	Deterministic	✗	$\widetilde{\mathcal{O}}\left(\frac{nL}{\mu}\right)$	$\mathcal{O}\left(\frac{nL}{\varepsilon}\right)$
Mirror-prox (Nemirovski, 2004)	Deterministic	✗	\	$\mathcal{O}\left(\frac{nL}{\varepsilon}\right)$
FBF (Tseng, 2000)	Deterministic	✗	\	$\mathcal{O}\left(\frac{nL}{\varepsilon}\right)$
FoRB (Malitsky and Tam, 2020)	Deterministic	✗	\	$\mathcal{O}\left(\frac{nL}{\varepsilon}\right)$
Mirror-prox (Juditsky et al., 2011)	Independent	✗	\	$\mathcal{O}\left(\frac{L}{\varepsilon}+\frac{1}{\varepsilon^2}\right)$
Extragradient (Beznosikov et al., 2020)	Independent	✗	$\widetilde{\mathcal{O}}\left(\frac{L}{\mu}+\frac{1}{\mu^2\varepsilon}\right)$	$\mathcal{O}\left(\frac{L}{\varepsilon}+\frac{1}{\varepsilon^2}\right)$
REG (Mishchenko et al., 2020b)	Independent	✗	$\widetilde{\mathcal{O}}\left(\frac{L}{\mu}+\frac{1}{\mu^2\varepsilon}\right)$	$\mathcal{O}\left(\frac{L}{\varepsilon}+\frac{1}{\varepsilon^2}\right)$
Extragradient (Carmon et al., 2019)	Independent	✓	\	$\widetilde{\mathcal{O}}\left(n+\frac{\sqrt{n}L}{\varepsilon}\right)$
Mirror-prox (Carmon et al., 2019)	Independent	✓	\	$\widetilde{\mathcal{O}}\left(n+\frac{\sqrt{n}L}{\varepsilon}\right)$
FBF (Palaniappan and Bach, 2016)	Independent	✓	$\widetilde{\mathcal{O}}\left(n+\frac{\sqrt{n}L}{\mu}\right)$	$\widetilde{\mathcal{O}}\left(n+\frac{\sqrt{n}L}{\varepsilon}\right)^{(1)}$
Extragradient (Chavdarova et al., 2019)	Independent	✓	$\widetilde{\mathcal{O}}\left(n+\frac{\overline{L}^2}{\mu^2}\right)$	$\widetilde{\mathcal{O}}\left(n+\frac{\overline{L}^2}{\varepsilon^2}\right)^{(1)}$
FoRB (Alacaoglu et al., 2021)	Independent	✓	\	$\mathcal{O}\left(n+\frac{n\overline{L}}{\varepsilon}\right)$
Extragradient (Alacaoglu and Malitsky, 2022)	Independent	✓	$\widetilde{\mathcal{O}}\left(n+\frac{\sqrt{n}L}{\mu}\right)$	$\mathcal{O}\left(n+\frac{\sqrt{n}L}{\varepsilon}\right)$
Mirror-prox (Alacaoglu and Malitsky, 2022)	Independent	✓	\	$\mathcal{O}\left(n+\frac{\sqrt{n}L}{\varepsilon}\right)$
Extragradient (this paper)	RR / SO	✗	$\widetilde{\mathcal{O}}\left(n+\frac{L}{\mu}+\frac{n^2}{\mu^2\varepsilon}\right)$	$\widetilde{\mathcal{O}}\left(n+\frac{L}{\varepsilon}+\frac{n^2}{\varepsilon^3}\right)^{(1)}$
Extragradient (this paper)	RR / SO	✓	$\widetilde{\mathcal{O}}\left(n\frac{L^2}{\mu^2}\right)$	$\widetilde{\mathcal{O}}\left(n\frac{L^2}{\varepsilon^2}\right)^{(1)}$

Columns: Sampling = Deterministic, if considered non-stochastic method, Independent, if method uses independent choice of operator's indexes, RR / SO if method uses shuffling heuristic, Assumption = assumption on operator F, VR? = whether the method uses variance reduction technique.
Notation: μ = constant of strong monotonicity, L = Lipschitz constant of F, $\overline{L}$ = Lipschitz in mean constant, i.e. $1/n\sum_{i=1}^{n}\|F_i(z_1)-F_i(z_2)\| \leqslant L\|z_1-z_2\|\ \forall z_1,z_2\in Z$, n = size of the dataset, ε = accuracy of the solution.
(1): This result is obtained with regularization trick: $\mu \sim \varepsilon/D^2$.

convergence rate. However, only strongly monotone VIs were considered in these works. Consequently, a notable paper in which the authors considered monotone operators was presented (Carmon et al., 2019). This work also falls under the Bregman setup but requires additional assumptions on the operator F and considers the matrix games setup. The current state-of-the-art in this area is the article (Alacaoglu and Malitsky, 2022), which improves the convergence estimates of previous studies. This work addresses various scenarios, including generally monotone and strongly monotone operators, as well as the Bregman and Euclidean setups. Convergence results from the papers highlighted above are summarized in Table 1.

In all these papers, the estimates were obtained in the formulation with an independent choice of the indexes of the operator at each step of the algorithm. As for the shuffling heuristic, there are many papers that explore methods suitable for solving classical finite-sum minimization problems. In the work (Mishchenko et al., 2020a), the authors proposed a classical SGD algorithm, and, by introducing a new notion of variance specific to RR/SO, they were able to match the lower bounds in such cases. In the work (Malinovsky et al., 2023), the SVRG method with RR was considered. The authors actively used results of the work (Mishchenko et al., 2020a) and obtain better rates. Besides, there are a set of works, that considered methods uses the VR technique in the shuffling setup (Huang et al., 2021; Mokhtari et al., 2018; Ying et al., 2020). However, so far there are no papers where the shuffling setting would be used to solve variational inequalities. We are filling this gap.

Contributions. Our main results can be summarized as follows.
• *Novel approach to proof.* Since shuffling methods do not have the property of unbiasedness of stochastic operators, it is necessary to propose new approaches to prove convergence. In this paper, we present a technique that allows us to "return" to the starting point of an epoch in which there is a property of unbiasedness.
• *Convergence estimates.* We provide the first theoretical convergence rates for shuffling methods applied to the finite-sum variational inequality problem. We consider two algorithms: Extragradient and Extragradient with variance reduction. Our comprehensive analysis establishes upper bounds on convergence rates, extending the theoretical framework to encompass this important class of algorithms. In the case of Extragradient, our estimate

on the linear term coincides with that for the method without shuffling, and in the case of Extragradient with VR, we are the first to obtain a linear convergence estimate for methods with shuffling in the VI problem.

• *Experiments.* We conduct comprehensive experiments, which emphasize the superiority of shuffling over the random index selection heuristic. We consider two classical practical applications: image denoising and adversarial training.

2 Setup

Assumptions. Now we present a list of assumptions within which we obtain the main statements.

Assumption 1. *Each operator F_i is L-Lipschitz, i.e., it satisfies $\|F_i(z_1) - F_i(z_2)\| \leq L\|z_1 - z_2\|$ for any $z_1, z_2 \in Z$.*

Assumption 2. *Each operator F_i is μ-strongly monotone, i.e., it satisfies $\langle F_i(z_1) - F_i(z_2), z_1 - z_2\rangle \geqslant \mu\|z_1 - z_2\|^2$ for any $z_1, z_2 \in Z$.*

Assumption 3. *Each stochastic operator F_i and full operator F is bounded at the point of the solution $z^* \in$ dom g, i.e. $\mathbb{E}\|F_i(z^*)\|^2 \leqslant \sigma_*^2, \|F(z^*)\|^2 \leqslant \sigma_*^2$.*

Proximal Algorithm. Earlier, we gave examples of the application of variational inequalities (Examples 1, 2). In many optimization problems, particularly in machine learning and signal processing, we often encounter the need to minimize the function of the same form, i.e. decomposed it into two parts: a smooth differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ and a possibly non-smooth function $g : \mathbb{R}^n \to \mathbb{R}$. To solve this problem, we can utilize the proximal gradient method. The core idea is to iteratively update the solution by combining gradient descent on the smooth part f and the proximal operator for the probably non-smooth part g. We also assume that g is proximal friendly, i.e. the computation of the proximal operator is done for free or costs very little. The proximal operator of the function g at a point x is defined as:

$$\text{prox}_g(z) = \arg\min_{y\in\mathbb{R}^n} \left\{ g(y) + \frac{1}{2}\|y - z\|^2 \right\},$$

where $\|\cdot\|$ denotes the Euclidean norm. Using the proximal operator, the update step for solving the optimization problem can be written as:

$$z^{t+1} = \text{prox}_{\alpha_t g}\left(z^t - \alpha_t \nabla f(z^t)\right).$$

For us, the proximal operator plays a role, since (1) also uses a regularizer.

3 Algorithms and convergence analysis

3.1 Extragradient

The setting of shuffling lies in the fact that we do not choose stochastic operator independently at each step of the method. Instead, we permute the sequence of indexes and, at each iteration, of the algorithm we choose operator according to the new sequence. In this work, we pay attention to the Random Reshuffling and Shuffle Once techniques and provide appropriate Extragradient methods (Algorithms 1, 2).

The analysis of shuffling methods has some specific details. The key difference between shuffling and independent choice is that shuffling methods do not have one essential feature – unbiasedness of stochastic operators:

$$\mathbb{E}_{\pi_s^t}\left[F_{\pi_s^t}(z_s^t)\right] \neq \frac{1}{n}\sum\nolimits_{i=1}^{n} F_{\pi_s^i}(z_s^t) = F(z_s^t).$$

This restriction leads us to a more complex analysis and non-standard techniques to prove convergence of the shuffling methods. Nevertheless, at two points – z_s^0 and z^*, this equality is true. Indeed, the point z_s^0 is the first point of the epoch and there we choose one random index out of n, and the point z^* does not depend on t. Thus, we can "go back" at the beginning of the epoch and take advantage of the unbiased operators. This technique is

Algorithm 1 RR Extragradient

Input: Starting point $z_0^0 \in \mathbb{R}^d$
Parameter: Stepsize γ
for $s = 0, 1, 2, \ldots, S-1$ **do**
  Generate a permutation $\pi_0, \pi_1, \ldots, \pi_{n-1}$ of sequence $\{1, 2, \ldots, n\}$
  for $t = 0, 1, 2, \ldots, n-1$ **do**
    $z_s^{t+\frac{1}{2}} = \text{prox}_{\gamma g}\left(z_s^t - \gamma F_{\pi_s^t}(z_s^t)\right)$
    $z_s^{t+1} = \text{prox}_{\gamma g}\left(z_s^t - \gamma F_{\pi_s^t}(z_s^{t+\frac{1}{2}})\right)$
  end for
  $z_s^n = z_{s+1}^0$
end for
Output: z_S^n

Algorithm 2 SO Extragradient

Input: Starting point $z_0^0 \in \mathbb{R}^d$
Parameter: Stepsize γ
Generate a permutation $\pi_0, \pi_1, \ldots, \pi_{n-1}$ of sequence $\{1, 2, \ldots, n\}$
for $s = 0, 1, 2, \ldots, S-1$ **do**
  for $t = 0, 1, 2, \ldots, n-1$ **do**
    $z_s^{t+\frac{1}{2}} = \text{prox}_{\gamma g}\left(z_s^t - \gamma F_{\pi_s^t}(z_s^t)\right)$
    $z_s^{t+1} = \text{prox}_{\gamma g}\left(z_s^t - \gamma F_{\pi_s^t}(z_s^{t+\frac{1}{2}})\right)$
  end for
  $z_s^n = z_{s+1}^0$
end for
Output: z_S^n

interesting not only in relation to shuffling methods. For example, it is applicable to methods that use Markov chains to select indexes, because there is also no unbiased property anywhere except at the correlation point of the chain. This is the key point of our analysis, and now, having shown it, we present the main result of this section.

Theorem 1. *Suppose Assumptions 1, 2 hold. Then for Algorithms 1, 2 with* $\gamma \leqslant \min\left\{\frac{1}{2\mu n}, \frac{1}{6L}\right\}$ *after* S *epochs,*

$$\|z_S^n - z^*\|^2 \leqslant (1 - \frac{\gamma\mu}{2})^{Sn}\|z_0^0 - z^*\|^2 + \frac{256\gamma n^2\sigma_*^2}{\mu}.$$

Corollary 1. *Suppose Assumptions 1, 2 hold. Then Algorithms 1, 2 with* $\gamma \leqslant \min\left\{\frac{1}{2\mu n}, \frac{1}{6L}, \frac{2\log\left(\max\left\{2, \frac{\mu^2\|z_0^0 - z^*\|^2 T}{512 n^2 \sigma_*^2}\right\}\right)}{\mu T}\right\}$*, to reach* ε*-accuracy, where* $\varepsilon \sim \|z_S^n - z^*\|^2$*, needs*

$$\widetilde{\mathcal{O}}\left(\left(n + \frac{L}{\mu}\right)\log\left(\frac{1}{\varepsilon}\right) + \frac{n^2\sigma_*^2}{\mu^2\varepsilon}\right) \quad \textit{iterations and oracle calls.}$$

Remark 1. *We can transform the obtained estimation for the case of monotone stochastic operators. To do this, we use a regularization trick with* $\mu \sim \frac{\varepsilon}{D}$*. Thus, solving the problem with the operator* $\hat{F}(z) = F(z) + \mu(z - z_0^0)$ *with the accuracy* $\frac{\varepsilon}{2}$*, we solve the problem (1) with the accuracy* ε *and obtain* $\widetilde{\mathcal{O}}\left(n + \frac{L}{\varepsilon} + \frac{n^2}{\varepsilon^3}\right)$ *iteration and oracle complexity. This is convergence in argument, it differs from the classical form.*

Let us explain the result of the theorem. The form of the estimate is classical and appears in all stochastic methods for strongly convex minimization (Moulines and Bach, 2011; Stich, 2019) and strongly monotone VIs (Beznosikov et al., 2020; Mishchenko et al., 2020b). Let us compare it with the results of related works. Our method is based on REG (Mishchenko et al., 2020b). In this work, authors obtain $\widetilde{\mathcal{O}}\left(\frac{L}{\mu} + \frac{1}{\mu^2\varepsilon}\right)$ oracle complexity. Therefore, our result is a great achievement in the shuffling theory, since despite the deterioration on n in the sublinear term, the estimation on the linear term coincides with that in the classical setting with independent choice of stochastic operators. Let us also compare the result with the work (Juditsky et al., 2011). The authors obtain $\mathcal{O}\left(\frac{L}{\varepsilon} + \frac{1}{\varepsilon^2}\right)$. However, uniform bounds on the variance were required in this work, and we bound the variance only at the optimum. Note that, according to current theory, shuffling methods are no more effective than methods with independent sampling for the classical minimization problem (Mishchenko et al., 2020a; Koloskova et al., 2024).

Let us pay attention to the second term in the estimation. In general, the sublinear term with σ_*^2 is not improved. However, for the finite-sum problem, this term can be eliminated by using additional techniques, such as variance reduction.

3.2 Extragradient with variance reduction

Now we use the variance reduction technique, which improves the convergence of the algorithms by reducing the influence of random fluctuations. This approach was not used in the previous algorithms presented. We introduce a version of the RR/SO Extragradient with variance reduction algorithm (Algorithm 3) and the convergence results for this method. Note the peculiarity in Line 11 of this algorithm. In the work (Malinovsky et al., 2023), where shuffling is investigated in variance reduction methods, the authors use a more classical version and compute $F(\omega_s^t)$ at the beginning of each epoch. We consider another option and compute this full operator randomly with probability p. We put $p = \frac{1}{n}$ not to increase the oracle complexity and obtain that on average we also update the full operator once per epoch.

Theorem 2. *Suppose that Assumptions 1, 2 hold. Then for Algorithm 3 with $\gamma \leqslant \frac{(1-\alpha)\mu}{6L^2}, p = \frac{1}{n}$ and $V_s^t = \mathbb{E}\|z_s^t - z^*\|^2 + \mathbb{E}\|\omega_s^t - z^*\|^2$ after T iterations,*

$$V_S^n \leqslant \left(1 - \frac{\gamma\mu}{4}\right)^T V_0^0.$$

Corollary 2. *Suppose that Assumptions 1, 2 hold. Then Algorithm 3 with $\gamma \leqslant \frac{(1-\alpha)\mu}{6L^2}, p = \frac{1}{n}$ and $V_s^t = \mathbb{E}\|z_s^t - z^*\|^2 + \mathbb{E}\|\omega_s^t - z^*\|^2$, to reach ε-accuracy, where $\varepsilon \sim V_S^n$, needs*

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

Remark 2. *Similarly to Remark 1, we can use our result in the monotone case by the regularization trick and obtain $\widetilde{\mathcal{O}}\big(n\frac{L^2}{\varepsilon^2}\big)$.*

Algorithm 3 RR/SO Extragradient with variance reduction

1: **Input: Parameters:** z_0^0, ω_0^0
2: **Parameter:** Stepsize γ, $\alpha \in (0,1)$
3: Generate a permutation $\pi_0, \pi_1, \ldots, \pi_{n-1}$ of sequence $\{1, 2, \ldots, n\}$ // SO heuristic
4: **for** $s = 0, 1, \ldots$ **do**
5: Generate a permutation $\pi_0, \pi_1, \ldots, \pi_{n-1}$ of sequence $\{1, 2, \ldots, n\}$ // RR heuristic
6: **for** $t = 0, 1, \ldots, n-1$ **do**
7: $\overline{z}_s^t = \alpha z_s^t + (1-\alpha)\omega_s^t$
8: $z_s^{t+1/2} = \mathrm{prox}_{\gamma g}\left(\overline{z}_s^t - \gamma F\left(\omega_s^t\right)\right)$
9: $\hat{F}(z_s^{t+1/2}) = F_{\pi_s^t}(z_s^{t+1/2}) - F_{\pi_s^t}(\omega_s^t) + F(\omega_s^t)$
10: $z_s^{t+1} = \mathrm{prox}_{\gamma g}\left(\overline{z}_s^t - \gamma\hat{F}\left(z_s^{t+1/2}\right)\right)$
11: $\omega_s^{t+1} = \begin{cases} z_s^t, & \text{with probability} \quad p \\ \omega_s^t & \text{with probability} \quad 1-p \end{cases}$
12: **end for**
13: $z_{s+1}^0 = z_s^n$
14: $\omega_{s+1}^0 = \omega_s^n$
15: **end for**
16: **Output:** z_S^n

We remove the variance that arose in Theorem 1 and obtain the linear convergence. Even though we get worse estimates than in the works that also use variance reduction technique, such as (Alacaoglu and Malitsky, 2022; Alacaoglu et al., 2021; Chavdarova et al., 2019; Palaniappan and Bach, 2016) (see Table 1), there is a distinct explanation for this. According to current theory, methods with the shuffling heuristic are worse than methods with independent sampling for the variance reduction methods (Malinovsky et al., 2023). Thus, we were unable to obtain theoretical convergence estimates for methods using shuffling heuristics that are equivalent to those for methods with independent index selection for the VI problem . Additionally, in the course of the work, no theoretical differences were revealed in the SO and RR techniques in relation to the problem (1).

4 Experiments

In this section, we evaluate the proposed algorithms to demonstrate their practical applications by conducting experiments in two cases: image denoising and adversarial training.

4.1 Image denoising

To formulate the image denoising problem (Chambolle and Pock, 2011), we consider the classic saddle point problem as we did 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 semicontinuous functions, and K is a continuous linear operator. To proceed to image denoising, we consider g is a given noisy image and u is a 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: $P = \{p \in \mathcal{Y} : \|p(x)\| \leq 1\}$. 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

$$\min_{u\in\mathcal{X}} \max_{p\in\mathcal{Y}} \left[-\langle u, \operatorname{div} p\rangle_{\mathcal{X}} + \lambda/2\|u - g\|_2^2 - \delta_P(p)\right]. \tag{7}$$

To bring the problem to the form of a finite sum (4), 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 the nearest neighbor.

We compare the RR/SO Extragradient with variance reduction (Algorithm 3) with Extragradient with variance reduction (Alacaoglu and Malitsky, 2022). Analogically we compare the RR/SO Extragradient (Algorithms 1, 2) and Extragradient (Juditsky et al., 2011). We select two images with different levels of additive zero-mean Gaussian noise: $\sigma = 0.05$ and $\sigma = 0.1$. Figures 1 and 2 provide a comparison of the proposed methods. Additional results for all considered methods on another image are presented in Figures 4, 5 in A.

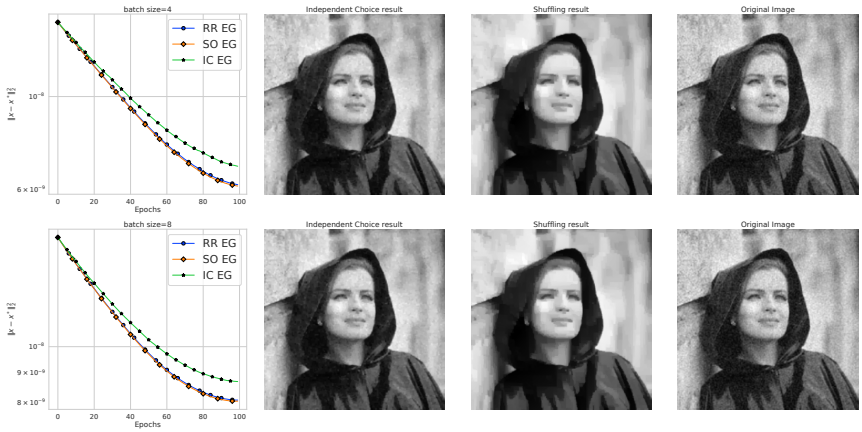

Figure 1: Extragradient convergence on image with $\sigma = 0.05$ on the problem (7).

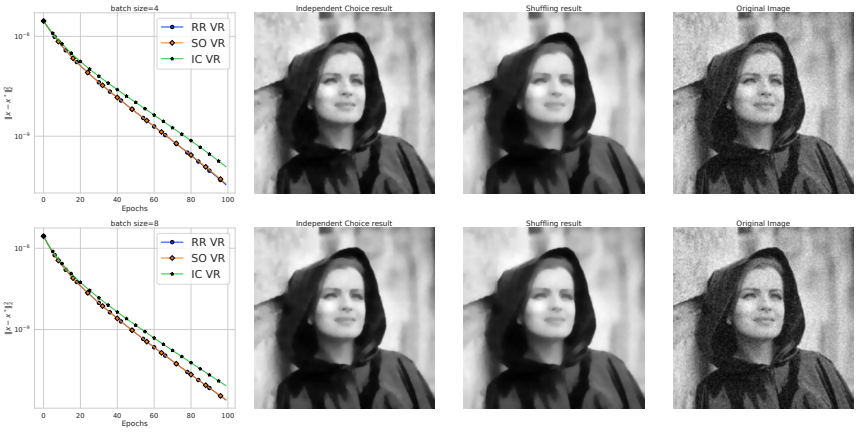

Figure 2: Extragradient with VR convergence on image with $\sigma = 0.05$ on the problem (7).

Comparing the images, it is evident that algorithms incorporating shuffling perform better than those that do not, even if the difference in the line graphs is subtle. Thus, using shuffling techniques allows us to achieve better results in solving such an important practical application as image denoising.

4.2 Adversarial training

Next, we address an adversarial training problem. We can formulate it in the following way:

$$\min_{w \in \mathbb{R}^d} \max_{\|r_i\| \leqslant D} \left[\frac{1}{2N} \sum_{i=1}^{N} \left(w^T (x_i + r_i) - y_i \right)^2 + \frac{\lambda}{2} \|w\|^2 - \frac{\beta}{2} \|r\|^2 \right], \tag{8}$$

where the samples corresponds to features x_i and targets y_i. We evaluate this issue across several datasets: `mushrooms`, `a9a`, and `w8a`, sourced from the LIBSVM library (Chang and Lin, 2011). A brief description of these datasets is provided in Table 2, A. The results are presented in Figure 3.

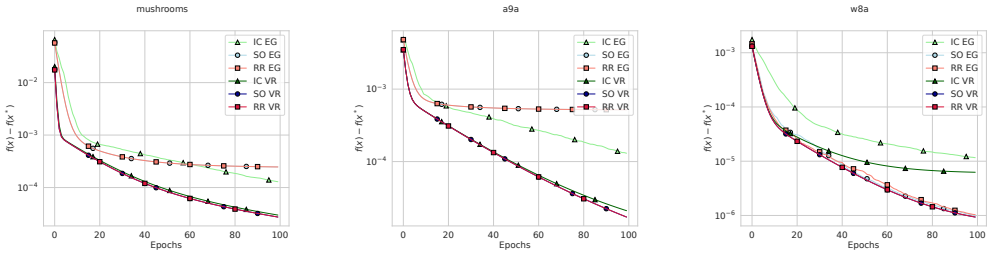

Figure 3: Extragradient with and without VR compared using various shuffling heuristics on `mushrooms`, `a9a` and `w8a` datasets on the problem (8).

As shown on plots, Extragradient and Extragradient with the VR algorithms using the independent choice of indexes demonstrate worse performance compared to those using shuffling methods. In these series of experiments, the RR exhibits better performance than other shuffling methods and significantly outperforms the non-shuffled versions.

References

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(2):321–346, 2021.

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.

Dimitri P Bertsekas and John N Tsitsiklis. Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642, 2000.

Aleksandr Beznosikov, Valentin Samokhin, and Alexander Gasnikov. Distributed saddle-point problems: Lower bounds, near-optimal and robust algorithms. *arXiv preprint arXiv:2010.13112*, 2020.

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.

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.

Yair Carmon, Yujia Jin, Aaron Sidford, and Kevin Tian. Variance reduction for matrix games. *Advances in Neural Information Processing Systems*, 32, 2019.

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.

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.

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.

Mert Gürbüzbalaban, Asu Ozdaglar, and Pablo A Parrilo. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186:49–84, 2021.

Jeff Haochen and Suvrit Sra. Random shuffling beats sgd after finite epochs. In *International Conference on Machine Learning*, pages 2624–2633. PMLR, 2019.

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.

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.

Xinmeng Huang, Kun Yuan, Xianghui Mao, and Wotao Yin. An improved analysis and rates for variance reduction under without-replacement sampling orders. *Advances in Neural Information Processing Systems*, 34:3232–3243, 2021.

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.

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

Anastasia Koloskova, Nikita Doikov, Sebastian U. Stich, and Martin Jaggi. On convergence of incremental gradient for non-convex smooth functions, 2024.

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.

Xiao Li, Zhihui Zhu, Anthony Man-Cho So, and Jason D Lee. Incremental methods for weakly convex optimization. *arXiv preprint arXiv:1907.11687*, 2019.

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.

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.

Grigory Malinovsky, Alibek Sailanbayev, and Peter Richtárik. Random reshuffling with variance reduction: New analysis and better rates. In *Uncertainty in Artificial Intelligence*, pages 1347–1357. PMLR, 2023.

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.

Olvi L Mangasarian and MV Solodov. Serial and parallel backpropagation convergence via nonmonotone perturbed minimization. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 1993.

Daniil Medyakov, Gleb Molodtsov, Aleksandr Beznosikov, and Alexander Gasnikov. Optimal data splitting in distributed optimization for machine learning. In *Doklady Mathematics*, volume 108, pages S465–S475. Springer, 2023.

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, Ahmed Khaled, and Peter Richtárik. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33: 17309–17320, 2020a.

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, 2020b.

Aryan Mokhtari, Mert Gurbuzbalaban, and Alejandro Ribeiro. Surpassing gradient descent provably: A cyclic incremental method with linear convergence rate. *SIAM Journal on Optimization*, 28(2):1420–1447, 2018.

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.

Dheeraj Nagaraj, Prateek Jain, and Praneeth Netrapalli. Sgd without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, pages 4703–4711. PMLR, 2019.

Angelia Nedic and Dimitri P Bertsekas. Incremental subgradient methods for nondifferentiable optimization. *SIAM Journal on Optimization*, 12(1):109–138, 2001.

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.

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.

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

Shashank Rajput, Anant Gupta, and Dimitris Papailiopoulos. Closing the convergence gap of sgd without replacement. In *International Conference on Machine Learning*, pages 7964–7973. PMLR, 2020.

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.

Itay Safran and Ohad Shamir. How good is sgd with random shuffling? In *Conference on Learning Theory*, pages 3250–3284. PMLR, 2020.

Sebastian U Stich. Unified optimal analysis of the (stochastic) gradient method. *arXiv preprint arXiv:1907.04232*, 2019.

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.

Bicheng Ying, Kun Yuan, and Ali H Sayed. Variance-reduced stochastic learning under random reshuffling. *IEEE Transactions on Signal Processing*, 68:1390–1408, 2020.

Appendix

Contents

1 Introduction **1**

2 Setup **5**

3 Algorithms and convergence analysis **5**

3.1 Extragradient 5

3.2 Extragradient with variance reduction 7

4 Experiments **8**

4.1 Image denoising 8

4.2 Adversarial training 9

References **13**

A Additional Experiments **15**

B Basic Inequalities **16**

C Extragradient **16**

D Extragradient with variance reduction **21**

A Additional Experiments

In this section, we present additional experiments that have been performed. Similar to the previous experiments, we observe a consistent pattern: methods incorporating shuffling techniques outperform those without shuffling. These results further confirm the effectiveness of shuffling techniques in solving the denoising problem on another image with higher σ.

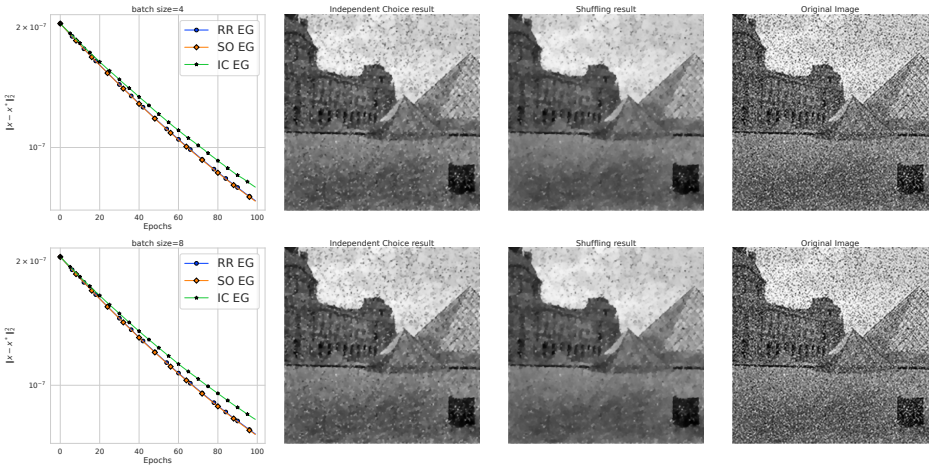

Figure 4: Extragradient convergence on image with $\sigma = 0.1$ on the problem (7).

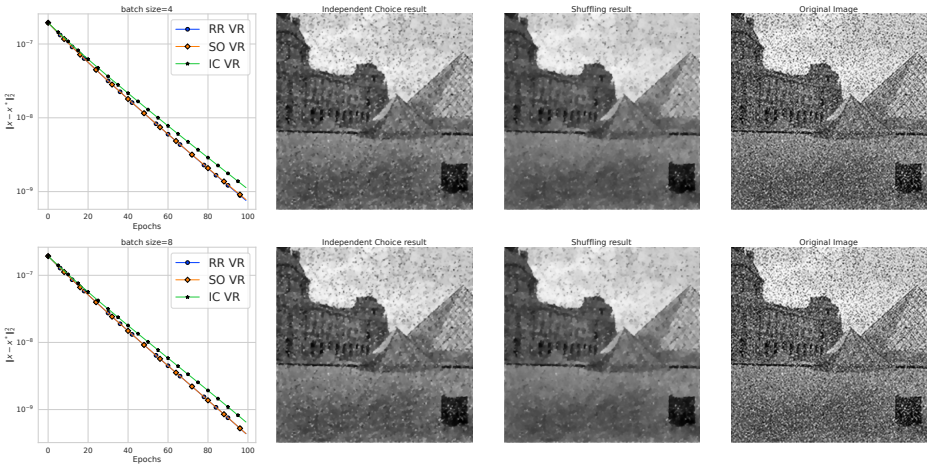

Figure 5: Extragradient with VR convergence on image with $\sigma = 0.1$ on the problem (7).

The datasets used for the experiments on the adversarial training include `mushrooms`, `a9a`, and `w8a`. These datasets vary in size and complexity, providing a comprehensive evaluation of our proposed algorithms in the context of adversarial training.

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

Table 2: Summary of Datasets

B Basic Inequalities

For all vectors $x, y, \{x_i\}_{i=1}^n$ in $\mathbb{R}^d$ with a positive scalar α, the following holds:

$$\begin{aligned}
\langle x, y\rangle &\leqslant \frac{\|x\|^2}{2\alpha} + \frac{\alpha\|y\|^2}{2}, && \text{(Scalar)}\\
2\langle x, y\rangle &= \|x+y\|^2 - \|x\|^2 - \|y\|^2, && \text{(Norm)}\\
-2\|x\|^2 &\leqslant -\|x+y\|^2 + 2\|y\|^2, && \text{(CS)}\\
\left\|\sum_{i=1}^n x_i\right\|^2 &\leqslant n\sum_{i=1}^n \|x_i\|^2, && \text{(Sum)}\\
\left\|\mathrm{prox}_{\gamma g}(x) - \mathrm{prox}_{\gamma g}(y)\right\|^2 &\leqslant \|x-y\|^2. && \text{(Prox)}
\end{aligned}$$

C Extragradient

Theorem 1. *Suppose Assumptions 1, 2 hold. Then for Algorithms 1, 2 with* $\gamma \leqslant \min\left\{\frac{1}{2\mu n}, \frac{1}{6L}\right\}$ *after* S *epochs,*

$$\|z_S^n - z^*\|^2 \leqslant (1 - \frac{\gamma\mu}{2})^{Sn}\|z_0^0 - z^*\|^2 + \frac{256\gamma n^2\sigma_*^2}{\mu}.$$

Proof. We start with the standard prox-inequality:

$$\hat{z} = \mathrm{prox}_g(z) \Longleftrightarrow \langle \hat{z} - z, u - \hat{z}\rangle \geqslant g(\hat{z}) - g(u), \quad \forall u \in \mathcal{Z}. \tag{9}$$

Substituting both steps of Algorithm 1 (Algorithm 2) into (9), we derive:

$$\begin{aligned}
\langle z_s^{t+1} - z_s^t + \gamma F_{\pi_s^t}(z_s^{t+1/2}), z^* - z_s^{t+1}\rangle &\geqslant \gamma(g(z_s^{t+1}) - g(z^*)),\\
\langle z_s^{t+1/2} - z_s^t + \gamma F_{\pi_s^t}(z_s^t), z_s^{t+1} - z_s^{t+1/2}\rangle &\geqslant \gamma(g(z_s^{t+1/2}) - g(z_s^{t+1})).
\end{aligned}$$

Summing inequalities, we get:

$$\begin{aligned}
\gamma(g(z_s^{t+1/2}) - g(z^*)) \leqslant{}& \langle z_s^{t+1} - z_s^t, z^* - z_s^{t+1}\rangle + \langle z_s^{t+1/2} - z_s^t, z_s^{t+1} - z_s^{t+1/2}\rangle\\
&+ \gamma\langle F_{\pi_s^t}(z_s^{t+1/2}), z^* - z_s^{t+1}\rangle + \gamma\langle F_{\pi_s^t}(z_s^t), z_s^{t+1} - z_s^{t+1/2}\rangle.
\end{aligned}$$

Now, we add and subtract $z_s^{t+1/2}$ to the right part of the third scalar product. Thus, rearranging terms, we arrive at:

$$\begin{aligned}
\gamma(g(z_s^{t+1/2}) - g(z^*)) \leqslant{}& \langle z_s^{t+1} - z_s^t, z^* - z_s^{t+1}\rangle + \langle z_s^{t+1/2} - z_s^t, z_s^{t+1} - z_s^{t+1/2}\rangle\\
&+ \gamma\langle F_{\pi_s^t}(z_s^{t+1/2}), z^* - z_s^{t+1/2}\rangle\\
&+ \gamma\langle F_{\pi_s^t}(z_s^{t+1/2}) - F_{\pi_s^t}(z_s^t), z_s^{t+1/2} - z_s^{t+1}\rangle.
\end{aligned}$$

We want to rewrite two first scalar products. We use (Norm). Thus, we come to:

$$\begin{aligned}
2\langle z_s^{t+1} - z_s^t, z^* - z_s^{t+1}\rangle &= \|z_s^t - z^*\|^2 - \|z_s^{t+1} - z_s^t\|^2 - \|z^* - z_s^{t+1}\|^2,\\
2\langle z_s^{t+1/2} - z_s^t, z_s^{t+1} - z_s^{t+1/2}\rangle &= \|z_s^{t+1} - z_s^t\|^2 - \|z_s^{t+1/2} - z_s^t\|^2 - \|z_s^{t+1} - z_s^{t+1/2}\|^2.
\end{aligned}$$

Substituting this, we get:

$$\begin{aligned}
\|z_s^{t+1} - z^*\|^2 \leqslant{}& \|z_s^t - z^*\|^2 - 2\gamma\left(\langle F_{\pi_s^t}(z_s^{t+1/2}), z_s^{t+1/2} - z^*\rangle + g(z_s^{t+1/2}) - g(z^*)\right)\\
&+ 2\gamma\langle F_{\pi_s^t}(z_s^{t+1/2}) - F_{\pi_s^t}(z_s^t), z_s^{t+1/2} - z_s^{t+1}\rangle\\
&- \|z_s^{t+1/2} - z_s^t\|^2 - \|z_s^{t+1} - z_s^{t+1/2}\|^2.
\end{aligned}$$

Now, applying (Scalar) and Assumption 1 to the second scalar product, we obtain:

$$
\begin{aligned}
\|z_s^{t+1} - z^*\|^2 &\leqslant \|z_s^t - z^*\|^2 - 2\gamma\big(\langle F_{\pi_s^t}(z_s^{t+1/2}), z_s^{t+1/2} - z^*\rangle \\
&\quad + g(z_s^{t+1/2}) - g(z^*)\big) + \gamma^2\|F_{\pi_s^t}(z_s^{t+1/2}) - F_{\pi_s^t}(z_s^t)\|^2 \\
&\quad + \|z_s^{t+1} - z_s^{t+1/2}\|^2 - \|z_s^{t+1/2} - z_s^t\|^2 - \|z_s^{t+1} - z_s^{t+1/2}\|^2 \\
&\overset{\text{Ass.1}}{\leqslant} \|z_s^t - z^*\|^2 + \left(\gamma^2 L^2 - 1\right)\|z_s^{t+1/2} - z_s^t\|^2 \\
&\quad - 2\gamma\left(\underbrace{\langle F_{\pi_s^t}(z_s^{t+1/2}), z_s^{t+1/2} - z^*\rangle + g(z_s^{t+1/2}) - g(z^*)}_{T_1}\right).
\end{aligned} \tag{10}
$$

To estimate the T_1 term, we take the expectation:

$$
\begin{aligned}
\mathbb{E}T_1 &= \mathbb{E}\langle F_{\pi_s^t}(z_s^{t+1/2}), z_s^{t+1/2} - z^*\rangle + g(z_s^{t+1/2}) - g(z^*) \\
&= \mathbb{E}\langle F_{\pi_s^t}(z_s^{t+1/2}) - F_{\pi_s^t}(z^*), z_s^{t+1/2} - z^*\rangle \\
&\quad + \mathbb{E}\langle F_{\pi_s^t}(z^*), z_s^{t+1/2} - z^*\rangle + g(z_s^{t+1/2}) - g(z^*) \\
&\overset{\text{Ass.2}}{\geqslant} \mu\mathbb{E}\|z_s^{t+1/2} - z^*\|^2 + \mathbb{E}\langle F_{\pi_s^t}(z^*), z_s^{t+1/2} - z_s^0\rangle \\
&\quad + \mathbb{E}\langle F_{\pi_s^t}(z^*), z_s^0 - z^*\rangle + g(z_s^{t+1/2}) - g(z^*).
\end{aligned}
$$

Now we pay attention to the second scalar product. Using tower property and unbiasedness of stochastic operator at the points z_s^0 and z^*, we have $\mathbb{E}\left[\mathbb{E}_t\left[F_{\pi_s^t}(z^*)|z_s^0 - z^*\right]\right] = F(z^*)$. Thus, we continue the estimation of $\mathbb{E}T_1$:

$$
\begin{aligned}
\mathbb{E}T_1 &\overset{\text{(Scalar)}}{\geqslant} \mu\mathbb{E}\|z_s^{t+1/2} - z^*\|^2 - \frac{\gamma}{2\beta}\mathbb{E}\|F_{\pi_s^t}(z^*)\|^2 - \frac{\beta}{2\gamma}\mathbb{E}\|z_s^{t+1/2} - z_s^0\|^2 \\
&\quad + \langle F(z^*), z_s^0 - z^*\rangle + g(z_s^{t+1/2}) - g(z^*) \\
&= \mu\mathbb{E}\|z_s^{t+1/2} - z^*\|^2 - \frac{\gamma}{2\beta}\mathbb{E}\|F_{\pi_s^t}(z^*)\|^2 - \frac{\beta}{2\gamma}\mathbb{E}\|z_s^{t+1/2} - z_s^0\|^2 \\
&\quad + \langle F(z^*), z_s^0 - z_s^{t+1/2}\rangle + \underbrace{\langle F(z^*), z_s^{t+1/2} - z^*\rangle + g(z_s^{t+1/2}) - g(z^*)}_{\geqslant 0 \ (1)} \\
&\overset{\text{(Scalar)}}{\geqslant} \mu\mathbb{E}\|z_s^{t+1/2} - z^*\|^2 - \frac{\gamma}{2\beta}\mathbb{E}\|F_{\pi_s^t}(z^*)\|^2 \\
&\quad - \frac{\gamma}{2\beta}\|F(z^*)\|^2 - \frac{\beta}{\gamma}\mathbb{E}\|z_s^{t+1/2} - z_s^0\|^2.
\end{aligned}
$$

Here we introduced $\beta > 0$, which we will define later. Substituting this inequality into (10), we obtain:

$$
\begin{aligned}
\mathbb{E}\|z_s^{t+1} - z^*\|^2 &\leqslant \mathbb{E}\|z_s^t - z^*\|^2 - 2\gamma\mu\mathbb{E}\|z_s^{t+\frac{1}{2}} - z^*\|^2 \\
&\quad + (\gamma^2 L^2 - 1)\mathbb{E}\|z_s^{t+\frac{1}{2}} - z_s^t\|^2 + \frac{\gamma^2}{\beta}\mathbb{E}\|F_{\pi_s^t}(z^*)\|^2 \\
&\quad + \frac{\gamma^2}{\beta}\|F(z^*)\|^2 + 2\beta\mathbb{E}\|z_s^{t+\frac{1}{2}} - z_s^0\|^2 \\
&\overset{\text{(Ass.3,}CS\text{)}}{\leqslant} (1 - \gamma\mu)\mathbb{E}\|z_s^t - z^*\|^2 + \frac{2\gamma^2}{\beta}\sigma_*^2 + 2\beta\mathbb{E}\|z_s^{t+\frac{1}{2}} - z_s^0\|^2 \\
&\quad + (\gamma^2 L^2 + 2\gamma\mu - 1)\mathbb{E}\|z_s^{t+\frac{1}{2}} - z_s^t\|^2.
\end{aligned} \tag{11}
$$

Now, we evaluate the $\|z_s^{t+\frac{1}{2}} - z_s^0\|^2$ term:

$$
\begin{aligned}
\|z_s^{t+\frac{1}{2}} - z_s^0\|^2 &\leqslant \left(1+\frac{1}{a}\right)\|z_s^{t+\frac{1}{2}} - z_s^t\|^2 + (1+a)\|z_s^t - z_s^0\|^2 \\
&\leqslant \left(1+\frac{1}{a}\right)\|z_s^{t+\frac{1}{2}} - z_s^t\|^2 + (1+a)\left(1+\frac{1}{b}\right)\|z_s^t - z_s^{t-\frac{1}{2}}\|^2 \\
&\quad + (1+a)(1+b)\|z_s^{t-\frac{1}{2}} - z_s^0\|^2 \\
&= \left(1+\frac{1}{a}\right)\|z_s^{t+\frac{1}{2}} - z_s^t\|^2 \\
&\quad + (1+a)\left(1+\frac{1}{b}\right)\Big\|\mathrm{prox}_{\gamma g}\left(z_s^{t-1} - \gamma F_{\pi_s^t}(z_s^{t-\frac{1}{2}})\right) \\
&\quad - \mathrm{prox}_{\gamma g}\left(z_s^{t-1} - \gamma F_{\pi_s^t}(z_s^{t-1})\right)\Big\|^2 + (1+a)(1+b)\|z_s^{t-\frac{1}{2}} - z_s^0\|^2 \\
&\overset{(\text{Prox})}{\leqslant} \left(1+\frac{1}{a}\right)\|z_s^{t+\frac{1}{2}} - z_s^t\|^2 \\
&\quad + (1+a)\left(1+\frac{1}{b}\right)\left\|\gamma F_{\pi_s^t}(z_s^{t-\frac{1}{2}}) - \gamma F_{\pi_s^t}(z_s^{t-1})\right\|^2 \\
&\quad + (1+a)(1+b)\|z_s^{t-\frac{1}{2}} - z_s^0\|^2 \\
&\leqslant \left(1+\frac{1}{a}\right)\|z_s^{t+\frac{1}{2}} - z_s^t\|^2 \\
&\quad + (1+a)\left(1+\frac{1}{b}\right)\gamma^2 L^2\|z_s^{t-\frac{1}{2}} - z_s^{t-1}\|^2 \\
&\quad + (1+a)(1+b)\|z_s^{t-\frac{1}{2}} - z_s^0\|^2 \\
&\leqslant \left(1+\frac{1}{a}\right)\|z_s^{t+\frac{1}{2}} - z_s^t\|^2 + \sum_{i=1}^{t-1}\|z_s^{i+\frac{1}{2}} - z_s^i\|^2 \\
&\quad \cdot \left(\left(1+\frac{1}{a}\right)(1+a)(1+b) + (1+a)\left(1+\frac{1}{b}\right)\gamma^2 L^2\right) \\
&\quad \cdot [(1+a)(1+b)]^{t-1-i} \\
&\quad + \left((1+a)\left(1+\frac{1}{b}\right)+1\right)[(1+a)(1+b)]^t\,\|z_s^{\frac{1}{2}} - z_s^0\|^2.
\end{aligned}
$$

We choose $a = b = \frac{1}{n}$ and consider coefficients before all three terms.

$$
\begin{aligned}
1+\frac{1}{a} &= 1+n, \\
\left(\left(1+\frac{1}{a}\right)+\frac{1}{b}\gamma^2 L^2\right)[(1+a)(1+b)]^{t-i} &= (1+n+n\gamma^2 L^2)\left(1+\frac{1}{n}\right)^{2(t-i)}, \\
\left((1+a)\left(1+\frac{1}{b}\right)+1\right)[(1+a)(1+b)]^{t-i}\Big|_{i=0} & \\
&= \left(3+\frac{1}{n}+n\right)\left(1+\frac{1}{n}\right)^{2(t-i)}\Big|_{i=0}.
\end{aligned}
$$

We can evaluate the smaller terms from above by the biggest one and write them into one sum:

$$
\|z_s^{t+\frac{1}{2}} - z_s^0\|^2 \leqslant \sum_{i=0}^{t}\|z_s^{i+\frac{1}{2}} - z_s^i\|^2\left(1+n+n\gamma^2 L^2\right)\left(1+\frac{1}{n}\right)^{2(t-i)}.
$$

Let us substitute the obtained inequality into (11).

$$\mathbb{E}\|z_s^{t+1} - z^*\|^2 \leqslant (1-\gamma\mu)\mathbb{E}\|z_s^t - z^*\|^2 + (\gamma^2L^2 + 2\gamma\mu - 1)\mathbb{E}\|z_s^{t+\frac{1}{2}} - z_s^t\|^2 + \frac{2\gamma^2}{\beta}\sigma_*^2 + 2\beta\sum_{i=0}^{t}\mathbb{E}\|z_s^{i+\frac{1}{2}} - z_s^i\|^2\left(1+n+n\gamma^2L^2\right)\left(1+\frac{1}{n}\right)^{2(t-i)}. \tag{12}$$

Now we define new sequence that contains iteration points in all epochs:

$$\widetilde{z}_k = z_{t+sn}.$$

Thus, additionally considering $\left(1+\frac{1}{n}\right)^{2(t-i)} \leqslant \left(1+\frac{1}{n}\right)^{2n} \leqslant e^2 \leqslant 8$, we can rewrite (12) in the following form:

$$\mathbb{E}\|\widetilde{z}_{k+1} - z^*\|^2 \leqslant (1-\gamma\mu)\mathbb{E}\|\widetilde{z}_k - z^*\|^2 + (\gamma^2L^2 + 2\gamma\mu - 1)\mathbb{E}\|\widetilde{z}_{k+\frac{1}{2}} - \widetilde{z}_k\|^2 + \frac{2\gamma^2}{\beta}\sigma_*^2 + 16\beta\sum_{i=0}^{n}\mathbb{E}\|\widetilde{z}_{k-i+\frac{1}{2}} - \widetilde{z}_{k-i}\|^2\left(1+n+n\gamma^2L^2\right).$$

Let us pay attention to the $\sum\limits_{i=0}^{n}\mathbb{E}\|\widetilde{z}_{k-i+\frac{1}{2}} - \widetilde{z}_{k-i}\|^2$ term in the obtained inequality. For the original sequence, this term represented the sum of the norms from the beginning of the current epoch to the current iteration t and could contain a maximum of n terms. Thus, a new expression that contains the sum of n norms up to the current iteration k is an upper bound and our expression is correct. Now we define $p_k = p^k = \left(1-\frac{\gamma\mu}{2}\right)^{-k}$ and summarize both sides over all iterations with coefficients p_k:

$$\begin{aligned}\sum_{k=0}^{Sn-1} p_k\mathbb{E}\|\widetilde{z}_{k+1} - z^*\|^2 \leqslant{}& (1-\gamma\mu)\sum_{k=0}^{Sn-1} p_k\mathbb{E}\|\widetilde{z}_k - z^*\|^2 \\ &+ (\gamma^2L^2 + 2\gamma\mu - 1)\sum_{k=0}^{Sn-1} p_k\mathbb{E}\|\widetilde{z}_{k+\frac{1}{2}} - \widetilde{z}_k\|^2 \\ &+ \sum_{k=0}^{Sn-1} p_k\sum_{i=0}^{n}\mathbb{E}\|\widetilde{z}_{k-i+\frac{1}{2}} - \widetilde{z}_{k-i}\|^2 \\ &\cdot 16\beta\left(1+n+n\gamma^2L^2\right) + \frac{2\gamma^2\sigma_*^2}{\beta}\sum_{k=0}^{Sn-1} p_k.\end{aligned} \tag{13}$$

Now we need to estimate the following term:

$$\begin{aligned}\sum_{k=0}^{Sn-1} p_k\sum_{i=0}^{n}\mathbb{E}\|\widetilde{z}_{k-i+\frac{1}{2}} - \widetilde{z}_{k-i}\|^2 &\leqslant p_n\sum_{k=0}^{Sn-1}\sum_{i=0}^{n} p_{k-i}\mathbb{E}\|\widetilde{z}_{k-i+\frac{1}{2}} - \widetilde{z}_{k-i}\|^2 \\ &\leqslant p_n n\sum_{k=0}^{Sn-1} p_k\mathbb{E}\|\widetilde{z}_{k+\frac{1}{2}} - \widetilde{z}_k\|^2.\end{aligned}$$

Note that we define points $\widetilde{z}_{-n}, \widetilde{z}_{-n+\frac{1}{2}}, \ldots, \widetilde{z}_{-\frac{1}{2}}$ by shifting the sequence $\{\widetilde{z}_k\}$ on n points. Since $p_n = \left(1-\frac{\gamma\mu}{2}\right)^{-n} = \left(1-\frac{\gamma\mu n}{2n}\right)^{-n} \leqslant e^{\frac{\gamma\mu n}{2}}$, we choose $\gamma \leqslant \frac{1}{2\mu n}$ and obtain $p_n \leqslant e^{\frac{1}{4}} \leqslant 2$. Substituting this into (13), we obtain:

$$\begin{aligned}\sum_{k=0}^{Sn-1} p_k\mathbb{E}\|\widetilde{z}_{k+1} - z^*\|^2 \leqslant{}& (1-\gamma\mu)\sum_{k=0}^{Sn-1} p_k\mathbb{E}\|\widetilde{z}_k - z^*\|^2 + \frac{2\gamma^2\sigma_*^2}{\beta}\sum_{k=0}^{Sn-1} p_k \\ &+ (\gamma^2L^2 + 2\gamma\mu - 1)\sum_{k=0}^{Sn-1} p_k\mathbb{E}\|\widetilde{z}_{k+\frac{1}{2}} - \widetilde{z}_k\|^2 \\ &+ 32\beta\left(1+n+n\gamma^2L^2\right)n\sum_{k=0}^{Sn-1} p_k\mathbb{E}\|\widetilde{z}_{k+\frac{1}{2}} - \widetilde{z}_k\|^2.\end{aligned}$$

We consider the coefficient before $\sum\limits_{k=0}^{Sn-1} p_k \mathbb{E}\|\widetilde{z}_{k+\frac{1}{2}} - \widetilde{z}_k\|^2$ and we make it negative by selecting γ and β.

$$32\beta(1+n+n\gamma^2L^2)n + \gamma^2L^2 + 2\gamma\mu - 1 \leqslant 0$$

We need $\gamma \leqslant \frac{1}{6L}, \beta = \frac{1}{64n^2}$. Then to satisfy the previous estimate on gamma we finally put $\gamma \leqslant \min\left\{\frac{1}{2\mu n}, \frac{1}{6L}\right\}$ and, assuming $n > 3$, have

$$\frac{1}{2n} + \frac{1}{2} + \frac{1}{72} + \frac{1}{36} + \frac{1}{3} - 1 \leqslant 0.$$

In this way,

$$\sum_{k=0}^{Sn-1} p_k \mathbb{E}\|\widetilde{z}_{k+1} - z^*\|^2 \leqslant (1-\gamma\mu) \sum_{k=0}^{Sn-1} p_k \mathbb{E}\|\widetilde{z}_k - z^*\|^2 + \frac{2\gamma^2\sigma_*^2}{\beta} \sum_{k=0}^{Sn-1} p_k.$$

Thus, substituting definition of p_t, we obtain:

$$\begin{aligned}
\sum_{k=0}^{Sn-1} \left(1-\frac{\gamma\mu}{2}\right)^{-k} \mathbb{E}\|\widetilde{z}_{k+1} - z^*\|^2 &\leqslant \sum_{k=0}^{Sn-1} \left(1-\frac{\gamma\mu}{2}\right)^{-k+1} \mathbb{E}\|\widetilde{z}_k - z^*\|^2 \\
&\quad + \frac{2\gamma^2\sigma_*^2}{\beta} \sum_{k=0}^{Sn-1} \left(1-\frac{\gamma\mu}{2}\right)^{-k}, \\
\left(1-\frac{\gamma\mu}{2}\right)^{-(Sn-1)} \mathbb{E}\|\widetilde{z}_{Sn} - z^*\|^2 &\leqslant \left(1-\frac{\gamma\mu}{2}\right) \mathbb{E}\|\widetilde{z}_0 - z^*\|^2 \\
&\quad + \frac{2\gamma^2\sigma_*^2}{\beta} \sum_{k=0}^{Sn-1} \left(1-\frac{\gamma\mu}{2}\right)^{-k}, \\
\mathbb{E}\|\widetilde{z}_{Sn} - z^*\|^2 &\leqslant \left(1-\frac{\gamma\mu}{2}\right)^{Sn} \mathbb{E}\|\widetilde{z}_0 - z^*\|^2 \\
&\quad + \frac{2\gamma^2\sigma_*^2}{\beta} \sum_{k=0}^{Sn-1} \left(1-\frac{\gamma\mu}{2}\right)^{Sn-k-1} \\
&= \left(1-\frac{\gamma\mu}{2}\right)^{Sn} \mathbb{E}\|\widetilde{z}_0 - z^*\|^2 \\
&\quad + \frac{2\gamma^2\sigma_*^2}{\beta} \sum_{k=0}^{Sn-1} \left(1-\frac{\gamma\mu}{2}\right)^{k}.
\end{aligned}$$

Finally, estimating geometric progression in the last term as $\sum\limits_{k=0}^{Sn-1} \left(1-\frac{\gamma\mu}{2}\right)^k \leqslant \frac{2}{\gamma\mu}$, we can write the final statement of the theorem:

$$\begin{aligned}
\mathbb{E}\|z_S^n - z^*\|^2 &\leqslant \left(1-\frac{\gamma\mu}{2}\right)^{Sn} \mathbb{E}\|z_0^0 - z^*\|^2 + \frac{4\gamma\sigma_*^2}{\beta\mu} \\
&= \left(1-\frac{\gamma\mu}{2}\right)^{Sn} \mathbb{E}\|z_0^0 - z^*\|^2 + \frac{256\gamma n^2\sigma_*^2}{\mu}.
\end{aligned}$$

□

Corollary 1. *Suppose Assumptions 1, 2 hold. Then Algorithms 1, 2 with* $\gamma \leqslant \min\left\{\frac{1}{2\mu n}, \frac{1}{6L}, \frac{2\log\left(\max\left\{2, \frac{\mu^2\|z_0^0 - z^*\|^2 T}{512n^2\sigma_*^2}\right\}\right)}{\mu T}\right\}$, *to reach* ε*-accuracy, where* $\varepsilon \sim \|z_S^n - z^*\|^2$, *needs*

$$\widetilde{\mathcal{O}}\left(\left(n + \frac{L}{\mu}\right)\log\left(\frac{1}{\varepsilon}\right) + \frac{n^2\sigma_*^2}{\mu^2\varepsilon}\right) \quad \textit{iterations and oracle calls.}$$

Proof. For the result obtained in Theorem 1, we utilize Lemma 2 from (Stich, 2019) and, using special tuning of γ, such as $\gamma \leqslant \min\left\{\frac{1}{2\mu n}, \frac{1}{6L}, \frac{2\log\left(\max\left\{2, \frac{\mu^2\|z_0^0 - z^*\|^2 T}{512 n^2 \sigma_*^2}\right\}\right)}{\mu T}\right\}$, we obtain that we need $\widetilde{\mathcal{O}}\left(\left(n + \frac{L}{\mu}\right)\log\left(\frac{1}{\varepsilon}\right) + \frac{n^2\sigma_*^2}{\mu^2\varepsilon}\right)$ iterations and oracle calls to reach ε-accuracy, where $\varepsilon \sim \|z_S^n - z^*\|^2$. □

D Extragradient with variance reduction

Theorem 2. *Suppose that Assumptions 1, 2 hold. Then for Algorithm 3 with $\gamma \leqslant \frac{(1-\alpha)\mu}{6L^2}, p = \frac{1}{n}$ and $V_s^t = \mathbb{E}\|z_s^t - z^*\|^2 + \mathbb{E}\|\omega_s^t - z^*\|^2$, after T iterations we have*

$$V_S^n \leqslant \left(1 - \frac{\gamma\mu}{4}\right)^T V_0^0.$$

Proof. We start with substituting both steps of Algorithm 3 to (9):

$$\langle z_s^{t+1} - \overline{z}_s^t + \gamma\hat{F}(z_s^{t+1/2}), z^* - z_s^{t+1}\rangle \geqslant \gamma(g(z_s^{t+1}) - g(z^*)),$$
$$\langle z_s^{t+1/2} - \overline{z}_s^t + \gamma F(\omega_s^t), z_s^{t+1} - z_s^{t+1/2}\rangle \geqslant \gamma(g(z_s^{t+1/2}) - g(z_s^{t+1})).$$

Let us summarize this two inequalities:

$$\begin{aligned}\gamma(g(z_s^{t+1/2}) - g(z^*)) \leqslant\ & \langle z_s^{t+1} - \overline{z}_s^t, z^* - z_s^{t+1}\rangle + \langle z_s^{t+1/2} - \overline{z}_s^t, z_s^{t+1} - z_s^{t+1/2}\rangle \\ & + \gamma\langle\hat{F}(z_s^{t+1/2}), z^* - z_s^{t+1}\rangle + \gamma\langle F(\omega_s^t), z_s^{t+1} - z_s^{t+1/2}\rangle.\end{aligned}$$

Now, we add and subtract $z_s^{t+1/2}$ to the right part of the third scalar product. Thus, rearranging terms and utilizing the definition of $\hat{F}(z_s^{t+1/2})$, we arrive at:

$$\begin{aligned}&\underbrace{\langle z_s^{t+1} - \overline{z}_s^t, z^* - z_s^{t+1}\rangle}_{T_1} + \underbrace{\langle z_s^{t+1/2} - \overline{z}_s^t, z_s^{t+1} - z_s^{t+1/2}\rangle}_{T_2} \\ &\quad + \underbrace{\gamma\langle F_{\pi_s^t}(\omega_s^t) - F_{\pi_s^t}(z_s^{t+1/2}), z_s^{t+1} - z_s^{t+1/2}\rangle}_{T_3} \\ &\quad + \underbrace{\gamma\langle\hat{F}(z_s^{t+1/2}), z^* - z_s^{t+1/2}\rangle + \gamma(g(z^*) - g(z_s^{t+1/2}))}_{T_4} \geqslant 0.\end{aligned} \tag{14}$$

We defined terms as T_1, T_2, T_3, T_4, respectively. Let us estimate them separately. We start with T_1 and T_2. To estimate them, firstly, we use the definition of $\overline{z}_s^t$ and, secondly, use (Norm). Thus, we obtain:

$$\begin{aligned}2T_1 &= 2\langle z_s^{t+1} - \overline{z}_s^t, z^* - z_s^{t+1}\rangle \\ &= 2\alpha\langle z_s^{t+1} - z_s^t, z^* - z_s^{t+1}\rangle + 2(1-\alpha)\langle z_s^{t+1} - \omega_s^t, z^* - z_s^{t+1}\rangle \\ &= \alpha(\|z^* - z_s^t\|^2 - \|z_s^{t+1} - z_s^t\|^2 - \|z^* - z_s^{t+1}\|^2) \\ &\quad + (1-\alpha)(\|z^* - \omega_s^t\|^2 - \|z_s^{t+1} - \omega_s^t\|^2 - \|z^* - z_s^{t+1}\|^2) \\ &= \alpha\|z_s^t - z^*\|^2 - \|z_s^{t+1} - z^*\|^2 + (1-\alpha)\|\omega_s^t - z^*\|^2 \\ &\quad - \alpha\|z_s^{t+1} - z_s^t\|^2 - (1-\alpha)\|z_s^{t+1} - \omega_s^t\|^2.\end{aligned}$$

The same holds for T_2:

$$\begin{aligned}2T_2 &= 2\langle z_s^{t+1/2} - \overline{z}_s^t, z_s^{t+1} - z_s^{t+1/2}\rangle \\ &= 2\alpha\langle z_s^{t+1/2} - z_s^t, z_s^{t+1} - z_s^{t+1/2}\rangle + 2(1-\alpha)\langle z_s^{t+1/2} - \omega_s^t, z_s^{t+1} - z_s^{t+1/2}\rangle \\ &= \alpha(\|z_s^{t+1} - z_s^t\|^2 - \|z_s^{t+1/2} - z_s^t\|^2 - \|z_s^{t+1} - z_s^{t+1/2}\|^2) \\ &\quad + (1-\alpha)(\|z_s^{t+1} - \omega_s^t\|^2 - \|z_s^{t+1/2} - \omega_s^t\|^2 - \|z_s^{t+1} - z_s^{t+1/2}\|^2) \\ &= \alpha\|z_s^{t+1} - z_s^t\|^2 - \|z_s^{t+1} - z_s^{t+1/2}\|^2 + (1-\alpha)\|z_s^{t+1} - \omega_s^t\|^2 \\ &\quad - \alpha\|z_s^{t+1/2} - z_s^t\|^2 - (1-\alpha)\|z_s^{t+1/2} - \omega_s^t\|^2.\end{aligned}$$

Now, we moving to the estimate of T_3:

$$
\begin{aligned}
2T_3 &= 2\gamma\langle F_{\pi_s^t}(\omega_s^t) - F_{\pi_s^t}(z_s^{t+1/2}), z_s^{t+1} - z_s^{t+1/2}\rangle \\
&\overset{\text{(Scalar)}}{\leqslant} \frac{\gamma^2}{\tau}\|F_{\pi_s^t}(\omega_s^t) - F_{\pi_s^t}(z_s^{t+1/2})\|^2 + \tau\|z_s^{t+1} - z_s^{t+1/2}\|^2 \\
&\overset{\text{(Sum)}}{\leqslant} \frac{\gamma^2 L^2}{\tau}\|z_s^{t+1/2} - \omega_s^t\|^2 + \tau\|z_s^{t+1} - z_s^{t+1/2}\|^2;
\end{aligned}
$$

Here we introduced $\tau > 0$, which we will define later. Last, we do the same for T_4:

$$
\begin{aligned}
2T_4 &= 2\gamma\langle \hat{F}(z_s^{t+1/2}), z^* - z_s^{t+1/2}\rangle + 2\gamma(g(z^*) - g(z_s^{t+1/2})) \\
&= 2\gamma\langle \hat{F}(z_s^{t+1/2}) - F(z_s^{t+1/2}), z^* - z_s^{t+1/2}\rangle \\
&\quad + 2\gamma\langle F(z_s^{t+1/2}) - F(z^*), z^* - z_s^{t+1/2}\rangle \\
&\quad + 2\gamma\left(\underbrace{\langle F(z^*), z^* - z_s^{t+1/2}\rangle + g(z^*) - g(z_s^{t+1/2})}_{\leqslant 0 \ (1)}\right) \\
&\overset{(\textit{Scalar},\text{Ass.1})}{\leqslant} \frac{4\gamma^2 L^2}{\tau}\|z_s^{t+1/2} - \omega_s^t\|^2 + \tau\|z_s^{t+1/2} - z^*\|^2 \\
&\quad - 2\gamma\langle F(z_s^{t+1/2}) - F(z^*), z_s^{t+1/2} - z^*\rangle \\
&\overset{\text{(Ass.2)}}{\leqslant} \frac{4\gamma^2 L^2}{\tau}\|z_s^{t+1/2} - \omega_s^t\|^2 + \tau\|z_s^{t+1/2} - z^*\|^2 - 2\gamma\mu\|z_s^{t+1/2} - z^*\|^2.
\end{aligned}
$$

Substituting all the obtained estimates into (14), we arrive at:

$$
\begin{aligned}
0 \leqslant\ & \alpha\|z_s^t - z^*\|^2 - \|z_s^{t+1} - z^*\|^2 + (1-\alpha)\|\omega_s^t - z^*\|^2 - \alpha\|z_s^{t+1} - z_s^t\|^2 \\
& - (1-\alpha)\|z_s^{t+1} - \omega_s^t\|^2 + \alpha\|z_s^{t+1} - z_s^t\|^2 - \|z_s^{t+1} - z_s^{t+1/2}\|^2 \\
& + (1-\alpha)\|z_s^{t+1} - \omega_s^t\|^2 - \alpha\|z_s^{t+1/2} - z_s^t\|^2 - (1-\alpha)\|z_s^{t+1/2} - \omega_s^t\|^2 \\
& + \frac{\gamma^2 L^2}{\tau}\|z_s^{t+1/2} - \omega_s^t\|^2 + \tau\|z_s^{t+1} - z_s^{t+1/2}\|^2 + \frac{4\gamma^2 L^2}{\tau}\|z_s^{t+1/2} - \omega_s^t\|^2 \\
& + \tau\|z_s^{t+1/2} - z^*\|^2 - 2\gamma\mu\|z_s^{t+1/2} - z^*\|^2.
\end{aligned}
$$

By grouping the coefficients of the same terms, we get:

$$
\begin{aligned}
\|z_s^{t+1} - z^*\|^2 \leqslant\ & \alpha\|z_s^t - z^*\|^2 + (1-\alpha)\|\omega_s^t - z^*\|^2 \\
& + \left(\frac{5\gamma^2 L^2}{\tau} - (1-\alpha)\right)\|z_s^{t+1/2} - \omega_s^t\|^2 - (1-\tau)\|z_s^{t+1} - z_s^{t+1/2}\|^2 \\
& - (2\gamma\mu - \tau)\|z_s^{t+1/2} - z^*\|^2 - \alpha\|z_s^{t+1/2} - z_s^t\|^2.
\end{aligned} \tag{15}
$$

Now, we want to estimate the $-(2\gamma\mu - \tau)\|z_s^{t+1/2} - z^*\|^2$ term. To do this we split it into two equal parts. To the first part we add and subtract ω_s^t, and to the second $-\ z_s^t$. After that we use (CS) for both terms:

$$
\begin{aligned}
-(2\gamma\mu - \tau)\|z_s^{t+1/2} - z^*\|^2 &= -\left(\gamma\mu - \frac{\tau}{2}\right)\|z_s^{t+1/2} - z^*\|^2 - \left(\gamma\mu - \frac{\tau}{2}\right)\|z_s^{t+1/2} - z^*\|^2 \\
&= -\left(\gamma\mu - \frac{\tau}{2}\right)\|z_s^{t+1/2} - \omega_s^t + \omega_s^t - z^*\|^2 \\
&\quad - \left(\gamma\mu - \frac{\tau}{2}\right)\|z_s^{t+1/2} - z_s^t + z_s^t - z^*\|^2 \\
&\leqslant \left(\gamma\mu - \frac{\tau}{2}\right)\|z_s^{t+1/2} - \omega_s^t\|^2 - \left(\frac{\gamma\mu}{2} - \frac{\tau}{4}\right)\|\omega_s^t - z^*\|^2 \\
&\quad + \left(\gamma\mu - \frac{\tau}{2}\right)\|z_s^{t+1/2} - z_s^t\|^2 - \left(\frac{\gamma\mu}{2} - \frac{\tau}{4}\right)\|z_s^t - z^*\|^2.
\end{aligned}
$$

Substituting this into (15),

$$\|z_s^{t+1} - z^*\|^2 \leqslant \left(\alpha - \frac{\gamma\mu}{2} + \frac{\tau}{4}\right)\|z_s^t - z^*\|^2 + \left(1 - \alpha - \frac{\gamma\mu}{2} + \frac{\tau}{4}\right)\|\omega_s^t - z^*\|^2$$
$$+ \left(\frac{5\gamma^2 L^2}{\tau} + \gamma\mu - \frac{\tau}{2} - (1-\alpha)\right)\|z_s^{t+1/2} - \omega_s^t\|^2$$
$$- (1-\tau)\|z_s^{t+1} - z_s^{t+1/2}\|^2 + \left(\gamma\mu - \frac{\tau}{2} - \alpha\right)\|z_s^{t+1/2} - z_s^t\|^2.$$

We want to choose parameters such that coefficients before the last three terms would be non-positive. Let us start with $\|z_s^{t+1/2} - \omega_s^t\|^2$ term.

$$\text{We pick } \tau = \gamma\mu;$$
$$\text{We want } 1 - \alpha \geqslant \frac{5\gamma L^2}{\mu} + \frac{\gamma\mu}{2};$$
$$\text{It is enough for us that } \gamma \leqslant \frac{(1-\alpha)\mu}{6L^2}.$$

Obviously, with this choice of γ and α, the last two terms are less than zero. In that way, we obtain:

$$\|z_s^{t+1} - z^*\|^2 \leqslant \left(\alpha - \frac{\gamma\mu}{4}\right)\|z_s^t - z^*\|^2 + \left(1 - \alpha - \frac{\gamma\mu}{4}\right)\|\omega_s^t - z^*\|^2.$$

According to the condition for updating the point ω_s^t,

$$\mathbb{E}\|\omega_s^{t+1} - z^*\|^2 = p\|z_s^t - z^*\|^2 + (1-p)\|\omega_s^t - z^*\|^2.$$

In that way:

$$\mathbb{E}\|z_s^{t+1} - z^*\|^2 + \frac{1-\alpha}{p}\mathbb{E}\|\omega_s^{t+1} - z^*\|^2 \leqslant \left(1 - \frac{\gamma\mu}{4}\right)\mathbb{E}\|z_s^t - z^*\|^2$$
$$+ \left((1-\alpha)\left(1 + \frac{1}{p} - 1\right) - \frac{\gamma\mu}{4}\right)\mathbb{E}\|\omega_s^{t+1} - z^*\|^2.$$

Now we put $p = 1 - \alpha$ and obtain:

$$\mathbb{E}\|z_s^{t+1} - z^*\|^2 + \mathbb{E}\|\omega_s^{t+1} - z^*\|^2 \leqslant \left(1 - \frac{\gamma\mu}{4}\right)\left(\mathbb{E}\|z_s^t - z^*\|^2 + \mathbb{E}\|\omega_s^t - z^*\|^2\right).$$

Denoting $V_s^t = \mathbb{E}\|z_s^t - z^*\|^2 + \mathbb{E}\|\omega_s^t - z^*\|^2$ and going into recursion over all epochs and iterations, we get:

$$V_S^n \leqslant \left(1 - \frac{\gamma\mu}{4}\right)^T V_0^0,$$

where T is the total number of iterations. □

Corollary 2 *Suppose that Assumptions 1, 2 hold. Then Algorithm 3 with* $\gamma \leqslant \frac{(1-\alpha)\mu}{6L^2}, p = \frac{1}{n}$ *and* $V_s^t = \mathbb{E}\|z_s^t - z^*\|^2 + \mathbb{E}\|\omega_s^t - z^*\|^2$*, to reach* ε*-accuracy, where* $\varepsilon \sim V_S^n$*, needs*

$$\mathcal{O}\left(n\frac{L^2}{\mu^2}\log\left(\frac{1}{\varepsilon}\right)\right) \quad \text{iterations and oracle calls.}$$

Proof. Substituting estimation of γ to the result of Theorem 2 we obtain, that method to converge to ε-accuracy, where $\varepsilon = V_S^n$, needs $\mathcal{O}\left(\frac{L^2}{p\mu^2}\log\left(\frac{1}{\varepsilon}\right)\right)$ iterations. At the same time each iteration costs $pn + 2$ calls to F_π. Thus, we obtain $\mathcal{O}\left(\left(n\frac{L^2}{\mu^2} + \frac{L^2}{p\mu^2}\right)\log\left(\frac{1}{\varepsilon}\right)\right)$ oracle complexity. Finally, the optimal choice $p = \frac{1}{n}$ gives $\mathcal{O}\left(n\frac{L^2}{\mu^2}\log\left(\frac{1}{\varepsilon}\right)\right)$ iteration and oracle complexity. This ends the proof. □