EFFICIENT STOCHASTIC ALGORITHMS FOR CONTINUAL FINITE-SUM VARIATIONAL INEQUALITIES

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper considers the continual finite-sum variational inequalities. Specifically, we seek a sequence $\boldsymbol{z}_1^*, \ldots, \boldsymbol{z}_n^*$ which corresponds to the solutions of prefix-sum variational inequalities $\{\frac{1}{i}\sum_{j=1}^{i}\langle F_j(\boldsymbol{z}_i^*), \boldsymbol{z} - \boldsymbol{z}_i^*\rangle \geq 0\}_{i=1}^n$, where each component operator $F_j \coloneqq \mathcal{Z} \to \mathbb{R}$ is strongly-monotone, and the feasible set $\mathcal{Z} \subseteq \mathbb{R}^d$ is convex and compact. We propose an efficient stochastic algorithm that finds a sequence of ϵ-approximate solutions for the continual finite-sum variational inequalities. In particular, our approach sparsely constructs the full operator across all stages, and it leverages a novel extragradient iteration to achieve a sharper incremental oracle complexity compared with existing methods. Furthermore, we conduct numerical experiments that demonstrate the effectiveness of our approaches.

1 INTRODUCTION

In recent years, we have witnessed a surge of interest in finite-sum variational inequality (VI) problems, that is, finding the solution $\boldsymbol{z}^* \in \mathcal{Z}$ that satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\langle F_i(\boldsymbol{z}^*), \boldsymbol{z} - \boldsymbol{z}^*\rangle \geq 0 \quad \text{for all } \boldsymbol{z} \in \mathcal{Z}, \tag{1}$$

where $\mathcal{Z} \subseteq \mathbb{R}^d$ is convex and compact. This formulation encompasses the following fundamental problems.

Problem 1.1. *The finite-sum minimax optimization problem*

$$\min_{\boldsymbol{x}\in\mathcal{X}}\max_{\boldsymbol{y}\in\mathcal{Y}}\frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}, \boldsymbol{y}),$$

where each component $f_i(\boldsymbol{x}, \boldsymbol{y})$ is strongly-convex in $\boldsymbol{x}$ and strongly-concave in $\boldsymbol{y}$. This formulation is equivalent to problem (1) by taking

$$\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{y}), \quad F_i(\boldsymbol{z}) \coloneqq \begin{pmatrix} \nabla_x f_i(\boldsymbol{x}, \boldsymbol{y}) \\ -\nabla_y f_i(\boldsymbol{x}, \boldsymbol{y}) \end{pmatrix}, \quad \mathcal{Z} \coloneqq \mathcal{X} \times \mathcal{Y}.$$

Problem 1.2. *The finite-sum minimization problem*

$$\min_{\boldsymbol{x}\in\mathcal{X}}\frac{1}{n}\sum_{i=1}^{n} f_i(\boldsymbol{x}),$$

where $f_i(\boldsymbol{x})$ is strongly-convex in $\boldsymbol{x}$. This formulation is equivalent to problem (1) by taking $\boldsymbol{z} = \boldsymbol{x}$, $F_i(\boldsymbol{z}) \coloneqq \nabla f_i(\boldsymbol{x})$ and $\mathcal{Z} \coloneqq \mathcal{X}$.

In many modern machine learning applications, new data keeps coming over time, and the objective function is dynamically evolving. The finite-sum objective function (1) fails to capture such a scenario since it requires all data to be accessible anytime. Therefore, we consider the continual finite-sum variational inequalities, which aim to build a continuously updated model that performs

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

Table 1: We compare the incremental oracle complexity for solving the continual finite-sum variational inequalities when each component function F_i is μ-strongly-monotone. We use the convergence metric $\mathbb{E}[\|\boldsymbol{z}_k - \boldsymbol{z}^*\|^2]$ for all methods.

METHODS	IO COMPLEXITY	REFERENCE
SVRG/SAGA	$\mathcal{O}\left(\left(n^2 + \frac{nL^2}{\mu^2}\right)\log\left(\frac{1}{\epsilon}\right)\right)$	PALANIAPPAN & BACH (2016)
A-SVRG/A-SAGA	$\tilde{\mathcal{O}}\left(\left(n^2 + \frac{n^{\frac{3}{2}}L}{\mu}\right)\log\left(\frac{1}{\epsilon}\right)\right)$	PALANIAPPAN & BACH (2016)
L-SVRE	$\mathcal{O}\left(\left(n^2 + \frac{n^{\frac{3}{2}}L}{\mu}\right)\log\left(\frac{1}{\epsilon}\right)\right)$	ALACAOGLU & MALITSKY (2022)
CSVRG	$\mathcal{O}\Big(\frac{L^2 G\log n}{\mu^{\frac{5}{2}}\epsilon^{\frac{1}{2}}} + \frac{nL^2}{\mu^2} + \frac{nG^{\frac{2}{3}}L^{\frac{2}{3}}\log n}{\mu\epsilon^{\frac{1}{3}}}\Big)$	COROLLARY 5.4
CSVRE	$\mathcal{O}\Big(\frac{LG\log n}{\mu^{\frac{3}{2}}\epsilon^{\frac{1}{2}}} + \frac{nL^2}{\mu^2} + \frac{nG^{\frac{2}{3}}L^{\frac{2}{3}}\log n}{\mu\epsilon^{\frac{1}{3}}}\Big)$	COROLLARY 5.7

equally well on both new and past data (Castro et al., 2018; Rosenfeld & Tsotsos, 2018; Mavrothalassitis et al., 2024; Wang et al., 2024). Formally, we aim to find the solution $\boldsymbol{z}^*$ that satisfies the following prefix-sum VI

$$\frac{1}{i}\sum_{j=1}^{i}\langle F_j(\boldsymbol{z}^*), \boldsymbol{z} - \boldsymbol{z}^*\rangle \geq 0 \quad \text{for all } \boldsymbol{z} \in \mathcal{Z} \tag{2}$$

at each stage $i \in [n]$, where each component operator F_j is μ-strongly-monotone, L-Lipschitz continuous, and its norm is bounded by G over the domain $\mathcal{Z}$. We denote the point $\boldsymbol{z}_i^*$ as the optimal solution of the prefix-sum variational inequalities (2) at stage i.

One straightforward solution for the continual finite-sum VI problem (2) is to apply existing stochastic methods for finite-sum VI on each prefix-sum problem. In particular, directly applying the SVRG/SAGA method (Palaniappan & Bach, 2016) to solve the problem (1) at each stage leads to a total incremental oracle (IO) complexity of $\mathcal{O}\big((n^2 + n^{3/2}\varkappa^2)\log(1/\epsilon)\big)$ where $\varkappa \coloneqq L/\mu$ is the condition number. Similarly, we can apply the loopless stochastic variance reduced extragradient (L-SVRE) (Alacaoglu & Malitsky, 2022) to achieve an improved complexity of $\mathcal{O}\big((n^2 + n^{3/2}\varkappa)\log(1/\epsilon)\big)$. In the continual learning setting where the number of functions n is extremely large and ϵ is mediocre, the IO complexities of both methods are prohibitively expensive.

In this work, we propose an efficient stochastic method called continual stochastic variance reduced gradient (CSVRG) for the continual finite-sum VI problem where each component operator $F_i(\boldsymbol{z})$ is strongly monotone. One of the major computational burdens of the aforementioned methods for the continual optimization problem is due to the evaluation of the full operator at the snapshot variable during each stage, which leads to a $\mathcal{O}(n^2)$ factor in the IO complexity. To alleviate this issue, we update the snapshot variable sparsely across all stages and construct the corresponding full operator at these variables. We show that our method obtains a sequence of ϵ-approximate solutions for the continual finite-sum VI problem with an IO complexity of $\tilde{\mathcal{O}}(L^2 G\mu^{-\frac{5}{2}}\epsilon^{-\frac{1}{2}} + n\varkappa^2 + nG^{\frac{2}{3}}L^{\frac{2}{3}}\mu^{-1}\epsilon^{-\frac{1}{3}})$. Our method is inspired by the CSVRG algorithm for continual finite-sum minimization proposed by Mavrothalassitis et al. (2024); however, the convergence analysis in our setting differs fundamentally from theirs. In particular, our analysis relies primarily on the monotonicity of the underlying operator, whereas the analysis in Mavrothalassitis et al. (2024) crucially depends on the convexity of the objective function—a property that does not hold for the variational inequality problem considered here. Despite addressing the more general continual VI setting, the IO complexity of our CSVRG method matches that of the CSVRG algorithm in Mavrothalassitis et al. (2024).

Furthermore, we propose a new stochastic algorithm called the continual stochastic variance-reduced extragradient (CSVRE) method, which extends CSVRG by incorporating a novel extragradient (EG)

iteration. This step balances historical and current information in updating the auxiliary variable $\boldsymbol{z}_i^{t+1/2}$, a key mechanism for achieving an improved IO complexity in the continual learning setting. The method computes a sequence of ϵ-saddle points for Problem (2) with a reduced IO complexity of $\tilde{\mathcal{O}}(LG\mu^{-\frac{3}{2}}\epsilon^{-\frac{1}{2}} + n\varkappa^2 + nG^{\frac{2}{3}}L^{\frac{2}{3}}\mu^{-1}\epsilon^{-\frac{1}{3}})$. We compare the proposed methods with existing algorithms for the continual finite-sum VI problem in Table 1. The results indicate that our methods attain strictly improved IO complexity relative to the existing baselines when the number of stages n is large, and the target accuracy parameter ϵ is of moderate scale. Finally, we conduct numerical experiments to substantiate the practical effectiveness of the proposed approaches.

Paper Organization In Section 2, we provide a literature review on continual learning and the finite-sum VI problem. In Section 3, we formalize the notations and assumptions of our problem. In Section 4, we propose the CSVRG and CSVRE methods for solving the continual finite-sum VI problem. In Section 5, we provide the theoretical convergence analysis of both CSVRG and CSVRE methods. In Section 6, we conduct numerical experiments to validate the effectiveness of our approaches. Finally, we conclude our work in Section 7.

2 Related Work

In this section, we review related work on continual learning and stochastic algorithms for finite-sum variational inequalities.

2.1 Continual/Incremental Learning

Although machine learning approaches have made significant progress in recent years, they are known to suffer from a phenomenon called catastrophic forgetting, a dramatic decrease in performance when new data or information is seen incrementally (Castro et al., 2018; Goodfellow et al., 2013; Kirkpatrick et al., 2017; McCloskey & Cohen, 1989; Mermillod et al., 2013). Several approaches (Jung et al., 2016; Li & Hoiem, 2017; Rusu et al., 2016; Terekhov et al., 2015; Kirkpatrick et al., 2017; Rebuffi et al., 2017) were introduced to mitigate this issue empirically without giving a theoretical guarantee. Recently, Mavrothalassitis et al. (2024) introduced efficient stochastic algorithms to address the finite-sum minimization problem in the continual learning setting, and established explicit bounds on the incremental first-order oracle (IFO) complexity. While continual learning shares similarities with online learning in updating models sequentially, the two frameworks differ in their fundamental objectives. Continual learning prioritizes long-term knowledge retention and stable performance across a sequence of tasks or stages (Wang et al., 2024), seeking to prevent forgetting previously acquired knowledge. In contrast, online learning focuses on rapid adaptation to newly arriving data, often under adversarial or non-stationary conditions, with performance measured by cumulative regret (Hazan et al., 2016). These differences highlight the distinct challenges and evaluation criteria that characterize each framework.

2.2 Finite-Sum Variational Inequalities

Stochastic first-order algorithms for the finite-sum minimization problem are well-studied in the literature (Allen-Zhu, 2018; Defazio et al., 2014; Fang et al., 2018; Johnson & Zhang, 2013; Lin et al., 2018). It has been shown that solving the convex and nonconvex minimization problems with the stochastic recursive gradient estimator achieves the optimal IFO complexity (Allen-Zhu, 2018; Woodworth & Srebro, 2016; Fang et al., 2018; Wang et al., 2019; Arjevani et al., 2023). For the finite-sum VI problem, stochastic variance-reduced algorithms similarly achieve the best-known IO complexities across various settings (Palaniappan & Bach, 2016; Chavdarova et al., 2019; Carmon et al., 2019; Alacaoglu et al., 2021; Alacaoglu & Malitsky, 2022). Considering the μ-strongly-monotone VI problem, Palaniappan & Bach (2016) leveraged the variance reduction methodology for the finite-sum VI problem (1) and they achieved an IO complexity of $\mathcal{O}((n + \varkappa^2)\log(1/\epsilon))$. Furthermore, they applied the catalyst acceleration framework to their methods and obtained an improved IO complexity of $\tilde{\mathcal{O}}((n + \varkappa\sqrt{n})\log(1/\epsilon))$. Later, the L-SVRE method (Alacaoglu & Malitsky, 2022) incorporated the extragradient iteration into the variance reduction framework and achieved an IO complexity of $\mathcal{O}((n+\varkappa\sqrt{n})\log(1/\epsilon))$, which matches the lower bound for the finite-sum minimax optimization problems (Han et al., 2024). Recently, Chen & Luo (2024) proposed a

stochastic algorithm called RAIN that attains a near-optimal first-order complexity of $\tilde{\mathcal{O}}(\sigma^2\epsilon^{-2})$ for computing ϵ-stationary points of the stochastic minimax optimization problem, measured by the gradient norm of the objective function. Their framework, however, is developed for the unconstrained setting, whereas our approach is designed for constrained VI problem. It is unclear whether their techniques can be extended to the constrained setting (Alacaoglu et al., 2024).

3 Notations and Assumptions

In this section, we formalize the notations and assumptions throughout this paper.

Definition 3.1. We say the operator $F : \mathbb{R}^d \to \mathbb{R}$ is L-Lipschitz continuous for some $L > 0$ if for any $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathbb{R}^d$, it holds that $\|F(\boldsymbol{z}_1) - F(\boldsymbol{z}_2)\| \leq L \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|$.

Definition 3.2. We say the operator $F : \mathbb{R}^d \to \mathbb{R}$ is μ-strongly-monotone for some $\mu > 0$ if for any $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathbb{R}^d$, it holds that $\langle F(\boldsymbol{z}_1) - F(\boldsymbol{z}_2), \boldsymbol{z}_1 - \boldsymbol{z}_2\rangle \geq \mu \|\boldsymbol{z}_1 - \boldsymbol{z}_2\|^2$.

Since the operator is μ-strongly monotone, each prefix-sum VI problem has a unique solution, and our goal is to find a sequence of approximate solutions sufficiently close to the solutions in terms of the weighted Euclidean distance.

Definition 3.3. For the μ-strongly monotone VI problem, a point $\hat{\boldsymbol{z}} \in \mathcal{Z}$ is said to be an ϵ-approximate solution if it satisfies $\mu \|\hat{\boldsymbol{z}} - \boldsymbol{z}^*\|^2 \leq \epsilon$, where $\boldsymbol{z}^*$ is the optimal solution to the problem.

In the remainder of this paper, we always suppose Problem (2) satisfies the following assumptions.

Assumption 3.4. The feasible set $\mathcal{Z} \subseteq \mathbb{R}^d$ in Problem (2) is convex and compact, and its diameter is bounded by D.

Assumption 3.5. For each $i \in [n]$, the function $F_i(\boldsymbol{z})$ is L-Lipschitz continuous, μ-strongly-monotone with $\mu > 0$, and it holds that $\|F_i(\boldsymbol{z})\| \leq G$ for any $\boldsymbol{z} \in \mathcal{Z}$.

4 Methodology

In this section, we propose stochastic algorithms for solving the continual finite-sum VI problem.

4.1 The Algorithm

A simple approach to solving the continual variational inequality problem is to directly apply a stochastic variance reduced algorithm, such as SVRG and L-SVRE (Palaniappan & Bach, 2016; Alacaoglu & Malitsky, 2022) to the prefix-sum problem (2) at each stage i. Specifically, at stage i, we maintain an aggregated snapshot variable $\tilde{\boldsymbol{z}}$ (updated every $\Theta(i)$ iterations) and compute the full prefix-sum operator $\tilde{\nabla}_i = \frac{1}{i}\sum_{j=1}^{i} F_j(\tilde{\boldsymbol{z}})$. Based on this snapshot, the L-SVRE method yields an ϵ-approximate solution with $\mathcal{O}((i + \sqrt{i}\varkappa)\log(1/\epsilon))$ IO calls at stage i. Summing $i = 1$ to n gives a total IO complexity of $\mathcal{O}((n^2 + n^{\frac{3}{2}}\varkappa)\log(1/\epsilon))$, which is dominated by the term n^2 when n is large and the required accuracy is moderate.

To reduce the expensive full operator evaluations required by L-SVRE, we extend the CSVRG method of Mavrothalassitis et al. (2024), originally developed for continual convex minimization, to our continual finite-sum VI problem. Following the standard CSVRG framework (Mavrothalassitis et al., 2024), we set a hyperparameter $\alpha > 0$ and update the aggregated snapshot variable, along with its full prefix-sum operator

$$\tilde{\nabla}_{i-1} \coloneqq \frac{1}{i-1}\sum_{j=1}^{i-1} F_j(\hat{\boldsymbol{z}}_{i-1}), \tag{3}$$

whenever condition $i - \mathrm{prev} \geq \alpha i$ holds, where prev is the stage at which the snapshot was last updated. Building on the operator (3), we construct inexpensive stochastic estimators for the full prefix-sum operator $\sum_{j=1}^{i} F_j(\boldsymbol{z})/i$ with the variance reduction technique. In particular, the stochastic estimator for the full operator takes the form

$$\nabla_i^t = \left(1 - \frac{1}{i}\right)\left(F_{u_t}(\boldsymbol{z}_i^t) - F_{u_t}(\hat{\boldsymbol{z}}_{\mathrm{prev}}) + \tilde{\nabla}_{i-1}\right) + \frac{1}{i}F_i(\boldsymbol{z}_i^t) \tag{4}$$

Algorithm 1: Continual Sparse Learning Method (CSL)

Input: $\hat{\boldsymbol{z}}_0 \in \mathcal{Z}$, prev $\leftarrow 0$, flag $\leftarrow$ false, sequences $\{T_i\}_{i=2}^n$, $\{\gamma_t\}$
$\hat{\boldsymbol{z}}_1 \leftarrow \text{ExtraGradient}(\hat{\boldsymbol{z}}_0)$, $\tilde{\nabla}_1 \leftarrow F_1(\hat{\boldsymbol{z}}_1)$
for $i = 2, \ldots, n$ **do**
  if $i - \text{prev} \geq \alpha \cdot i$ **then**
    $\tilde{\nabla}_{i-1} \leftarrow \frac{1}{i-1} \sum_{j=1}^{i-1} F_j(\hat{\boldsymbol{z}}_{i-1})$
    prev $\leftarrow i-1$
    flag $\leftarrow$ true
  end
  Option I: $\hat{\boldsymbol{z}}_i \leftarrow \text{SVRG}(\hat{\boldsymbol{z}}_{\text{prev}}, \tilde{\nabla}_{i-1}, T_i, \{\gamma_t\}, \hat{\boldsymbol{z}}_{i-1})$
  Option II: $\hat{\boldsymbol{z}}_i \leftarrow \text{SVRE}(\hat{\boldsymbol{z}}_{\text{prev}}, \tilde{\nabla}_{i-1}, T_i, \{\gamma_t\}, \hat{\boldsymbol{z}}_{i-1})$
  Output: $\hat{\boldsymbol{z}}_i$.
  if flag **then**
    $\tilde{\nabla}_i \leftarrow \frac{1}{i} \sum_{j=1}^{i} F_j(\hat{\boldsymbol{z}}_i)$
    prev $\leftarrow i$
    flag $\leftarrow$ false
  else
    $\tilde{\nabla}_i \leftarrow (1 - \frac{1}{i})\tilde{\nabla}_{i-1} + \frac{1}{i} F_i(\hat{\boldsymbol{z}}_{\text{prev}})$
  end
end

Algorithm 2: SVRG$(\hat{\boldsymbol{z}}_{\text{prev}}, \tilde{\nabla}_{i-1}, T_i, \{\gamma_t\}, \hat{\boldsymbol{z}}_{i-1})$

$\boldsymbol{z}_i^0 \leftarrow \hat{\boldsymbol{z}}_{i-1}$
for $t = 0, \ldots, T_i - 1$ **do**
  Select $u_t \sim \text{Unif}(1, \ldots, i-1)$
  $\nabla_i^t \leftarrow \left(1 - \frac{1}{i}\right)\left(h_{u_t}(\boldsymbol{z}_i^t) - h_{u_t}(\hat{\boldsymbol{z}}_{\text{prev}}) + \tilde{\nabla}_{i-1}\right) + \frac{1}{i} h_i(\boldsymbol{z}_i^t)$
  $\boldsymbol{z}_i^{t+1} \leftarrow \Pi_{\mathcal{Z}}(\boldsymbol{z}_i^t - \gamma_t \nabla_i^t)$
end
return $\hat{\boldsymbol{z}}_i \leftarrow \boldsymbol{z}_i^{T_i}$.

where u_t is drawn uniformly from $[i-1]$. We then perform one projected stochastic gradient descent step on $\boldsymbol{z}_i^t$ to obtain the next iterate $\boldsymbol{z}_i^{t+1}$. We refer to the whole procedure as the continual stochastic variance-reduced gradient (CSVRG) method, described in detail in Algorithm 1 under Option I. When $F_j(\boldsymbol{z}) = \nabla f_j(\boldsymbol{x})$ where f_j is a strongly convex function, our CSVRG algorithm coincides with the CSVRG method proposed by Mavrothalassitis et al. (2024). In contrast, if $F_j(\boldsymbol{z}) = (\nabla_x f_j(\boldsymbol{x}, \boldsymbol{y}), -\nabla_y f_j(\boldsymbol{x}, \boldsymbol{y}))$ where f_j is strongly-convex in $\boldsymbol{x}$ and strongly-concave in $\boldsymbol{y}$, our CSVRG method naturally applies to the continual finite-sum minimax optimization problem.

4.2 Acceleration with Extragradient

In this subsection, we introduce a novel extragradient iteration into the continual learning framework (Algorithm 1) to obtain a more efficient stochastic algorithm. The main intuition of the stochastic variance reduced extragradient is to use the full operator evaluated at the snapshot variable to find a mid-point, and then use the stochastic estimator of the full operator at the mid-point to find the next iterate. In particular, we modify the subroutine SVRG method (Algorithm 2) by introducing two auxiliary variables $\bar{\boldsymbol{z}}_i^t$ and $\boldsymbol{z}_i^{t+1/2}$ at each iteration. The variable $\bar{\boldsymbol{z}}_i^t$ is defined as the weighted

Algorithm 3: SVRE($\hat{\boldsymbol{z}}_{\rm prev}, \tilde{\nabla}_{i-1}, T_i, \{\gamma_t\}, \hat{\boldsymbol{z}}_{i-1}$)

$\boldsymbol{z}_i^0 \leftarrow \hat{\boldsymbol{z}}_{i-1}$
for $t = 0, \ldots, T_i - 1$ **do**
  $\bar{\boldsymbol{z}}_i^t \leftarrow \eta_t \boldsymbol{z}_i^t + (1-\eta_t)\hat{\boldsymbol{z}}_{\rm prev}$
  $\boldsymbol{z}_i^{t+1/2} \leftarrow \Pi_{\mathcal{Z}}\left(\bar{\boldsymbol{z}}_i^t - \gamma_t\left(\left(1-\frac{1}{i}\right)\tilde{\nabla}_{i-1} + \frac{1}{i}h_i(\hat{\boldsymbol{z}}_{\rm prev})\right)\right)$
  Select $u_t \sim \mathrm{Unif}(1, \ldots, i-1)$
  $\nabla_i^t \leftarrow \left(1-\frac{1}{i}\right)\left(h_{u_t}\left(\boldsymbol{z}_i^{t+1/2}\right) - h_{u_t}(\hat{\boldsymbol{z}}_{\rm prev}) + \tilde{\nabla}_{i-1}\right) + \frac{1}{i}h_i\left(\boldsymbol{z}_i^{t+1/2}\right)$
  $\boldsymbol{z}_i^{t+1} \leftarrow \Pi_{\mathcal{Z}}(\bar{\boldsymbol{z}}_i^t - \gamma_t \nabla_i^t)$
end
return: $\hat{\boldsymbol{z}}_i \leftarrow \boldsymbol{z}_i^{T_i}$.

average of $\boldsymbol{z}_i^t$ and $\hat{\boldsymbol{z}}_{\rm prev}$, given by

$$\bar{\boldsymbol{z}}_i^t = \eta_t \boldsymbol{z}_i^t + (1-\eta_t)\hat{\boldsymbol{z}}_{\rm prev},$$

where $\eta_t \in [0, 1)$ is a weighting parameter. Then we find the mid-point $\boldsymbol{z}_i^{t+1/2}$ with the following novel extragradient iteration:

$$\boldsymbol{z}_i^{t+1/2} = \Pi_{\mathcal{Z}}\Big(\bar{\boldsymbol{z}}_i^t - \gamma_t\Big(\Big(1-\frac{1}{i}\Big)\tilde{\nabla}_{i-1} + \frac{1}{i}F_i(\hat{\boldsymbol{z}}_{\rm prev})\Big)\Big). \tag{5}$$

We remark that $\sum_{j=1}^{i} F_j(\hat{\boldsymbol{z}}_{\rm prev})/i = (1-1/i)\,\tilde{\nabla}_{i-1} + F_i(\hat{\boldsymbol{z}}_{\rm prev})/i$ represents the full prefix-sum operator evaluated at $\hat{\boldsymbol{z}}_{\rm prev}$. Recall that the classical extragradient method computes the intermediate point $\boldsymbol{z}_i^{t+1/2}$ solely based on the aggregated operator $\tilde{\nabla}_{i-1}$. By contrast, update (5) incorporates a carefully weighted combination of both the historical information $\tilde{\nabla}_{i-1}$ and the operator $F_i(\hat{\boldsymbol{z}}_{\rm prev})$ at current stage. This refinement is crucial for obtaining an improved IO complexity bound in the continual learning setting. Building upon this idea, we then construct an efficient stochastic estimator of the full prefix-sum operator $\sum_{j=1}^{i} F_j(\boldsymbol{z})/i$ at the midpoint $\boldsymbol{z}_i^{t+1/2}$ in a manner analogous to formula (4), such that

$$\nabla_i^t = \Big(1-\frac{1}{i}\Big)\left(h_{u_t}\left(\boldsymbol{z}_i^{t+1/2}\right) - h_{u_t}(\hat{\boldsymbol{z}}_{\rm prev}) + \tilde{\nabla}_{i-1}\right) + \frac{1}{i}h_i\left(\boldsymbol{z}_i^{t+1/2}\right). \tag{6}$$

Finally, we perform one step of stochastic gradient descent to obtain the next iterate $\boldsymbol{z}_i^{t+1}$. We call the whole procedure continual stochastic variance-reduced extragradient (CSVRE) and present its details in Algorithm 1 with Option II. In the next section, we will show that the CSVRE method obtains an improved IO complexity compared with the CSVRG method.

5 Theoretical Analysis

In this section, we establish the IO complexity of the proposed methods in Section 4 for solving the continual finite-sum VI problem where each operator F_i is strongly-monotone. We start our analysis with the following two lemmas, which show that the stochastic estimators of the full prefix-sum operators defined in (4) and (6) are unbiased and have controlled variance.

Lemma 5.1. *Let u_t be some index uniformly sampled from $[i-1]$. Then for any $\boldsymbol{z} \in \mathcal{Z}$, the stochastic estimator*

$$\bar{\nabla} = \left(1-\frac{1}{i}\right)\left(F_{u_t}(\boldsymbol{z}) - F_{u_t}(\hat{\boldsymbol{z}}_{\rm prev}) + \tilde{\nabla}_{i-1}\right) + \frac{1}{i}F_i(\boldsymbol{z})$$

that satisfies $\mathbb{E}\left[\bar{\nabla}\right] = \sum_{j=1}^{i} F_j(\boldsymbol{z})/i$.

We provide the upper bound for the variance of the stochastic estimator as follows.

Lemma 5.2. *For any $\boldsymbol{z} \in \mathcal{Z}$, then the stochastic estimator $\bar{\nabla}$ defined in Lemma 5.1 satisfies*

$$\mathbb{E}\left[\left\|\bar{\nabla} - \frac{1}{i}\sum_{j=1}^{i} F_j(\boldsymbol{z})\right\|^2\right] \leq 2L^2\mathbb{E}\left[\|\boldsymbol{z} - \boldsymbol{z}_i^*\|^2\right] + \frac{4G^2L^2\alpha^2}{\mu^2} + 4L^2\mathbb{E}\left[\|\hat{\boldsymbol{z}}_{\text{prev}} - \boldsymbol{z}_{\text{prev}}^*\|^2\right],$$

where α is some positive hyperparameter, $\boldsymbol{z}_i^*$ and $\boldsymbol{z}_{\text{prev}}^*$ are the solutions of the VI problems at stages i and prev*, respectively.*

Lemma 5.2 implies that the variance of the stochastic estimator decreases when the variable $\boldsymbol{z}$ is close to the optimum $\boldsymbol{z}_i^*$. Based on this result, we can perform the convergence analysis for each stage and obtain a sequence of ϵ-approximate solutions under appropriate choices of hyperparameters for the CSVRG method.

Theorem 5.3. *Under Assumption 3.4 and Assumption 3.5, running the CSVRG method (Algorithm 1 with Option I) with $\alpha = \mu\epsilon^{\frac{1}{3}}G^{-\frac{2}{3}}L^{-\frac{2}{3}}$ and the subroutine SVRG method (Algorithm 2) with*

$$\gamma_t = \frac{2}{\mu(t+\beta)},\ \beta = \frac{6L^2}{\mu^2},\ T_i = \mathcal{O}\left(\frac{L^2G}{\mu^{\frac{5}{2}}i\epsilon^{\frac{1}{2}}} + \frac{L^2}{\mu^2} + \frac{G^{\frac{2}{3}}L^{\frac{2}{3}}}{\mu\epsilon^{\frac{1}{3}}}\right),$$

then the output $\hat{\boldsymbol{z}}_i$ satisfies $\mathbb{E}\left[\mu\|\hat{\boldsymbol{z}}_i - \boldsymbol{z}_i^*\|^2\right] \leq \epsilon$ for each $i \in [n]$.

Theorem 5.3 implies the following IO complexity of Algorithm 1 with Option I.

Corollary 5.4. *Under Assumption 3.4 and Assumption 3.5, the CSVRG method (Algorithm 1 with Option I) requires at most*

$$\mathcal{O}\left(\frac{L^2G\log n}{\mu^{\frac{5}{2}}\epsilon^{\frac{1}{2}}} + \frac{nL^2}{\mu^2} + \frac{nG^{\frac{2}{3}}L^{\frac{2}{3}}\log n}{\mu\epsilon^{\frac{1}{3}}}\right)$$

IO calls to obtain a sequence of ϵ-saddle points for solving the Problem (2).

Remark 5.5. We would like to highlight that our analysis of CSVRG substantially differs from that of Mavrothalassitis et al. (2024). Our analysis is mainly based on controlling the Euclidean distance between the current iterate and the optimal solution. In contrast, Mavrothalassitis et al. (2024) derives a recursion in terms of the function-value gap. It is worth noting that the upper bound of the function value gap can only work for the analysis of the minimization problem, which cannot be extended to the more general VI setting.

The extragradient method is known to achieve improved convergence rates for the finite-sum VI problem. A natural question is whether our novel extragradient step, inspired by the classical variant, can further reduce the IO complexity in the continual learning setting. Specifically, we show that the CSVRE method attains a sequence of ϵ-approximate solutions under the following hyperparameter choices.

Theorem 5.6. *Under Assumption 3.4 and Assumption 3.5, running the CSVRE method (Algorithm 1 with Option II) with $\alpha = \mu\epsilon^{\frac{1}{3}}G^{-\frac{2}{3}}L^{-\frac{2}{3}}$ and the subroutine SVRE method (Algorithm 3) with*

$$\gamma_t = \frac{2}{\mu(t+\beta)}, \quad \eta_t = 1 - 2\gamma_t^2L^2, \quad \beta = \frac{8L}{\mu},$$
$$T_i = \mathcal{O}\left(\frac{LG}{\mu^{\frac{3}{2}}i\epsilon^{\frac{1}{2}}} + \frac{L^2}{\mu^2} + \frac{G^{\frac{2}{3}}L^{\frac{2}{3}}}{\mu\epsilon^{\frac{1}{3}}}\right),$$

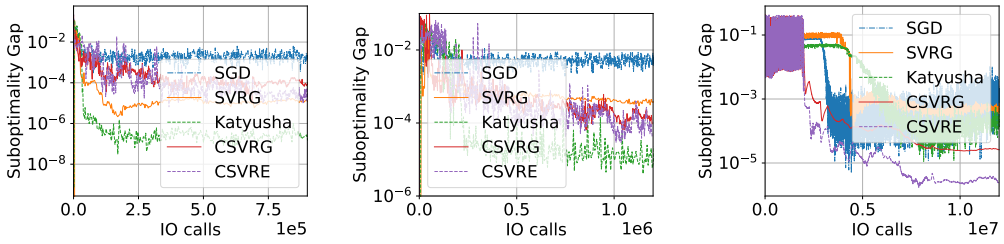
then the output $\hat{\boldsymbol{z}}_i$ satisfies $\mathbb{E}\left[\mu\|\hat{\boldsymbol{z}}_i - \boldsymbol{z}_i^*\|^2\right] \leq \epsilon$ for each $i \in [n]$.

Theorem 5.6 implies the following IO complexity of Algorithm 1 with Option II.

Corollary 5.7. *Under Assumption 3.4 and Assumption 3.5, the CSVRE method (Algorithm 1 with Option II) requires at most*

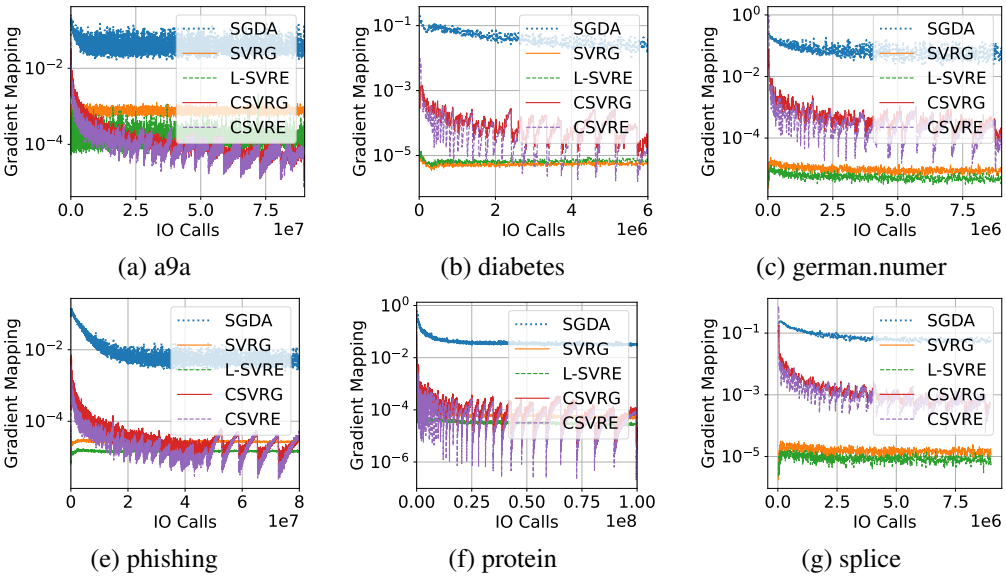
$$\mathcal{O}\left(\frac{LG\log n}{\mu^{\frac{3}{2}}\epsilon^{\frac{1}{2}}} + \frac{nL^2}{\mu^2} + \frac{nG^{\frac{2}{3}}L^{\frac{2}{3}}\log n}{\mu\epsilon^{\frac{1}{3}}}\right)$$

IO calls to obtain a sequence of ϵ-saddle points for solving the Problem (2).



(a) diabetes (b) german.numer (c) cod-rna

Figure 1: Suboptimality gap vs. the number of IO calls for the ridge regression problem.



(a) a9a (b) diabetes (c) german.numer

(e) phishing (f) protein (g) splice

Figure 2: Gradient mapping vs. the number of IO calls for the robust linear regression problem.

6 Experiments

In this section, we conduct numerical experiments to validate the effectiveness of the proposed methods on the continual finite-sum VI problem. Recall that both minimization and minimax optimization can be formulated as special cases of VI problems. We consider ridge regression for continual minimization, as well as robust linear regression for continual minimax optimization. We leave the experiments on the fairness-aware machine learning problem in Appendix E.

6.1 Ridge Regression Problem

We consider the prefix-sum ridge regression problem

$$\min_{\|\boldsymbol{x}\| \leq R} g_i(x) \coloneqq \frac{1}{i} \sum_{j=1}^{i} f_j(\boldsymbol{x}),$$

where each component function is defined as

$$f_j(\boldsymbol{x}) = (\boldsymbol{x}^\top a_j - b_j)^2 + \lambda \|\boldsymbol{x}\|^2 .$$

The random variable $\boldsymbol{x} \in \mathbb{R}^d$ is the model weight, a_j is the data feature, $b_j \in \mathbb{R}$ is the corresponding label.

We compare our methods CSVRG and CSVRE with baseline methods including SGD (Nemirovski, 2004; Bottou, 2010), SVRG (Johnson & Zhang, 2013), and Katyusha (Allen-Zhu, 2018). We test all

algorithms on three real-world datasets ("diabetes", "german.numer", "cod-rna") from the LIBSVM repository (Chang & Lin, 2011). We set hyperparameters $\lambda = 10^{-3}$, $R = 100.0$ across all the datasets. For both CSVRG and CSVRE methods, we set the sparsity parameter $\alpha = 0.1$. We choose $T_i = 400$ for the CSVRG method and $T_i = 300$ for the CSVRE method so that the per-stage IO calls of the two methods match. At each stage, we choose the outer iteration to be 10 and the inner iterations to be 200 for the SVRG and L-SVRE methods. The step size for each algorithm is tuned from the set $\{10^{-5}, 3 \times 10^{-5}, \ldots, 0.03, 0.1\}$ for each algorithm. Following the setup of Mavrothalassitis et al. (2024), we reveal a new data point at each stage $i \in [n]$.

The experimental results are presented in Figure 1. Our methods may underperform relative to the baselines on datasets with small sample sizes (e.g., "diabete" with $n = 750$, "german.numer" with $n = 1,000$). However, they demonstrate clear and consistent superiority on larger datasets (e.g., "cod-rna" with $n = 59,535$). This empirical trend aligns well with our theoretical results.

6.2 Robust Linear Regression

We consider the prefix-sum robust linear regression problem

$$\min_{\|\boldsymbol{x}\| \le R_x} \max_{\|\boldsymbol{y}\| \le R_y} g_i(\boldsymbol{x}, \boldsymbol{y}) \coloneqq \frac{1}{i} \sum_{j=1}^{i} f_j(\boldsymbol{x}, \boldsymbol{y}),$$

where each component function is defined as

$$f_i(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}(\boldsymbol{x}^\top(\boldsymbol{a}_j + \boldsymbol{y}) - b_j)^2 + \lambda \|\boldsymbol{x}\|^2 - \beta \|\boldsymbol{y}\|^2 .$$

Here the random variable $\boldsymbol{x} \in \mathbb{R}^d$ is the model weight, $\boldsymbol{y} \in \mathbb{R}^d$ describes the noise, $\boldsymbol{a}_j \in \mathbb{R}^d$ is the data feature, and $b_j \in \mathbb{R}$ is the corresponding label.

We compare our methods CSVRG and CSVRE with baseline methods including SGDA (Nemirovski, 2004; Korpelevich, 1976), SVRG (Palaniappan & Bach, 2016), and L-SVRE (Alacaoglu & Malitsky, 2022). We test all algorithms on six real-world datasets ("a9a", "diabetes", "german.numer", "phishing", "protein", "splice") from the LIBSVM repository (Chang & Lin, 2011). We set hyperparameters $\lambda = 2.0$, $\beta = 2.0$, $R_x = 1.0$ and $R_y = 0.1$ across all the datasets. For both CSVRG and CSVRE methods, we set the sparsity parameter $\alpha = 0.1$. All other hyperparameters, including the stepsize and the number of iterations, are chosen following the same configuration used in the ridge regression experiments.

The experimental results are presented in Figure 2. Similar to the results of ridge regression problem, we observe that our methods may underperform relative to the baselines on datasets with small sample sizes (e.g., "diabete" with $n = 750$, "german.numer" with $n = 1,000$, "splice" with $n = 1,000$). In contrast, they demonstrate clear and consistent superiority on larger datasets (e.g., "a9a" with $n = 13,000$, "phishing" with $n = 12,000$, "protein" with $n = 12,000$). These empirical findings are consistent with our theory, since the performance of our methods improves as the dataset size increases.

7 Conclusion

In this paper, we propose an efficient stochastic algorithm, termed CSVRG, for solving continual finite-sum variational inequality problems. Our theoretical analysis shows that the proposed method generates a sequence of approximate solutions using significantly fewer incremental oracle (IO) calls than existing approaches. Moreover, we introduce a novel extragradient-based iteration that further sharpens the IO complexity compared with CSVRG. We validate the effectiveness of the proposed methods through numerical experiments on several real-world tasks.

For future work, it would be interesting to further investigate the gap between the upper and lower complexity bounds for the continual finite-sum variational inequality problem. Another promising direction is the development of efficient stochastic algorithms for continual finite-sum nonconvex optimization problems.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the writing / implement methods / design or provide feedback on research methodology or experiments / interpret results / formulate mathematical claims. We have not used generative AI tools for developing theoretical models or conceptual frameworks / providing critical ingredients for proving mathematical claims / support qualitative and thematic data analysis / provide critical ingredients for proving mathematical claims / assist in the writing of proofs / propose or refine hypotheses and generate synthetic data sets / assist with translation / clean and reformat dataset are not applicable to this work. Additionally, we used generative AI tools for suggest experimental parameters / create or edit software code / summarize or analyse existing literature / discover research topics or identify gaps. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by us, LLM-recommended references are double-checked by us. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Ahmet Alacaoglu and Yura Malitsky. Stochastic variance reduction for variational inequality methods. In *Conference on Learning Theory*, pp. 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.

Ahmet Alacaoglu, Donghwan Kim, and Stephen J Wright. Revisiting inexact fixed-point iterations for min-max problems: Stochasticity and structured nonconvexity. *arXiv preprint arXiv:2402.05071*, 2024.

Zeyuan Allen-Zhu. Katyusha: The first direct acceleration of stochastic gradient methods. *Journal of Machine Learning Research*, 18(221):1–51, 2018.

Yossi Arjevani, Yair Carmon, John Duchi, Dylan Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1):165–214, 2023.

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*, pp. 177–186. Springer, 2010.

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

Francisco M Castro, Manuel J Marín-Jiménez, Nicolás Guil, Cordelia Schmid, and Karteek Alahari. End-to-end incremental learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 233–248, 2018.

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 and Luo Luo. Near-optimal algorithms for making the gradient small in stochastic minimax optimization. *Journal of Machine Learning Research*, 25(387):1–44, 2024.

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, 2014.

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.

Ian J Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks. *arXiv preprint arXiv:1312.6211*, 2013.

Yuze Han, Guangzeng Xie, and Zhihua Zhang. Lower complexity bounds of finite-sum optimization problems: The results and construction. *Journal of Machine Learning Research*, 25(2):1–86, 2024.

Elad Hazan et al. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

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

Heechul Jung, Jeongwoo Ju, Minju Jung, and Junmo Kim. Less-forgetting learning in deep neural networks. *arXiv preprint arXiv:1607.00122*, 2016.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.

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

Tai Le Quy, Arjun Roy, Vasileios Iosifidis, Wenbin Zhang, and Eirini Ntoutsi. A survey on datasets for fairness-aware machine learning. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 12(3):e1452, 2022.

Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947, 2017.

Hongzhou Lin, Julien Mairal, and Zaid Harchaoui. Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research*, 18(212):1–54, 2018.

Chengchang Liu and Luo Luo. Quasi-newton methods for saddle point problems. *Advances in Neural Information Processing Systems*, 35:3975–3987, 2022.

Ioannis Mavrothalassitis, Stratis Skoulakis, Leello Tadesse Dadi, and Volkan Cevher. Efficient continual finite-sum minimization. In *The Twelfth International Conference on Learning Representations*, 2024.

Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pp. 109–165. Elsevier, 1989.

Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: Investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, 4:504, 2013.

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.

Yurii Nesterov et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

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

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.

Amir Rosenfeld and John Tsotsos. Incremental learning through deep adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(3):651–663, 2018.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.

Alexander V Terekhov, Guglielmo Montone, and J Kevin O'Regan. Knowledge transfer in deep block-modular neural networks. In *Biomimetic and Biohybrid Systems: 4th International Conference, Living Machines 2015, Barcelona, Spain, July 28-31, 2015, Proceedings 4*, pp. 268–279. Springer, 2015.

Liyuan Wang, Xingxing Zhang, Hang Su, and Jun Zhu. A comprehensive survey of continual learning: Theory, method and application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383, 2024.

Zhe Wang, Kaiyi Ji, Yi Zhou, Yingbin Liang, and Vahid Tarokh. Spiderboost and momentum: Faster variance reduction algorithms. *Advances in Neural Information Processing Systems*, 32, 2019.

Blake Woodworth and Nati Srebro. Tight complexity bounds for optimizing composite objectives. *Advances in Neural Information Processing Systems*, 29, 2016.

APPENDIX

The appendix is organized as follows. Appendix A provides a discussion of the lower bound for solving the continual finite-sum variational inequality problem. Appendix B introduces several key lemmas essential for the convergence analysis of the proposed methods for solving the continual minimax optimization problem. Appendix C provides the convergence analysis and total time complexity of the CSVRG method for solving the continual problem in the strongly-convex-strongly-concave setting. Appendix D provides the convergence analysis of the CSVRE method and demonstrates its improved time complexity by incorporating the extragradient iteration. Appendix E conducts additional experiments on the prefix-sum fairness-aware machine learning problem.

A DISCUSSION ABOUT LOWER BOUND

We now study lower bounds for solving the continual finite-sum variational inequality (VI) problem using incremental algorithms. Since finite-sum minimization can be viewed as a special case of the VI problem, the lower bounds established for continual finite-sum minimization in Mavrothalassitis et al. (2024) also apply directly to the continual finite-sum VI problem considered in this work. In particular, we rely on the lower bounds established by Mavrothalassitis et al. (2024), which we restate below for completeness.

Theorem A.1 (Mavrothalassitis et al. (2024, Theorem 3))**.** *For any $\alpha > 0$, there is no stochastic first-order method for the continual finite-sum minimization problem with $\mathcal{O}(n^{2-\alpha}\log(1/\epsilon))$ IFO complexity. Moreover, for any $\alpha < 1/4$, there is no stochastic first-order method for the continual finite-sum minimization problem with $\mathcal{O}(n\epsilon^{-\alpha})$ IFO complexity.*

Theorem A.1 implies that any stochastic algorithm that finds an ϵ-approximate solution to the corresponding continual VI problem must make at least $\Omega(n\epsilon^{-1/4})$ IO calls. Recall from Corollary 5.7 that the IO complexity of our CSVRE method for the continual finite-sum VI problem is dominated by $\mathcal{O}(n\epsilon^{-1/3})$, which is close to the lower bound $\Omega(n\epsilon^{-1/4})$. Closing the remaining gap of order $\epsilon^{-1/12}$ for VI problems remains an open question.

We further note that directly applying L-SVRE to the continual finite-sum VI problem results in an IO complexity of $\mathcal{O}(n^2\log(1/\epsilon))$. On the other hand, Theorem A.1 implies that, for any $\alpha > 0$, no IO algorithm for the continual finite-sum VI problem can achieve an IO complexity better than $\mathcal{O}(n^{2-\alpha}\log(1/\epsilon))$. Consequently, the trade-off between the dependence on n and ϵ is fundamental in the VI setting. By relaxing the logarithmic dependence on n, our proposed CSVRE method improves the dependence on n and attains an IO complexity of $\mathcal{O}(n\epsilon^{-1/3})$ for the continual finite-sum VI problem. In contrast, the aforementioned lower bound indicates that it is impossible to simultaneously maintain a logarithmic dependence on $1/\epsilon$ while still achieving a linear dependence on n in this class of VI problems.

B ESTABLISHED RESULTS

Firstly, we present some useful tools for the analysis of the constrained optimization.

Lemma B.1 (Nesterov et al. (2018))**.** *Given a convex and compact set $\mathcal{C} \subseteq \mathbb{R}^d$, and any $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d$, we have*

$$\|\Pi_{\mathcal{C}}(\boldsymbol{u}) - \Pi_{\mathcal{C}}(\boldsymbol{v})\| \leq \|\boldsymbol{u} - \boldsymbol{v}\|. \tag{7}$$

Lemma B.2 (Nesterov et al. (2018))**.** *Given a convex and compact set $\mathcal{C} \subseteq \mathbb{R}^d$, for any $\boldsymbol{u} \in \mathbb{R}^d$ and $\boldsymbol{v} \in \mathcal{C}$, we have*

$$\langle \Pi_{\mathcal{C}}(\boldsymbol{u}) - \boldsymbol{u}, \Pi_{\mathcal{C}}(\boldsymbol{u}) - \boldsymbol{v} \rangle \leq 0. \tag{8}$$

C CONVERGENCE ANALYSIS OF CSVRG

In this section, we present the convergence analysis of the CSVRG method for solving the continual finite-sum VI problem.

C.1 SUPPORTING LEMMAS

Firstly, we show that the optimal solutions of prefix-sum objective functions $g_i(\boldsymbol{x}, \boldsymbol{y})$ and $g_j(\boldsymbol{x}, \boldsymbol{y})$ are similar when i and j are close to each other.

Lemma C.1. *For all $i \in [n-1]$ and $j \in [n-i]$, it holds that*

$$\sqrt{\mu \left\| \boldsymbol{z}_{i+j}^* - \boldsymbol{z}_i^* \right\|^2} \leq \frac{jG}{\sqrt{\mu}(i+j)}.$$

Proof. Assumption 3.5 implies that

$$\begin{aligned}
\mu \left\| \boldsymbol{z}_{i+j}^* - \boldsymbol{z}_i^* \right\|^2 \leq & \frac{1}{i+j} \sum_{k=1}^{i+j} \langle F_k(\boldsymbol{z}_{i+j}^*) - F_k(\boldsymbol{z}_i^*), \boldsymbol{z}_{i+j}^* - \boldsymbol{z}_i^* \rangle \\
\leq & \frac{1}{i+j} \sum_{k=1}^{i+j} \langle F_k(\boldsymbol{z}_i^*), \boldsymbol{z}_i^* - \boldsymbol{z}_{i+j}^* \rangle \\
= & \frac{1}{i+j} \sum_{k=1}^{i} \langle F_k(\boldsymbol{z}_i^*), \boldsymbol{z}_i^* - \boldsymbol{z}_{i+j}^* \rangle + \frac{1}{i+j} \sum_{k=i+1}^{i+j} \langle F_k(\boldsymbol{z}_i^*), \boldsymbol{z}_i^* - \boldsymbol{z}_{i+j}^* \rangle \\
\leq & \frac{1}{i+j} \sum_{k=i+1}^{i+j} \langle F_k(\boldsymbol{z}_i^*), \boldsymbol{z}_i^* - \boldsymbol{z}_{i+j}^* \rangle \\
\leq & \frac{jG}{i+j} \left\| \boldsymbol{z}_i^* - \boldsymbol{z}_{i+j}^* \right\|,
\end{aligned}$$

where the second and third inequalities are due to the optimality of $\boldsymbol{z}_{i+j}^*$ and $\boldsymbol{z}_i^*$, respectively. The last inequality follows from Cauchy–Schwarz inequality and Assumption 3.5. Rearranging the terms, we can show that

$$\sqrt{\mu \left\| \boldsymbol{z}_{i+j}^* - \boldsymbol{z}_i^* \right\|^2} \leq \frac{jG}{\sqrt{\mu}(i+j)}.$$

□

The above lemma implies the following result.

Lemma C.2. *Let $\hat{\boldsymbol{z}}_j \in \mathcal{Z}$ be the approximate solution at stage $j \in [n]$. Then for all $i \in \{j+1, \ldots, n\}$,*

$$\mu \left\| \hat{\boldsymbol{z}}_j - \boldsymbol{z}_i^* \right\|^2 \leq \frac{2(G(i-j))^2}{\mu i^2} + 2\mu \left\| \hat{\boldsymbol{z}}_j - \boldsymbol{z}_j^* \right\|^2.$$

Proof. Applying Young's Inequality, we obtain

$$\begin{aligned}
& \mu \left\| \hat{\boldsymbol{z}}_j - \boldsymbol{z}_i^* \right\|^2 \\
\leq & 2\mu \left(\left\| \boldsymbol{z}_j^* - \boldsymbol{z}_i^* \right\|^2 + \left\| \boldsymbol{z}_j^* - \hat{\boldsymbol{z}}_j \right\|^2 \right) \\
\leq & \frac{2(G(i-j))^2}{\mu i^2} + 2\mu \left\| \hat{\boldsymbol{z}}_j - \boldsymbol{z}_j^* \right\|^2.
\end{aligned}$$

The last inequality follows from Lemma C.1. □

Before studying the stochastic estimator of CSVRG, we first show some properties about the gradient estimator $\tilde{\nabla}$.

Lemma C.3. *Let* prev *be the stage index at which a full prefix-sum gradient is computed; then*

$$\tilde{\nabla}_{i-1} = \frac{1}{i-1}\sum_{j=1}^{i-1} F_j(\hat{\boldsymbol{z}}_{\text{prev}}). \tag{9}$$

Proof. We will prove the lemma by induction. Note that after stage $i = 1$, we set $\tilde{\nabla}_1 = F_1(\hat{\boldsymbol{z}}_1)$.

Induction Hypothesis Let Eq. (9) holds for stage $i - 1$.

Induction Step We will prove Eq. (9) holds for stage i. We consider three cases as below.

(a). $i - \text{prev} \geq \alpha i$. In this case, we have

$$\tilde{\nabla}_{i-1} = \frac{1}{i-1}\sum_{j=1}^{i-1} F_j(\hat{\boldsymbol{z}}_{i-1}),$$

In addition, we set $\text{prev} = i - 1$, therefore Eq. (9) holds at the current iteration.

(b). $(i-1) - \text{prev} \geq \alpha(i-1)$. It means that prev was updated in the previous stage. In this case, we infer from Algorithm 1 such that

$$\tilde{\nabla}_{i-1} = \frac{1}{i-1}\sum_{j=1}^{i-1} F_j(\hat{\boldsymbol{z}}_{i-1}).$$

Since we set $\text{prev} = i - 1$ in the last stage, Eq. (9) holds.

(c). $i - \text{prev} < \alpha i$ and $(i-1) - \text{prev} < \alpha(i-1)$. From the inductive hypothesis, we have

$$\tilde{\nabla}_{i-2} = \frac{1}{i-2}\sum_{j=1}^{i-2} F_j(\hat{\boldsymbol{z}}_{\text{prev}}). \tag{10}$$

Consequently, we reach step 16 of Algorithm 1, and it follows that

$$\begin{aligned}\tilde{\nabla}_{i-1} = \left(1 - \frac{1}{i-1}\right)\tilde{\nabla}_{i-2} + \frac{1}{i-1}F_{i-1}(\hat{\boldsymbol{z}}_{\text{prev}}) = \quad & \frac{1}{i-1}\sum_{j=1}^{i-2} F_j(\hat{\boldsymbol{z}}_{\text{prev}}) + \frac{1}{i-1}F_{i-1}(\hat{\boldsymbol{z}}_{\text{prev}}) \\ = & \frac{1}{i-1}\sum_{j=1}^{i-1} F_j(\hat{\boldsymbol{z}}_{\text{prev}}).\end{aligned}$$

□

Now we can show the stochastic estimator $\bar{\nabla}$ is an unbiased estimator of the full prefix-sum operator, and it has bounded variance.

C.2 Proof of Lemma 5.1

Proof. According to the definition of $\bar{\nabla}$, one has

$$\begin{aligned}\mathbb{E}[\bar{\nabla}] &= \left(1 - \frac{1}{i}\right)\left(\mathbb{E}\left[F_{u_t}(\boldsymbol{z}) - F_{u_t}(\hat{\boldsymbol{z}}_{\text{prev}})\right] + \tilde{\nabla}_{i-1}\right) + \frac{1}{i}F_i(\boldsymbol{z}) \\ &= \frac{i-1}{i}\left(\frac{1}{i-1}\sum_{j=1}^{i-1}\left(F_j(\boldsymbol{z}) - F_j(\hat{\boldsymbol{z}}_{\text{prev}})\right) + \tilde{\nabla}_{i-1}\right) + \frac{1}{i}F_i(\boldsymbol{z}) \\ &= \frac{1}{i}\sum_{j=1}^{i-1} F_j(\boldsymbol{z}) + \frac{1}{i}F_i(\boldsymbol{z}) \\ &= \frac{1}{i}\sum_{j=1}^{i} F_j(\boldsymbol{z}),\end{aligned}$$

where we use the result of Lemma C.3 for the third equality. □

C.3 Proof of Lemma 5.2

Proof. By the definition of $\bar{\nabla}$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\bar{\nabla}-\frac{1}{i}\sum_{j=1}^{i}F_j(\boldsymbol{z})\right\|^2\right]\\
=&\left(1-\frac{1}{i}\right)^2\mathbb{E}\left[\left\|F_{u_t}(\boldsymbol{z})-F_{u_t}(\hat{\boldsymbol{z}}_{\mathrm{prev}})+\tilde{\nabla}_{i-1}-\frac{1}{i-1}\sum_{j=1}^{i-1}F_j(\boldsymbol{z})\right\|^2\right]\\
\leq&\left(1-\frac{1}{i}\right)^2\mathbb{E}\left[\left\|(F_{u_t}(\boldsymbol{z})-F_{u_t}(\hat{\boldsymbol{z}}_{\mathrm{prev}}))\right\|^2\right]\\
\leq&\left(1-\frac{1}{i}\right)^2L^2\mathbb{E}\left[\left\|\boldsymbol{z}-\hat{\boldsymbol{z}}_{\mathrm{prev}}\right\|^2\right]\\
\leq&\left(1-\frac{1}{i}\right)^2 2L^2\mathbb{E}\left[\left\|\boldsymbol{z}-\boldsymbol{z}_i^*\right\|^2+\left\|\boldsymbol{z}_i^*-\hat{\boldsymbol{z}}_{\mathrm{prev}}\right\|^2\right].
\end{aligned}
$$

The first inequality follows from $\mathbb{E}[\|\boldsymbol{x}-\mathbb{E}[\boldsymbol{x}]\|^2]\leq\mathbb{E}[\|\boldsymbol{x}\|^2]$ for any random variable $\boldsymbol{x}$. The second inequality is derived using Assumption 3.5. Lemma C.2 implies that

$$
\begin{aligned}
&\mathbb{E}\left[\left\|\hat{\boldsymbol{z}}_{\mathrm{prev}}-\boldsymbol{z}_i^*\right\|^2\right]\\
=&\frac{1}{\mu}\mathbb{E}\left[\mu\left\|\hat{\boldsymbol{z}}_{\mathrm{prev}}-\boldsymbol{z}_i^*\right\|^2\right]\\
\leq&\frac{1}{\mu}\mathbb{E}\left[\frac{2(G(i-\mathrm{prev}))^2}{\mu i^2}+2\mu\left\|\hat{\boldsymbol{z}}_{\mathrm{prev}}-\boldsymbol{z}_{\mathrm{prev}}^*\right\|^2\right]\\
\leq&\frac{2G^2\alpha^2}{\mu^2}+\frac{2}{\mu}\mathbb{E}\left[\mu\left\|\hat{\boldsymbol{z}}_{\mathrm{prev}}-\boldsymbol{z}_{\mathrm{prev}}^*\right\|^2\right].
\end{aligned}
$$

The second inequality is due to the condition for full prefix-sum operator computation in Algorithm 1, then one has

$$
\mathbb{E}\left[\left\|\bar{\nabla}-\frac{1}{i}\sum_{j=1}^{i}F_j(\boldsymbol{z})\right\|^2\right]\leq 2L^2\mathbb{E}\left[\left\|\boldsymbol{z}-\boldsymbol{z}_i^*\right\|^2\right]+\frac{4G^2L^2\alpha^2}{\mu^2}+\frac{4L^2}{\mu}\mathbb{E}\left[\mu\left\|\hat{\boldsymbol{z}}_{\mathrm{prev}}-\boldsymbol{z}_{\mathrm{prev}}^*\right\|^2\right].
$$

□

Now we can characterize the convergence rate of the CSVRG method at each stage with the following lemma.

Lemma C.4. *We suppose that $\mathbb{E}\left[\mu\left\|\hat{\boldsymbol{z}}_j-\boldsymbol{z}_j^*\right\|^2\right]\leq\epsilon$ holds for $j\in[i-1]$, running the CSVRG methods (Algorithm 1 with Option I) with $\gamma_t=2/(\mu(t+\beta))$ where $\beta=6L^2/\mu^2$, then the output $\hat{\boldsymbol{z}}_i$ holds that*

$$
\mathbb{E}\left[\mu\left\|\hat{\boldsymbol{z}}_i-\boldsymbol{z}_i^*\right\|^2\right]\leq\frac{(\beta-1)(\beta-2)}{(T_i+\beta-1)(T_i+\beta-2)}\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^0-\boldsymbol{z}_i^*\right\|^2\right]+\frac{16G^2L^2\alpha^2}{\mu^3T_i}+\frac{16L^2\epsilon}{\mu^2T_i}.
$$

Proof. Observe that $\boldsymbol{z}_i^* = \Pi_{\mathcal{Z}}\left(\boldsymbol{z}_i^* - \frac{\gamma_t}{i}\sum_{j=1}^{i} F_j(\boldsymbol{z}_i^*)\right)$, we have

$$\begin{aligned}
&\left\|\boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^*\right\|^2 \\
=&\left\|\Pi_{\mathcal{Z}}(\boldsymbol{z}_i^t - \gamma_t \nabla_i^t) - \Pi_{\mathcal{Z}}\left(\boldsymbol{z}_i^* - \frac{\gamma_t}{i}\sum_{j=1}^{i} F_j(\boldsymbol{z}_i^*)\right)\right\|^2 \\
\leq&\left\|\boldsymbol{z}_i^t - \gamma_t \nabla_i^t - \boldsymbol{z}_i^* + \frac{\gamma_t}{i}\sum_{j=1}^{i} F_j(\boldsymbol{z}_i^*)\right\|^2 \\
=&\left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^* - \frac{\gamma_t}{i}\sum_{j=1}^{i}\left(F_j(\boldsymbol{z}_i^t) - F_j(\boldsymbol{z}_i^*)\right) - \gamma_t\left(\nabla_i^t - \frac{1}{i}\sum_{j=1}^{i} F_j(\boldsymbol{z}_i^t)\right)\right\|^2.
\end{aligned}$$

Taking expectations and multiplying μ on both sides of the inequality, we have

$$\begin{aligned}
&\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^*\right\|^2\right] \\
\leq&\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2 - \frac{2\mu\gamma_t}{i}\sum_{j=1}^{i}\langle F_j(\boldsymbol{z}_i^t) - F_j(\boldsymbol{z}_i^*), \boldsymbol{z}_i^t - \boldsymbol{z}_i^*\rangle + \frac{\mu\gamma_t^2}{i}\sum_{j=1}^{i}\left\|F_j(\boldsymbol{z}_i^t) - F_j(\boldsymbol{z}_i^*)\right\|^2\right] \\
&+ \mu\gamma_t^2\mathbb{E}\left[\left\|\nabla_i^t - \frac{1}{i}\sum_{j=1}^{i} F_j(\boldsymbol{z}_i^t)\right\|^2\right].
\end{aligned}$$

Assumption 3.5 implies that

$$\begin{aligned}
&\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^*\right\|^2\right] \\
\leq&(1 - 2\mu\gamma_t)\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2\right] + \mu\gamma_t^2 L^2\mathbb{E}\left[\left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2\right] + \mu\gamma_t^2\mathbb{E}\left[\left\|\nabla_i^t - \frac{1}{i}\sum_{j=1}^{i} F_j(\boldsymbol{z}_i^t)\right\|^2\right] \\
=&(1 - 2\mu\gamma_t + \gamma_t^2 L^2)\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2\right] + \mu\gamma_t^2\mathbb{E}\left[\left\|\nabla_i^t - \frac{1}{i}\sum_{j=1}^{i} F_j(\boldsymbol{z}_i^t)\right\|^2\right].
\end{aligned}$$

By applying Lemma 5.2, we obtain

$$\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^*\right\|^2\right] \leq (1 - 2\mu\gamma_t + 3\gamma_t^2 L^2)\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2\right] + \frac{4G^2L^2\alpha^2\gamma_t^2}{\mu} + 4L^2\gamma_t^2\mathbb{E}\left[\mu\left\|\hat{\boldsymbol{z}}_{\text{prev}} - \boldsymbol{z}_{\text{prev}}^*\right\|^2\right].$$

If we choose $\gamma_t = 2/(\mu(t+\beta))$, where $\beta = 6L^2/\mu^2$, then

$$3\gamma_t L^2 = \frac{6L^2}{\mu(t+\beta)} \leq \frac{6L^2}{\mu\beta} = \mu.$$

Consequently, it follows that

$$\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^*\right\|^2\right] \leq \frac{t+\beta-2}{t+\beta}\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2\right] + \frac{16G^2L^2\alpha^2}{\mu^3(t+\beta)^2} + \frac{16L^2}{\mu^2(t+\beta)^2}\mathbb{E}\left[\mu\left\|\hat{\boldsymbol{z}}_{\text{prev}} - \boldsymbol{z}_{\text{prev}}^*\right\|^2\right].$$

Multiply both sides by $(t+\beta)(t+\beta-1)$, we obtain

$$\begin{aligned}
&(t+\beta)(t+\beta-1)\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^*\right\|^2\right] \\
\leq&(t+\beta-1)(t+\beta-2)\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2\right] + \frac{16G^2L^2\alpha^2}{\mu^3} + \frac{16L^2}{\mu^2}\mathbb{E}\left[\mu\left\|\hat{\boldsymbol{z}}_{\text{prev}} - \boldsymbol{z}_{\text{prev}}^*\right\|^2\right] \\
\leq&(t+\beta-1)(t+\beta-2)\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2\right] + \frac{16G^2L^2\alpha^2}{\mu^3} + \frac{16L^2\epsilon}{\mu^2}.
\end{aligned}$$

The last inequality is due to the induction hypothesis that $\mathbb{E}\left[\mu\left\|\hat{\boldsymbol{z}}_{\text{prev}}-\boldsymbol{z}_{\text{prev}}^*\right\|^2\right]\leq\epsilon$. By summing up the above equality for t from 0 to T_i-1, we have

$$(T_i+\beta-1)(T_i+\beta-2)\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^{T_i}-\boldsymbol{z}_i^*\right\|^2\right]\leq(\beta-1)(\beta-2)\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^0-\boldsymbol{z}_i^*\right\|^2\right]+\frac{16G^2L^2\alpha^2T_i}{\mu^3}+\frac{16L^2\epsilon T_i}{\mu^2}.$$

Divide both sides of the above inequality by $(T_i+\beta-1)(T_i+\beta-2)$, one has

$$\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^{T_i}-\boldsymbol{z}_i^*\right\|^2\right]\leq\frac{(\beta-1)(\beta-2)}{(T_i+\beta-1)(T_i+\beta-2)}\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^0-\boldsymbol{z}_i^*\right\|^2\right]+\frac{16G^2L^2\alpha^2}{\mu^3T_i}+\frac{16L^2\epsilon}{\mu^2T_i}.$$

The inequality holds due to the observation $\beta\geq 1$. Noting that $\hat{\boldsymbol{z}}_i=\boldsymbol{z}_i^{T_i}$, we get the desired bound. □

After characterizing the convergence rate of the CSVRG of a single stage, we wish to compute the total time complexity of the CSVRG method. The following two lemmas are instrumental in establishing the total complexity of the CSVRG method.

Lemma C.5 (Mavrothalassitis et al. (2024)). *Over a sequence of n stages in Algorithm 1, the condition $i-\text{prev}\geq\alpha i$ is satisfied for $\lceil\log n/\alpha\rceil$ times.*

Corollary C.6 (Mavrothalassitis et al. (2024)). *Over a sequence of n stages in Algorithm 1, it requires $\mathcal{O}(\sum_{i=1}^n T_i+n\lceil\log n/\alpha\rceil)$ FOs.*

We can show that under appropriate choices of hyperparameters, the CSVRG method obtains a sequence of ϵ-approximate solutions for the continual VI problem.

C.4 Proof of Theorem 5.3

Proof. We prove the theorem by induction.
Induction Hypothesis. At epoch 1, Algorithm 1 performs extragradient on F_1 to produce $\hat{\boldsymbol{z}}_1$, and we obtain

$$\mathbb{E}[\mu\left\|\hat{\boldsymbol{z}}_1-\boldsymbol{z}_1^*\right\|^2]\leq\epsilon.$$

with $\mathcal{O}(L/\mu\log(1/\epsilon))$ FOs. Now we assume

$$\mathbb{E}\left[\mu\left\|\hat{\boldsymbol{z}}_j-\boldsymbol{z}_j^*\right\|^2\right]\leq\epsilon$$

holds for epochs $j\in[i-1]$.

Induction Step. By Lemma C.2, we have

$$\begin{aligned}&\mu\left\|\boldsymbol{z}_i^0-\boldsymbol{z}_i^*\right\|^2\\=&\mu\left\|\hat{\boldsymbol{z}}_{i-1}-\boldsymbol{z}_i^*\right\|^2\\\leq&\frac{2G^2}{\mu i^2}+2\mu\mathbb{E}\left[\left\|\hat{\boldsymbol{z}}_{i-1}-\boldsymbol{z}_{i-1}^*\right\|^2\right]\\\leq&\frac{2G^2}{\mu i^2}+2\epsilon.\end{aligned}$$

The last inequality is due to the induction hypothesis. By Lemma C.4, we have

$$\begin{aligned}&\mathbb{E}\left[\mu\left\|\hat{\boldsymbol{z}}_i-\boldsymbol{z}_i^*\right\|^2\right]\\\leq&\frac{\beta^2}{T_i^2}\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^0-\boldsymbol{z}_i^*\right\|^2\right]+\frac{16G^2L^2\alpha^2}{\mu^3T_i}+\frac{16L^2\epsilon}{\mu^2T_i}\\=&\frac{36L^4}{\mu^4T_i^2}\mathbb{E}\left[\mu\left\|\boldsymbol{z}_i^0-\boldsymbol{z}_i^*\right\|^2\right]+\frac{16G^2L^2\alpha^2}{\mu^3T_i}+\frac{16L^2\epsilon}{\mu^2T_i}\\\leq&\frac{72L^4G^2}{\mu^5T_i^2i^2}+\frac{72L^4\epsilon}{\mu^4T_i^2}+\frac{16G^2L^2\alpha^2}{\mu^3T_i}+\frac{16L^2\epsilon}{\mu^2T_i}.\end{aligned}$$

By taking $\alpha = \mu\epsilon^{\frac{1}{3}}/(G^{\frac{2}{3}}L^{\frac{2}{3}})$, and we choose

$$T_i = \mathcal{O}\left(\frac{L^2 G}{\mu^{\frac{5}{2}} i\epsilon^{\frac{1}{2}}} + \frac{L^2}{\mu^2} + \frac{G^{\frac{2}{3}}L^{\frac{2}{3}}}{\mu\epsilon^{\frac{1}{3}}}\right),$$

then it holds that

$$\mathbb{E}\left[\mu \left\|\hat{\boldsymbol{z}}_i - \boldsymbol{z}_i^*\right\|^2\right] \leq \epsilon.$$

□

Accordingly, the total time complexity of the CSVRG method is presented in the following corollary.

C.5 Proof of Corollary 5.4

Proof. By adding up T_i from $i = 1$ to n, we have

$$\sum_{i=1}^{n} T_i = \mathcal{O}\left(\frac{L^2 G \log n}{\mu^{\frac{5}{2}}\epsilon^{\frac{1}{2}}} + \frac{nL^2}{\mu^2} + \frac{nG^{\frac{2}{3}}L^{\frac{2}{3}}}{\mu\epsilon^{\frac{1}{3}}}\right).$$

In addition, recall that we choose $\alpha = \mu\epsilon^{\frac{1}{3}}/G^{\frac{2}{3}}L^{\frac{2}{3}}$, then

$$\frac{n\log n}{\alpha} = \mathcal{O}\left(\frac{nG^{\frac{2}{3}}L^{\frac{2}{3}}\log n}{\mu\epsilon^{\frac{1}{3}}}\right).$$

Corollary C.6 implies the total IO calls. □

D Convergence Analysis of the CSVRE Method

In this section, we present the convergence analysis of the CSVRE method for solving the continual VI problem. We first show the convergence of the CSVRE method at each stage as follows.

Lemma D.1. *We suppose that $\mathbb{E}[\mu \left\|\hat{\boldsymbol{z}}_j - \boldsymbol{z}_j^*\right\|^2] \leq \epsilon$ holds for $j \in [i-1]$, running the CSVRE methods (Algorithm 1 with Option II) with $\gamma_t = 2/(\mu(t+\beta))$ where $\beta = 8L/\mu$, and $\eta_t = 1 - 2\gamma_t^2 L^2$, then the output $\hat{\boldsymbol{z}}_i$ holds that*

$$\mu \left\|\hat{\boldsymbol{z}}_i - \boldsymbol{z}_i^*\right\|^2 \leq \frac{72L^2 \left\|\boldsymbol{z}_i^0 - \boldsymbol{z}_i^*\right\|^2}{\mu T_i^2} + \frac{18L^2G^2\alpha^2}{\mu^3 T_i} + \frac{18L^2\epsilon}{\mu^2 T_i}.$$

Proof. Let $\tilde{F}_i(\boldsymbol{z}) = \frac{1}{i}\sum_{j=1}^{i} F_j(\boldsymbol{z})$, the proximal update of $\boldsymbol{z}_i^{t+1/2}$ and $\boldsymbol{z}_i^{t+1}$ and Lemma B.2 imply that

$$\left\langle \boldsymbol{z}_i^{t+1/2} - \bar{\boldsymbol{z}}_i^t + \gamma_t\left(\left(1-\frac{1}{i}\right)\tilde{F}_{i-1}(\hat{\boldsymbol{z}}_{\text{prev}}) + \frac{1}{i}F_i(\hat{\boldsymbol{z}}_{\text{prev}})\right), \boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^{t+1/2}\right\rangle \geq 0,$$

and

$$\left\langle \boldsymbol{z}_i^{t+1} - \bar{\boldsymbol{z}}_i^t + \gamma_t\left(\left(1-\frac{1}{i}\right)\left(F_{u_t}(\boldsymbol{z}_i^{t+1/2}) - F_{u_t}(\hat{\boldsymbol{z}}_{\text{prev}}) + \tilde{F}_{i-1}(\hat{\boldsymbol{z}}_{\text{prev}})\right) + \frac{1}{i}F_i(\boldsymbol{z}_i^{t+1/2})\right), \boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\rangle \geq 0.$$

Summing up these two inequalities, we have

$$\begin{aligned}
0 \leq & \langle \boldsymbol{z}_i^{t+1/2} - \bar{\boldsymbol{z}}_i^t, \boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^{t+1/2}\rangle + \langle \boldsymbol{z}_i^{t+1} - \bar{\boldsymbol{z}}_i^t, \boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\rangle \\
& + \gamma_t\left\langle\left(1-\frac{1}{i}\right)\left(F_{u_t}(\boldsymbol{z}_i^{t+1/2}) - F_{u_t}(\hat{\boldsymbol{z}}_{\text{prev}}) + \tilde{F}_{i-1}(\hat{\boldsymbol{z}}_{\text{prev}})\right) + \frac{1}{i}F_i(\boldsymbol{z}_i^{t+1/2}), \boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1/2}\right\rangle \\
& + \gamma_t\left\langle\left(1-\frac{1}{i}\right)\left(F_{u_t}(\boldsymbol{z}_i^{t+1/2}) - F_{u_t}(\hat{\boldsymbol{z}}_{\text{prev}})\right) + \frac{1}{i}\left(F_i(\boldsymbol{z}_i^{t+1/2}) - F_i(\hat{\boldsymbol{z}}_{\text{prev}})\right), \boldsymbol{z}_i^{t+1/2} - \boldsymbol{z}_i^{t+1}\right\rangle.
\end{aligned} \tag{11}$$

Using $2\langle \boldsymbol{a}, \boldsymbol{b}\rangle = \|\boldsymbol{a}+\boldsymbol{b}\|^2 - \|\boldsymbol{a}\|^2 - \|\boldsymbol{b}\|^2$ and recall that $\bar{\boldsymbol{z}}_i^t = \eta_t \boldsymbol{z}_i^t + (1-\eta_t)\hat{\boldsymbol{z}}_{\text{prev}}$, the first term on the r.h.s. of (11) can be written as

$$
\begin{aligned}
&2\left\langle \boldsymbol{z}_i^{t+1/2} - \bar{\boldsymbol{z}}_i^t, \boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^{t+1/2}\right\rangle \\
=&2\eta_t \left\langle \boldsymbol{z}_i^{t+1/2} - \boldsymbol{z}_i^t, \boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^{t+1/2}\right\rangle + 2(1-\eta_t)\left\langle \boldsymbol{z}_i^{t+1/2} - \hat{\boldsymbol{z}}_{\text{prev}}, \boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^{t+1/2}\right\rangle \\
=&\eta_t\left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^{t+1}\right\|^2 - \eta_t\left\|\boldsymbol{z}_i^{t+1/2} - \boldsymbol{z}_i^t\right\|^2 + (1-\eta_t)\left\|\hat{\boldsymbol{z}}_{\text{prev}} - \boldsymbol{z}_i^{t+1}\right\|^2 \\
&-(1-\eta_t)\left\|\boldsymbol{z}_i^{t+1/2} - \hat{\boldsymbol{z}}_{\text{prev}}\right\|^2 - \left\|\boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^{t+1/2}\right\|^2.
\end{aligned}
\tag{12}
$$

Similarly, the second term of (11) can be written as

$$
\begin{aligned}
&2\left\langle \boldsymbol{z}_i^{t+1} - \bar{\boldsymbol{z}}_i^t, \boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\rangle \\
=&2\left\langle \boldsymbol{z}_i^{t+1} - \eta_t\boldsymbol{z}_i^t - (1-\eta_t)\hat{\boldsymbol{z}}_{\text{prev}}, \boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\rangle \\
=&2\eta_t\left\langle \boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^t, \boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\rangle + 2(1-\eta_t)\left\langle \boldsymbol{z}_i^{t+1} - \hat{\boldsymbol{z}}_{\text{prev}}, \boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\rangle \\
=&\eta_t\left(\left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2 - \left\|\boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^t\right\|^2 - \left\|\boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\|^2\right) \\
&+(1-\eta_t)\left(\|\hat{\boldsymbol{z}}_{\text{prev}} - \boldsymbol{z}_i^*\|^2 - \left\|\boldsymbol{z}_i^{t+1} - \hat{\boldsymbol{z}}_{\text{prev}}\right\|^2 - \left\|\boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\|^2\right) \\
=&\eta_t\left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2 - \eta_t\left\|\boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^t\right\|^2 + (1-\eta_t)\left\|\hat{\boldsymbol{z}}_{\text{prev}} - \boldsymbol{z}_i^*\right\|^2 \\
&-(1-\eta_t)\left\|\boldsymbol{z}_i^{t+1} - \hat{\boldsymbol{z}}_{\text{prev}}\right\|^2 - \left\|\boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\|^2.
\end{aligned}
\tag{13}
$$

Using the fact $\mathbb{E}\left[F_{u_t}(\boldsymbol{z})\right] = \tilde{F}_{i-1}(\boldsymbol{z})$ for $\forall \boldsymbol{z} \in \mathcal{Z}$, we can bound the third term of (11) as follows

$$
\begin{aligned}
&2\mathbb{E}\left[\left\langle\left(1-\frac{1}{i}\right)\left(F_{u_t}(\boldsymbol{z}_i^{t+1/2}) - F_{u_t}(\hat{\boldsymbol{z}}_{\text{prev}}) + \tilde{F}_{i-1}(\hat{\boldsymbol{z}}_{\text{prev}})\right) + \frac{1}{i}F_i(\boldsymbol{z}_i^{t+1/2}), \boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1/2}\right\rangle\right] \\
=&2\left\langle\left(1-\frac{1}{i}\right)\tilde{F}_{i-1}\left(\boldsymbol{z}_i^{t+1/2}\right) + \frac{1}{i}F_i\left(\boldsymbol{z}_i^{t+1/2}\right), \boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1/2}\right\rangle \\
=&2\left\langle\tilde{F}_i\left(\boldsymbol{z}_i^{t+1/2}\right), \boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1/2}\right\rangle \\
\leq&2\left\langle\tilde{F}_i\left(\boldsymbol{z}_i^*\right), \boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1/2}\right\rangle - 2\mu\left\|\boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1/2}\right\|^2 \\
\leq&-\mu\left\|\boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\|^2 + 2\mu\left\|\boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^{t+1/2}\right\|^2.
\end{aligned}
\tag{14}
$$

The last inequality is due to $\|\boldsymbol{a}+\boldsymbol{b}\|^2 \leq 2\|\boldsymbol{a}\|^2 + 2\|\boldsymbol{b}\|^2$ and the optimality of $\boldsymbol{z}_i^*$. Furthermore, using Young's inequality, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[2\gamma_t\left\langle\left(1-\frac{1}{i}\right)\left(F_{u_t}(\boldsymbol{z}_i^{t+1/2}) - F_{u_t}(\hat{\boldsymbol{z}}_{\text{prev}})\right) + \frac{1}{i}\left(F_i(\boldsymbol{z}_i^{t+1/2}) - F_i(\hat{\boldsymbol{z}}_{\text{prev}})\right), \boldsymbol{z}_i^{t+1/2} - \boldsymbol{z}_i^{t+1}\right\rangle\right] \\
=&\mathbb{E}\left[2\gamma_t\left\langle\left(1-\frac{1}{i}\right)\left(\tilde{F}_{i-1}(\boldsymbol{z}_i^{t+1/2}) - \tilde{F}_{i-1}(\hat{\boldsymbol{z}}_{\text{prev}})\right) + \frac{1}{i}\left(F_i(\boldsymbol{z}_i^{t+1/2}) - F_i(\hat{\boldsymbol{z}}_{\text{prev}})\right), \boldsymbol{z}_i^{t+1/2} - \boldsymbol{z}_i^{t+1}\right\rangle\right] \\
=&\mathbb{E}\left[2\gamma_t\left\langle\tilde{F}_i(\boldsymbol{z}_i^{t+1/2}) - \tilde{F}_i(\hat{\boldsymbol{z}}_{\text{prev}}), \boldsymbol{z}_i^{t+1/2} - \boldsymbol{z}_i^{t+1}\right\rangle\right] \\
\leq&2\gamma_t^2\mathbb{E}\left[\left\|\tilde{F}_i(\boldsymbol{z}_i^{t+1/2}) - \tilde{F}_i(\hat{\boldsymbol{z}}_{\text{prev}})\right\|^2\right] + \frac{1}{2}\mathbb{E}\left[\left\|\boldsymbol{z}_i^{t+1/2} - \boldsymbol{z}_i^{t+1}\right\|^2\right] \\
\leq&2\gamma_t^2L^2\mathbb{E}\left[\left\|\boldsymbol{z}_i^{t+1/2} - \hat{\boldsymbol{z}}_{\text{prev}}\right\|^2\right] + \frac{1}{2}\mathbb{E}\left[\left\|\boldsymbol{z}_i^{t+1/2} - \boldsymbol{z}_i^{t+1}\right\|^2\right],
\end{aligned}
\tag{15}
$$

where the first inequality is due to Young's inequality and the last inequality follows from Assumption 3.5. Substituting (12)-(15) into eq. (11), we can show that

$$\begin{aligned}
&\eta_t \left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2 + (1-\eta_t)\left\|\hat{\boldsymbol{z}}_{\text{prev}} - \boldsymbol{z}_i^*\right\|^2 - \left\|\boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\|^2 - \eta_t \left\|\boldsymbol{z}_i^{t+1/2} - \boldsymbol{z}_i^t\right\|^2 \\
&- (1-\eta_t)\left\|\boldsymbol{z}_i^{t+1/2} - \hat{\boldsymbol{z}}_{\text{prev}}\right\|^2 - \left\|\boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^{t+1/2}\right\|^2 \\
&- \gamma_t\mu\left\|\boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\|^2 + 2\gamma_t\mu\left\|\boldsymbol{z}_i^{t+1} - \boldsymbol{z}_i^{t+1/2}\right\|^2 \\
&+ 2\gamma_t^2 L^2 \mathbb{E}\left[\left\|\boldsymbol{z}_i^{t+1/2} - \hat{\boldsymbol{z}}_{\text{prev}}\right\|^2\right] + \frac{1}{2}\mathbb{E}\left[\left\|\boldsymbol{z}_i^{t+1/2} - \boldsymbol{z}_i^{t+1}\right\|^2\right] \geq 0.
\end{aligned}$$

Rearrange the terms, we have

$$\begin{aligned}
&(1+\gamma_t\mu)\left\|\boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\|^2 \\
\leq&\eta_t \left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2 + (1-\eta_t)\left\|\hat{\boldsymbol{z}}_{\text{prev}} - \boldsymbol{z}_i^*\right\|^2 - \left(\frac{1}{2} - 2\gamma_t\mu\right)\left\|\boldsymbol{z}_i^{t+1/2} - \boldsymbol{z}_i^{t+1}\right\|^2 \\
&- \left(1 - \eta_t - 2\gamma_t^2 L^2\right)\left\|\boldsymbol{z}_i^{t+1/2} - \hat{\boldsymbol{z}}_{\text{prev}}\right\|^2.
\end{aligned}$$

If we choose $\gamma_t = 2/(\mu(t+\beta))$ and $\eta_t = 1 - 2\gamma_t^2 L^2$ where $\beta = 8L/\mu$, such that

$$\frac{1}{2} - 2\gamma_t\mu \geq 0 \quad \text{and} \quad 1 - 2\gamma_t^2 L^2 \geq 0.$$

We can show that

$$\begin{aligned}
&(1+\gamma_t\mu)\left\|\boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\|^2 \\
\leq&\eta_t \left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2 + (1-\eta_t)\left\|\hat{\boldsymbol{z}}_{\text{prev}} - \boldsymbol{z}_i^*\right\|^2 \\
\leq&\eta_t \left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2 + (1-\eta_t)\left[\frac{2G^2 i^2\alpha^2}{\mu^2 i^2} + 2\left\|\hat{\boldsymbol{z}}_{\text{prev}} - \boldsymbol{z}_{\text{prev}}^*\right\|^2\right] \\
\leq&\eta_t \left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2 + \frac{(1-\eta_t)2G^2\alpha^2}{\mu^2} + \frac{2(1-\eta_t)\epsilon}{\mu}.
\end{aligned}$$

The second inequality follows from Lemma C.2. The last inequality is due to the induction hypothesis. Since we choose $\gamma_t = 2/(\mu(t+\beta))$ and $\eta_t = 1 - 2\gamma_t^2 L^2$ where $\beta = 8L/\mu$, then

$$\frac{t+\beta+2}{t+\beta}\left\|\boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\|^2 \leq \left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2 + \frac{16L^2G^2\alpha^2}{\mu^4(t+\beta)^2} + \frac{16L^2\epsilon}{\mu^3(t+\beta)^2}.$$

Multiply both sides by $(t+\beta)(t+\beta+1)$, we have

$$\begin{aligned}
&(t+\beta+2)(t+\beta+1)\left\|\boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\|^2 \\
\leq&(t+\beta)(t+\beta+1)\left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2 + \frac{16L^2G^2\alpha^2(t+\beta+1)}{\mu^4(t+\beta)} + \frac{16L^2\epsilon(t+\beta+1)}{\mu^3(t+\beta)}.
\end{aligned}$$

Observe that

$$\frac{t+\beta+1}{t+\beta} \leq \frac{\beta+1}{\beta} \leq \frac{9}{8},$$

we can deduce that

$$\begin{aligned}
&(t+\beta+2)(t+\beta+1)\mu\left\|\boldsymbol{z}_i^* - \boldsymbol{z}_i^{t+1}\right\|^2 \\
\leq&(t+\beta)(t+\beta+1)\mu\left\|\boldsymbol{z}_i^t - \boldsymbol{z}_i^*\right\|^2 + \frac{18L^2G^2\alpha^2}{\mu^3} + \frac{18L^2\epsilon}{\mu^2}.
\end{aligned}$$

Sum up the above inequalities from $t=0$ to $T_i - 1$, we can obtain that

$$(T_i+\beta+1)(T_i+\beta)\mu\left\|\boldsymbol{z}_i^* - \boldsymbol{z}_i^{T_i}\right\|^2 \leq \beta(\beta+1)\mu\left\|\boldsymbol{z}_i^0 - \boldsymbol{z}_i^*\right\|^2 + \frac{18L^2G^2\alpha^2T_i}{\mu^3} + \frac{18L^2\epsilon T_i}{\mu^2}.$$

Dividing both sides by $(T_i+\beta+1)(T_i+\beta)$ and substituting $\beta = 8L/\mu$, one has

$$\begin{aligned}\mu\left\|\boldsymbol{z}_i^{T_i}-\boldsymbol{z}_i^*\right\|^2 \leq & \frac{\beta(\beta+1)\mu\left\|\boldsymbol{z}_i^0-\boldsymbol{z}_i^*\right\|^2}{(T_i+\beta+1)(T_i+\beta)}+\frac{18L^2G^2\alpha^2T_i}{\mu^3(T_i+\beta+1)(T_i+\beta)}+\frac{18L^2\epsilon T_i}{\mu^2(T_i+\beta+1)(T_i+\beta)} \\ \leq & \frac{72L^2\left\|\boldsymbol{z}_i^0-\boldsymbol{z}_i^*\right\|^2}{\mu T_i^2}+\frac{18L^2G^2\alpha^2}{\mu^3T_i}+\frac{18L^2\epsilon}{\mu^2T_i}.\end{aligned}$$

Noting $\hat{\boldsymbol{z}}_i = \boldsymbol{z}_i^{T_i}$, it completes the proof. □

We can show that under appropriate choices of hyperparameters, the CSVRE method obtains a sequence of ϵ-approximate solutions for the continual finite-sum VI problem.

D.1 Proof of Theorem 5.6

Proof. We prove the theorem by induction.

Induction Hypothesis. At epoch 1, Algorithm 1 performs the extragradient method on F_1 to produce $\hat{\boldsymbol{z}}_1$, and we obtain

$$\mathbb{E}[\mu\left\|\hat{\boldsymbol{z}}_1-\boldsymbol{z}_1^*\right\|^2] \leq \epsilon.$$

with $\mathcal{O}(L/\mu\log(1/\epsilon))$ FOs. Now we assume $\mathbb{E}\left[\mu\left\|\hat{\boldsymbol{z}}_j-\boldsymbol{z}_j^*\right\|^2\right] \leq \epsilon$ holds for epochs $j \in [i-1]$.

Induction Step. By Lemma C.2, we have

$$\begin{aligned}&\mu\left\|\boldsymbol{z}_i^0-\boldsymbol{z}_i^*\right\|^2 \\ =&\mu\left\|\hat{\boldsymbol{z}}_{i-1}-\boldsymbol{z}_i^*\right\|^2 \\ \leq&\frac{2G^2}{\mu i^2}+2\mu\left\|\hat{\boldsymbol{z}}_{i-1}-\boldsymbol{z}_{i-1}^*\right\|^2 \\ \leq&\frac{2G^2}{\mu i^2}+2\epsilon.\end{aligned}$$

The last inequality is due to the induction hypothesis. By Lemma D.1, we have

$$\begin{aligned}&\mu\left\|\hat{\boldsymbol{z}}_i-\boldsymbol{z}_i^*\right\|^2 \\ \leq&\frac{72L^2}{\mu^2T_i^2}\left(\frac{2G^2}{\mu i^2}+2\epsilon\right)+\frac{18L^2G^2\alpha^2}{\mu^3T_i}+\frac{18L^2\epsilon}{\mu^2T_i} \\ =&\frac{144L^2G^2}{\mu^3i^2T_i^2}+\frac{144L^2\epsilon}{\mu^2T_i^2}+\frac{18L^2G^2\alpha^2}{\mu^3T_i}+\frac{18L^2\epsilon}{\mu^2T_i}.\end{aligned}$$

By taking $\alpha = \mu\epsilon^{\frac{1}{3}}/(G^{\frac{2}{3}}L^{\frac{2}{3}})$ and choose

$$T_i = \mathcal{O}\left(\frac{LG}{\mu^{\frac{3}{2}}i\epsilon^{\frac{1}{2}}}+\frac{L^2}{\mu^2}+\frac{G^{\frac{2}{3}}L^{\frac{2}{3}}}{\mu\epsilon^{\frac{1}{3}}}\right),$$

we obtain

$$\mu\left\|\hat{\boldsymbol{z}}_i-\boldsymbol{z}_i^*\right\|^2 \leq \epsilon.$$

□

D.2 Proof of Corollary 5.7

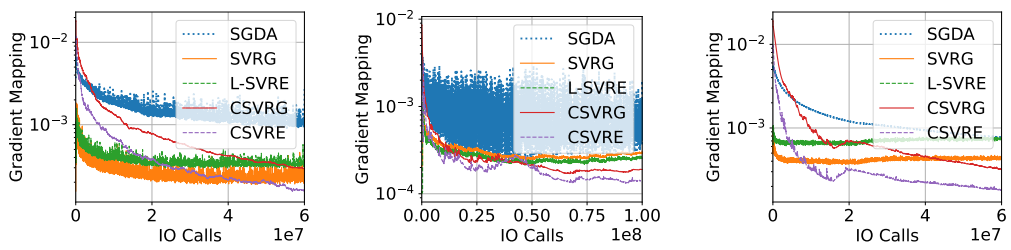
Proof. By summing up T_i from $i=1$ to n, we have

$$\sum_{i=1}^n T_i = \mathcal{O}\left(\frac{LG\log n}{\mu^{\frac{3}{2}}\epsilon^{\frac{1}{2}}}+\frac{L^2n}{\mu^2}+\frac{G^{\frac{2}{3}}L^{\frac{2}{3}}n}{\mu\epsilon^{\frac{1}{3}}}\right).$$

In addition, by choosing $\alpha = \mu\epsilon^{\frac{1}{3}}/(G^{\frac{2}{3}}L^{\frac{2}{3}})$, one has

$$\frac{n\log n}{\alpha} = \mathcal{O}\left(\frac{nG^{\frac{2}{3}}L^{\frac{2}{3}}\log n}{\mu\epsilon^{\frac{1}{3}}}\right).$$

Applying Corollary C.6 implies the total IO calls. □



(a) adult (b) bank market (c) law school

Figure 3: Gradient mapping vs the number of IO calls for the fairness-aware machine learning problem.

E Experiments on Fairness-Aware Machine Learning

We consider the following prefix-sum fairness-aware machine learning problem

$$\min_{\boldsymbol{x}\in\mathbb{R}^d}\max_{y\in\mathbb{R}} g_i(\boldsymbol{x},y) := \frac{1}{i}\sum_{j=1}^{i} f_j(\boldsymbol{x},y),$$

where each component function is

$$f_j(\boldsymbol{x},y) = \big(l(\boldsymbol{a}_j,b_j,\boldsymbol{x}) - \beta l(\boldsymbol{a}_j^\top \boldsymbol{x},c_j,y)\big) + \lambda\|\boldsymbol{x}\|^2 - \gamma y^2,$$

and $l(\boldsymbol{a},b,\boldsymbol{c}) = \log(1+\exp(-b\boldsymbol{a}^\top\boldsymbol{c}))$. Here, the tuple $(\boldsymbol{a}_j,b_j,c_j)$ is a training data point where $\boldsymbol{a}_j\in\mathbb{R}^d$ is the input variable, $b_j\in\mathbb{R}$ is the output, and $c_j\in\mathbb{R}$ is the input variable that we want to protect and make it unbiased.

Our experiments are conducted on the fairness-aware binary classification datasets "adult", "bank market", and "law school" (Le Quy et al., 2022; Liu & Luo, 2022). We set the parameters β, λ and γ as 0.5, 10^{-4} and 10^{-4}, respectively. For other hyperparameters, including stepsize and the number of iterations T_i, we use the same parameter setting as in the robust linear regression experiment.

The experimental results for this task are presented in Figure 3. As the number of I/O calls increases, both the CSVRG and CSVRE methods demonstrate consistently superior performance compared with the baseline approaches, highlighting their effectiveness in leveraging larger datasets.