OPT2024: 16th Annual Workshop on Optimization for Machine Learning

Stochastic Quasi-Variational Inequalities: Convergence Analysis Beyond Strong Monotonicity

Zeinab Alizadeh ZALIZADEH@ARIZONA.EDU
Afrooz Jalilzadeh AFROOZ@ARIZONA.EDU
The University of Arizona, USA

Abstract

Variational Inequality is a well-established framework for Nash equilibrium and saddle-point problems. However, its generalization, Quasi-Variational Inequalities, where the constraint set depends on the decision variable, is less understood, with existing results focused on strongly monotone cases. This paper proposes an extra-gradient method for a class of monotone Stochastic Quasi-Variational Inequality (SQVI) and provides the first convergence rate analysis for the non-strongly monotone setting. Our approach not only advances the theoretical understanding of SQVI but also demonstrates its practical applicability.

Keywords: Quasi-Variational Inequalities, Generalized Nash Equilibrium, Extra-Gradient Method

1. Introduction

Variational Inequality (VI) problems have applications in areas like Nash games, traffic, and economic equilibrium [9]. The Stochastic Variational Inequality (SVI) extends VI theory to handle decision-making under uncertainty [12]. Quasi-Variational Inequality (QVI) arises when the constraint set depends on the decision variable, capturing interdependencies in shared-resource games. This paper focuses on the Stochastic QVI (SQVI) problem. In particular, the goal is to find $x^* \in K(x^*)$ such that the following holds:

$$\langle F(x^*), y - x^* \rangle \geq 0, \quad \forall y \in K(x^*), \tag{SQVI}$$

where $K : X \to 2^X$ is a set-valued mapping with non-empty, closed and convex values such that $K(x) \subseteq X$ for all $x \in X$, $X \subseteq \mathbb{R}^n$ is a convex and compact set, $F(x) \triangleq \mathbb{E}[G(x, \xi)]$, $\xi : \Omega \to \mathbb{R}^d$, $G : X \times \mathbb{R}^d \to \mathbb{R}^n$, and the associated probability space is denoted by $(\Omega, \mathcal{F}, \mathbb{P})$.

Although the theoretical results and algorithm development for VIs are rich and fruitful [2, 7, 11, 15, 18, 20, 21, 23, 27, 29, 40], research studies on QVIs remain limited and most of the existing methods for solving VIs are not amendable for solving (**SQVI**) which calls for the development of new techniques and iterative methods. In particular, the primary focus of existing research studies for QVIs is on solution existence [35] and the development of algorithms often requires restrictive assumptions such as strong monotonicity [25]. To fill this gap, in this paper, we aim to develop efficient inexact iterative methods for solving (**SQVI**) under less restrictive assumptions with convergence rate guarantees. In the deterministic setting, several studies have explored numerical approaches for solving QVIs [3, 10, 24, 32–34, 36, 37]. Notably, Mijajlović et al. [25] demonstrated a linear convergence rate for the strongly monotone QVI problem (see also Nesterov and Scrimali [30]). In the stochastic regime, Alizadeh et al. [1] obtained a linear convergence rate under strong

monotonicity assumptions. In this paper, we propose an extra-gradient method for solving monotone SQVIs, where the operator F satisfies the quadratic growth property (see Definition 1). To the best of our knowledge, this is the first-rate result for non-strongly monotone QVIs. Table 1 shows a summary of existing complexity results for VI and QVI problems.

Reference	Problem	Setting	Operator	Complexity
[30]	VI	Deterministic	Strongly Monotone	$\mathcal{O}(\log(1/\epsilon))$
[15]	VI	Stochastic	Strongly Monotone	$\mathcal{O}(1/\epsilon)$
[28]	VI	Deterministic	Monotone	$\mathcal{O}(1/\epsilon)$
[17]	VI	Stochastic	Monotone	$\mathcal{O}(1/\epsilon^2)$
[30]	QVI	Deterministic	Strongly Monotone	$\mathcal{O}(\log(1/\epsilon))$
[1]	QVI	Stochastic	Strongly Monotone	$\mathcal{O}(1/\epsilon^2)$
This Paper	QVI	Stochastic	Monotone & Quadratic Growth	$\mathcal{O}(1/\epsilon^2)$

Table 1: Compression of Complexity Results for VI and QVI

2. Applications

Generalized Nash Equilibruim. Nash equilibrium (NE) is a key game theory concept where a group of selfish agents compete, each optimizing their own objective. An NE occurs when no player can lower their cost by unilaterally changing their strategy. The Generalized Nash Equilibrium Problem (GNEP) extends NE by allowing each player's strategy set to depend on others' strategies, which often occurs when sharing a common resource (e.g., a communication link or power grid). GNEP is widely used in fields like economics and operations research [8, 19]. Consider N players each with cost function $f_i(x_i, x_{(-i)}) \triangleq \mathbb{E}[h(x_i, x_{(-i)}, \xi)]$ for $i = 1 \ldots, N$, where x_i is the strategy of player i and $x_{(-i)}$ is the strategy of other players. Each player i's objective is to solve the following optimization problem: $\min_{x_i} f_i(x_i, x_{(-i)})$ such that $x_i \in K_i(x_{(-i)})$, where $K_i(x_{(-i)}) = \{x_i \in \mathbb{R}^{n_i} | g_i(x_i, x_{(-i)}) \leq 0\}$ is a closed convex set-valued map. $f_i, g_i : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ are continuously differentiable. By defining $K(x) = \prod_{i=1}^N K_i(x_{(-i)})$ and $F(x) = [\nabla_{x_i} f_i(x)]_{i=1}^N$, finding a GNE will be equivalent to SQVI problem: $\langle F(x^*), x - x^* \rangle \geq 0, \forall x \in K(x^*)$.

Saddle Point Problems with Coupling Constraints. Saddle point problems, $\min_u \max_w f(u, w)$, have gained attention due to their relevance in machine learning applications like reinforcement learning, GANs, fairness, and adversarial imitation learning. Convex-concave minimax problems can also be viewed through a game theory lens, where one player minimizes and the other maximizes the payoff. A saddle point (u^*, w^*) represents both a minimum in the u-direction and a maximum in the w-direction. Here, w^* is the inner player's best response to the opponent's strategy u^*, and a saddle point (u^*, w^*) is also called a Nash equilibrium (NE). Here, we consider a more general SP problem where the constraint depends on the decisions of both players, i.e.,

$$\min_{u \in U} \max_{w \in W} f(u, w) \quad \text{s.t.} \quad g(u, w) \leq 0, \tag{1}$$

where $f(u, w) \triangleq \mathbb{E}[h(u, w, \xi)]$, U and W are convex sets. Such problems have numerous applications in various fields such as adversarial attacks in network flow problems [39]. Because of the dependency of the constraint on both variables, if g is not jointly convex in both x and y

then this problem cannot be formulated as traditional VI but we can reformulate it as QVI. From the first-order optimality condition of (1), we have that $\langle \nabla_x f(u^*, w^*), u - u^* \rangle \geq 0, \forall u \in \{u \in U \mid g(u, w^*) \leq 0\}$ and $\langle \nabla_w f(u^*, w^*), w^* - w \rangle \geq 0, \forall w \in \{w \in W | g(u^*, w) \leq 0\}$. Defining $F(x^*) = \begin{bmatrix} \nabla_u f(x^*) & -\nabla_w f(x^*) \end{bmatrix}^T$ and $K(x^*) \triangleq U(w^*) \times W(u^*)$, solving (1) will be equivalent to solving the following SQVI problem: $\langle F(x^*), x - x^* \rangle \geq 0, \forall x \in K(x^*)$.

3. Preliminaries

Notations. Throughout the paper, $\|x\|$ denotes the Euclidean vector norm, i.e., $\|x\| = \sqrt{x^T x}$. $\mathcal{P}_K[x]$ is the projection of x onto the set K, i.e. $\mathcal{P}_K[x] = \text{argmin}_{y \in K} \|y - x\|$. $\mathbb{E}[x]$ is used to denote the expectation of a random variable x. We let X^* denote the set of optimal solution of (**SQVI**) problem, which is assumed to be nonempty.

Assumptions and Technical Lemmas. In this paper, we consider a monotone operator F that has a quadratic growth property, which is defined next.

Definition 1 *An operator F has a quadratic growth (QG) property on set X if there exists a constant $\mu_F > 0$ such that for any $x \in X$ and $y = \mathcal{P}_{X^*}(x)$, we have $\langle F(x) - F(y), x - y \rangle \geq \mu_F \|x - y\|^2$, for all $x \in X$.*

It is worth noting that, unlike the strong monotonicity assumption, the QG property does not imply a unique solution. In fact, QG property is a weaker assumption than strong monotonicity [26]. As an example of a QG operator, consider function $f(x) \triangleq g(Ax) + c^T x$, where g is a smooth and strongly convex function, $A \in \mathbb{R}^{n \times m}$ is a nonzero general matrix and $c \in \mathbb{R}^n$. One can show that $\nabla f(x)$ satisfies the QG property [26] while $\nabla f(x)$ may not be strongly monotone unless A has a full column rank.

Assumption 1 *(i) The set of optimal solution, X^* is nonempty. (ii) Operator $F : X \to \mathbb{R}^n$ is monotone, i.e., $\langle F(x) - F(y), x - y \rangle \geq 0$ for all $x, y \in X$, and satisfies the QG property. (iii) F is L-Lipschitz continuous on X, i.e., $\|F(x) - F(y)\| \leq L\|x - y\|$ for all $x, y \in X$.*

If $\mathcal{F}_k$ denotes the information history at epoch k, then we have the following requirements on the associated filtrations where $\bar{w}_{k,N_k} \triangleq \frac{1}{N_k} \sum_{j=1}^{N_k} (G(x_k, \xi_{j,k}) - F(x_k))$.

Assumption 2 *There exists $\nu > 0$ such that $\mathbb{E}[\bar{w}_{k,N_k} \mid \mathcal{F}_k] = 0$ and $\mathbb{E}[\|\bar{w}_{k,N_k}\|^2 \mid \mathcal{F}_k] \leq \frac{\nu^2}{N_k}$ holds almost surely for all k, where $\mathcal{F}_k \triangleq \sigma\{x_0, x_1, \ldots, x_{k-1}\}$.*

Gap Function. Now, we define a gap function to measure the quality of the solution obtained from the algorithm. In particular, for a given iterate x we use $\text{dist}(x, X^*) \triangleq \|x - \mathcal{P}_{X^*}(x)\|$ to find the distance of the solution obtained by the algorithm from the optimal solution set X^*. Moreover, we call x to be an ϵ-solution if $\|x - \bar{x}\| \leq \epsilon$ where $\bar{x} \triangleq \mathcal{P}_{X^*}(x)$.

4. Proposed Method

A popular method for solving SVI problems is the stochastic Extra-gradient (SEG) method, originally proposed by Korpelevich [18]. When $K(x) = K$ is a closed and convex set, (**SQVI**) reduces to an SVI. The challenge in solving SQVI lies in the dynamic nature of the constraint set, which evolves during iterations. To address this, we impose a condition on the projection operator to

ensure that $K(x)$ does not change drastically as x varies, guaranteeing the projection remains contractive. This assumption is fundamental for convergence in QVI problems and is present in all existing results, indicating its necessity for current approaches [1, 30, 31].

Assumption 3 *There exists $\gamma > 0$ such that $\|\mathbf{P}_{K(x)}[u] - \mathbf{P}_{K(y)}[u]\| \leq \gamma\|x-y\|$ for all $x, y, u \in X$ and $\gamma + \sqrt{1-\mu_F{}^2/L^2} < 1$.*

Algorithm 1 inexact Extra-gradient SQVI (iEG-SQVI)

Input: $x_0 \in X$, $\eta > 0$, $\{N_k\}_k$, $\{t_k\}_k$, $\{b_k\}_k$, $\{\alpha_k\}_k$ and Algorithm $\mathcal{M}$ satisfying Assumption 4;
for $k = 0, \dots T-1$ **do**
(1) Use Algorithm $\mathcal{M}$ with t_k iterations to find an approximated solution d_k of

$$\min_{x \in K(x_k)} \left\| x - \left(x_k - \eta \frac{\sum_{j=1}^{N_k} G(x_k, \xi_{j,k})}{N_k} \right) \right\|^2;$$

(2) $u_k \leftarrow (1-b_k)x_k + b_k d_k$;
(3) Use Algorithm $\mathcal{M}$ with t_k iterations to find an approximated solution s_k of

$$\min_{x \in K(u_k)} \left\| x - \left(u_k - \eta \frac{\sum_{j=1}^{N_k} G(u_k, \xi'_{j,k})}{N_k} \right) \right\|^2;$$

(4) $x_{k+1} \leftarrow (1-\alpha_k)x_k + \alpha_k s_k$;
end for

To ensure convergence, a retraction step [25], $(1-\alpha)x_k + \alpha s_k$, is introduced for some $\alpha \in [0, 1]$. Moreover, the exact computation of the projection onto the constraint set can be computationally expensive or infeasible. To address this, we propose using approximation techniques to obtain practical solutions. Specifically, we assume the constraints are defined by a smooth nonlinear function $g : X \times X \to \mathbb{R}^m$, with $K(x) = \{y \in X \mid g(x, y) \leq 0\}$, where $g(x, \cdot)$ is convex for any $x \in X$. In Algorithm 1, we introduce the inexact Extra-gradient SQVI (iEG-SQVI) method, which approximates the projection using an inner algorithm, $\mathcal{M}$, running for t_k inner iterations. To ensure fast convergence, $\mathcal{M}$ must satisfy the following property.

Assumption 4 *For any $x \in \mathbb{R}^n$, any closed and convex set $K \subseteq \mathbb{R}^n$, and an initial point u_0, $\mathcal{M}$ can generate an output $u \in \mathbb{R}^n$ such that $\|u - \tilde{u}\|^2 \leq C/t^2$ for some $C > 0$ satisfying $\tilde{u} = \text{argmin}_{y \in K}\{\frac{1}{2}\|y - x\|^2\}$.*

Next, we discuss that several optimization methods satisfy the condition outlined in Assumption 4.

Remark 2 *When the constraint set $K(x)$ involves (non)linear convex constraints, steps (1) and (3) of Algorithm 1 require inexact computation of the projection, which entails solving a strongly convex problem with convex constraints. Efficient first-order primal-dual methods, such as those in [13, 14, 22], achieve a convergence rate of $\mathcal{O}(1/t^2)$ in terms of suboptimality and infeasibility, where t is the number of iterations.*

4.1. Convergence Analysis

Next, we introduce a crucial lemma for our convergence analysis. As previously discussed, the problem is not strong monotone and may not possess a unique solution. Therefore, we define the gap function as $\text{dist}(x, X^*) \triangleq \|x - \bar{x}\|$, where $\bar{x} = \mathcal{P}_{X^*}(x)$. Since the optimal solutions are not explicitly available, we need to express $\bar{x}$ based on its first-order optimality condition. This representation will be utilized in the subsequent convergence analysis of the algorithm.

Lemma 3 *Let X^* denote the set of optimal solutions of problem (SQVI). Moreover, for any $x \in X$ define $\bar{x} \triangleq \mathcal{P}_{X^*}(x)$. Then, $\bar{x}$ satisfies the following for any $\eta > 0$:*

$$\bar{x} = \mathcal{P}_{K(\bar{x})}(\bar{x} - \eta F(\bar{x})). \tag{2}$$

Next, we first increase the sample size at each iteration, then use a constant mini-batch to demonstrate linear convergence and determine the oracle complexity.

Theorem 4 (Increasing sample-size) *Let $\{x_k\}_{k\geq 0}$ be the iterates generated by Algorithm 1 using step-size $\eta > 0$ satisfying $|\eta - \frac{\mu_F}{L^2}| < \frac{\sqrt{\mu_F^2 - L^2(2\gamma - \gamma^2)}}{L^2}$ and retraction parameters $\alpha_k = \bar{\alpha} \in (0, 1)$ and $b_k = \bar{b} \in (0, \frac{1}{1-\beta})$ for $k \geq 0$, where $\beta \triangleq \gamma + \sqrt{1 + L^2\eta^2 - 2\eta\mu_F}$. Suppose Assumptions 1-3 hold, by selecting the number of inner steps for algorithm $\mathcal{M}$ as $t_k = \frac{(k+1)\log^2(k+2)}{\rho^k}$ and choosing the number of sample sizes at iteration k as $N_k = \lceil \rho^{-2k} \rceil$ where $\rho > 1 - q$, we obtain:*

(i) For any $T \geq 1$, $\mathbb{E}[\|x_T - \bar{x}_T\|] \leq \mathcal{O}(\rho^T)$.

(ii) An ϵ-solution x_T, i.e., $\mathbb{E}[\|x_T - \bar{x}_T\|] \leq \epsilon$, can be achieved within $T = \mathcal{O}(\log(1/\epsilon))$ iterations which requires $\sum_{k=0}^{T-1} N_k \geq \mathcal{O}(1/\epsilon^2)$ sample operator evaluations and $\sum_{k=0}^{T-1} t_k = \mathcal{O}(\frac{1}{\epsilon}\log(1/\epsilon))$ number of total inner iterations.

Theorem 5 (Constant mini-batch) *Under premises of Theorem 4, choosing $t_k = \frac{(k+1)\log^2(k+2)}{(1-q)^k}$ and $N_k = N$, then*

(i) For any $T \geq 1$, $\mathbb{E}\left[\|x_T - \bar{x}_T\|\right] \leq \mathcal{O}\left((1-q)^T + \frac{1}{q\sqrt{N}}\right)$.

(ii) Let mini-batch size $N = \mathcal{O}(1/(q^2\epsilon^2))$. An ϵ-solution x_T, i.e., $\mathbb{E}[\|x_T - \bar{x}_T\|] \leq \epsilon$, can be achieved within $T = \mathcal{O}(\frac{1}{q}\log(1/\epsilon))$ iterations which requires $NT = \mathcal{O}(\frac{1}{q^3\epsilon^2}\log(1/\epsilon))$ sample operator evaluations.

5. Numerical Experiment

Over-parameterized Regression Game. In a regression problem, the goal is to find a parameter vector $x \in \mathbb{R}^d$ that minimizes the loss function $\ell^{\text{tr}}(x)$ over the training dataset D^{tr}. Without explicit regularization, an over-parameterized regression problem exhibits multiple global minima over the training dataset, and not all optimal regression coefficients perform equally well. Considering a secondary objective, such as minimizing the loss over a validation set D^{val}, helps in selecting a model parameter that performs well on both training and validation datasets.

Consider a collection of N players each having a model parameter $x_i \in \mathbb{R}^d$. Define $\mathbf{x} \triangleq [x_i]_{i=1}^N$, and suppose there is a shared training dataset D^{tr} and each player possesses an individual validation

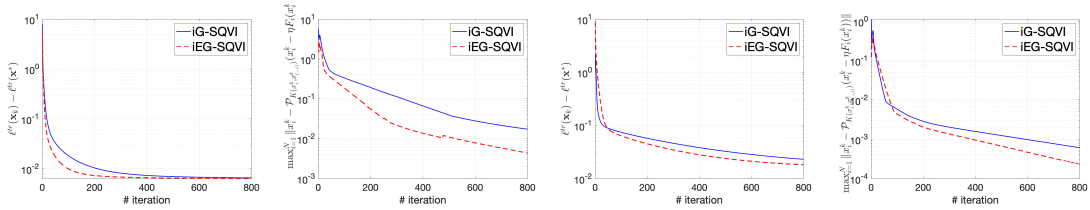

Figure 1: The two figures on the left are for the `triazines` dataset (180 data points, 60 features), and the two on the right are for the `eunite2001` dataset (320 data points, 16 features).

dataset D_i^{val}. The goal is to find a model parameter $\mathbf{x}$ by minimizing the training loss $\ell^{\text{tr}}(x)$ while each player improves its model parameter based on their validation set by minimizing $\ell_i^{\text{val}}(x_i)$. This problem can be formulated as bilevel GNE:

$$\min_{x_i \in \mathbb{R}^d} \ell_i^{\text{val}}(x_i), \quad x_i \in \text{argmin}_{x_i \in K_i} \ell^{\text{tr}}\left(x_i, x^*_{(-i)}\right).$$

In this experiment, we define $\ell_i^{\text{val}}(x_i) \triangleq \frac{1}{2}\|A_i^{\text{val}} x_i - b_i^{\text{val}}\|^2$, where $A_i^{\text{val}} \in \mathbb{R}^{n \times d}$ and $b_i^{\text{val}} \in \mathbb{R}^{n \times 1}$, and $\ell^{\text{tr}}(x) \triangleq \frac{1}{2}\|A^{\text{tr}}\mathbf{x} - b^{\text{tr}}\|^2$, where $A^{\text{tr}} \in \mathbb{R}^{Nn \times Nd}$ and $b^{\text{tr}} \in \mathbb{R}^{Nn \times 1}$ and $X_i = \{x_i \mid \|x_i\| \leq \lambda\}$ for some $\lambda > 0$.

One can show that this problem can be formulated as (**SQVI**) by choosing

$$K(x) = \Pi_{i=1}^N K_i(x_i, x_{(-i)}), \text{ where } K_i(x_i, x_{(-i)}) = \text{argmin}_{x_i \in X_i} \frac{1}{2}\left\|A_i^{\text{tr}} x_i + A_{(-i)}^{\text{tr}} x_{(-i)} - b_i^{\text{tr}}\right\|^2,$$

$$F(x) = [F_i(x_i)]_{i=1}^N \text{ where } F_i(x_i) = (A_i^{\text{val}})^T (A_i^{\text{val}} x_i - b_i^{\text{val}}).$$

Note that the operator F is monotone and satisfies the quadratic growth property, but it may not be strongly monotone. Since no existing methods address the non-strongly monotone setting, we implemented only our proposed extra-gradient method and its gradient variant in the numerical results. Figure 1 compares the suboptimality of the lower-level problem and the gap function based on the optimality condition (2). More details, including parameter choices, datasets, and additional synthetic dataset results, are in the appendix.

References

[1] Zeinab Alizadeh, Brianna M Otero, and Afrooz Jalilzadeh. An inexact variance-reduced method for stochastic quasi-variational inequality problems with an application in healthcare. In *2022 Winter Simulation Conference (WSC)*, pages 3099–3109. IEEE, 2022.

[2] Zeinab Alizadeh, Afrooz Jalilzadeh, and Farzad Yousefian. Randomized lagrangian stochastic approximation for large-scale constrained stochastic nash games. *Optimization Letters*, 18(2): 377–401, 2024.

[3] Anatoly Sergeevich Antipin, N Mijajlović, and M Jaćimović. A second-order iterative method for solving quasi-variational inequalities. *Computational Mathematics and Mathematical Physics*, 53(3):258, 2013.

[4] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[5] DP Bertsekas, A Nedić, and A Ozdaglar. *Convex analysis and optimization, ser*, volume 1. Athena Scientific, 2003.

[6] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[7] Sayantan Choudhury, Eduard Gorbunov, and Nicolas Loizou. Single-call stochastic extragradient methods for structured non-monotone variational inequalities: Improved analysis under weaker conditions. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Francisco Facchinei and Christian Kanzow. Generalized nash equilibrium problems. *Annals of Operations Research*, 175(1):177–211, 2010.

[9] Francisco Facchinei and Jong-Shi Pang. *Finite-dimensional variational inequalities and complementarity problems*. Springer Science & Business Media, 2007.

[10] Francisco Facchinei, Christian Kanzow, and Simone Sagratella. Solving quasi-variational inequalities via their kkt conditions. *Mathematical Programming*, 144(1):369–412, 2014.

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

[12] Gul Gurkan, A Yonca Ozge, and Stephen M Robinson. Sample-path solution of stochastic variational inequalities, with applications to option pricing. In J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, editors, *Proceedings Winter Simulation Conference*, pages 337–344. IEEE, 1996.

[13] Erfan Yazdandoost Hamedani and Necdet Serhat Aybat. A primal-dual algorithm with line search for general convex-concave saddle point problems. *SIAM Journal on Optimization*, 31 (2):1299–1329, 2021.

[14] Niao He, Anatoli Juditsky, and Arkadi Nemirovski. Mirror prox algorithm for multi-term composite minimization and semi-separable problems. *Computational Optimization and Applications*, 61(2):275–319, 2015.

[15] Afrooz Jalilzadeh and Uday V Shanbhag. A proximal-point algorithm with variable sample-sizes (PPAWSS) for monotone stochastic variational inequality problems. In N Mustafee, K.-H. G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, editors, *2019 Winter Simulation Conference (WSC)*, pages 3551–3562. IEEE, 2019.

[16] Ruichen Jiang, Nazanin Abolfazli, Aryan Mokhtari, and Erfan Yazdandoost Hamedani. A conditional gradient-based method for simple bilevel optimization with convex lower-level problem. In *International Conference on Artificial Intelligence and Statistics*, pages 10305–10323. PMLR, 2023.

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

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

[19] Suad Krilašević and Sergio Grammatico. Learning generalized nash equilibria in monotone games: A hybrid adaptive extremum seeking control approach. *Automatica*, 151:110931, 2023.

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

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

[22] Yura Malitsky. Proximal extrapolated gradient methods for variational inequalities. *Optimization Methods and Software*, 33(1):140–164, 2018.

[23] 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.

[24] N Mijajlović and M Jacimović. A proximal method for solving quasi-variational inequalities. *Computational Mathematics and Mathematical Physics*, 55(12):1981, 2015.

[25] Nevena Mijajlović, Milojica Jaćimović, and Muhammad Aslam Noor. Gradient-type projection methods for quasi-variational inequalities. *Optimization Letters*, 13(8):1885–1896, 2019.

[26] Ion Necoara, Yu Nesterov, and Francois Glineur. Linear convergence of first order methods for non-strongly convex optimization. *Mathematical Programming*, 175:69–107, 2019.

[27] 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. doi: 10.1137/S1052623403425629.

[28] 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.

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

[30] Yurii Nesterov and Laura Scrimali. Solving strongly monotone variational and quasi-variational inequalities. *Discrete and Continuous Dynamical Systems*, 31(4):1383–1396, 2011.

[31] M Aslam Noor and Werner Oettli. On general nonlinear complementarity problems and quasi-equilibria. *Le Matematiche*, 49(2):313–331, 1994.

[32] Muhammad Aslam Noor. New approximation schemes for general variational inequalities. *Journal of Mathematical Analysis and applications*, 251(1):217–229, 2000.

[33] Muhammad Aslam Noor. Existence results for quasi variational inequalities. *Banach Journal of Mathematical Analysis*, 1(2):186–194, 2007.

[34] Jong-Shi Pang and Masao Fukushima. Quasi-variational inequalities, generalized nash equilibria, and multi-leader-follower games. *Computational Management Science*, 2(1):21–56, 2005.

[35] Uma Ravat and Uday V Shanbhag. On the existence of solutions to stochastic quasi-variational inequality and complementarity problems. *Mathematical Programming*, 165(1):291–330, 2017.

[36] Irina Prokof'evna Ryazantseva. First-order methods for certain quasi-variational inequalities in a hilbert space. *Computational Mathematics and Mathematical Physics*, 47(2):183–190, 2007.

[37] Hanif Salahuddin. Projection methods for quasi-variational inequalities. *Mathematical and Computational Applications*, 9(2):125–131, 2004.

[38] Sepideh Samadi, Daniel Burbano, and Farzad Yousefian. Achieving optimal complexity guarantees for a class of bilevel convex optimization problems. *arXiv preprint arXiv:2310.12247*, 2023.

[39] Ioannis Tsaknakis, Mingyi Hong, and Shuzhong Zhang. Minimax problems with coupled linear constraints: Computational complexity and duality. *SIAM Journal on Optimization*, 33 (4):2675–2702, 2023.

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

Appendix A.

The following lemma is essential for analyzing the convergence rate.

Lemma 6 *[5] Let $X \subseteq \mathbb{R}^n$ be a nonempty closed and convex set. Then the following hold: (a) $\|\mathcal{P}_X[u] - \mathcal{P}_X[v]\| \leq \|u - v\|$ for all $u, v \in \mathbb{R}^n$; (b) $(\mathcal{P}_X[u] - u)^T(x - \mathcal{P}_X[u]) \geq 0$ for all $u \in \mathbb{R}^n$ and $x \in X$.*

Next, we provide the proof of Lemma 3 stated in the main body of the paper.

Proof of Lemma 3. Note that $\bar{x} \in X^*$ implies that for any $x \in K(\bar{x})$ we have that $\langle F(\bar{x}), x - \bar{x}\rangle \geq 0$. Multiplying both sides of the last inequality by $\eta > 0$ we obtain that for any $x \in K(\bar{x})$, $\langle \eta F(\bar{x}), x - \bar{x}\rangle \geq 0$ which due to convexity of set $K(\bar{x})$ is equivalent to $\bar{x} = \operatorname{argmin}_{x \in K(\bar{x})} \|x - (\bar{x} - \eta F(\bar{x}))\|^2 = \mathcal{P}_{K(\bar{x})}(\bar{x} - \eta F(\bar{x}))$. □

Before stating our main results, we need to define a few notations to facilitate the rate results.

Definition 7 *At each iteration of $k \geq 0$, we define the error of sample operator F as $\bar{w}_{k,N_k} \triangleq \frac{1}{N_k}\sum_{j=1}^{N_k}(G(x_k,\xi_{j,k}) - F(x_k))$, and $\bar{w}'_{k,N_k} \triangleq \frac{1}{N_k}\sum_{j=1}^{N_k}(G(u_k,\xi_{j,k}) - F(u_k))$. Moreover, the error of approximating the projection is defined by $e_k \triangleq d_k - \mathcal{P}_{K(x_k)}\left(x_k - \eta\frac{\sum_{j=1}^{N_k} G(x_k,\xi_{j,k})}{N_k}\right)$ and $e'_k \triangleq s_k - \mathcal{P}_{K(u_k)}\left(u_k - \eta\frac{\sum_{j=1}^{N_k} G(u_k,\xi_{j,k})}{N_k}\right)$ in step (1) and (3) of Algorithm 1, respectively.*

In the next theorem, we establish a bound on the expected solution error, which is expressed in terms of errors associated with the sample operator and the projection approximations. Subsequently, in Theorem 4, we provide the rate and complexity statements for Algorithm 1.

Theorem 8 *Let $\{x_k\}_{k\geq 0}$ be the iterates generated by Algorithm 1 using step-size $\eta > 0$ satisfying $|\eta - \frac{\mu_F}{L^2}| < \frac{\sqrt{\mu_F^2 - L^2(2\gamma-\gamma^2)}}{L^2}$ and retraction parameters $\alpha_k = \bar{\alpha} \in (0,1)$ and $b_k = \bar{b} \in (0, \frac{1}{1-\beta})$ for $k \geq 0$, where $\beta \triangleq \gamma + \sqrt{1 + L^2\eta^2 - 2\eta\mu_F}$. Suppose Assumptions 1-3 hold, then for any $T \geq 1$ we have that*

$$\begin{aligned}\|x_T - \bar{x}_T\| \leq (1-q)^T\|x_0 - \bar{x}_0\| + \bar{\alpha}\beta\bar{b}\sum_{k=0}^{T-1}(1-q)^{T-k-1}\left(\|e'_k\| + \eta\|\bar{w}'_{k,N_k}\|\right)\\ + \bar{\alpha}\sum_{k=0}^{T-1}(1-q)^{T-k-1}\left(\|e_k\| + \eta\|\bar{w}_{k,N_k}\|\right),\end{aligned} \tag{3}$$

where $q \triangleq \bar{\alpha}(1-\beta)(1+\beta\bar{b}) \in (0,1)$.

Proof For any $k \geq 0$, we define $\bar{x}_k \triangleq \mathcal{P}_{X^*}(x_k)$ where X^* denotes the set of optimal solutions of problem (**SQVI**). From Lemma 3 we conclude that $\bar{x}_k = \mathcal{P}_{K(\bar{x}_k)}[\bar{x}_k - \eta F(\bar{x}_k)]$. Using the update rule of x_{k+1} in Algorithm 1 and the fact that e_k denotes the error of computing the projection operator, we obtain the following.

$$\begin{aligned}\|x_{k+1} - \bar{x}_k\| &= \|(1-\alpha_k)x_k + \alpha_k\mathcal{P}_{K(u_k)}\left[u_k - \eta(F(u_k) + \bar{w}_{k,N_k})\right]\\ &\quad + \alpha_k e_k - (1-\alpha_k)\bar{x}_k - \alpha_k\mathcal{P}_{K(\bar{x}_k)}\left[\bar{x}_k - \eta F(\bar{x}_k)\right]\|\\ &= \|(1-\alpha_k)x_k + \alpha_k\mathcal{P}_{K(u_k)}\left[u_k - \eta(F(u_k) + \bar{w}_{k,N_k})\right]\\ &\quad + \alpha_k e_k - (1-\alpha_k)\bar{x}_k - \alpha_k\mathcal{P}_{K(\bar{x}_k)}\left[\bar{x}_k - \eta F(\bar{x}_k)\right]\\ &\quad \pm \alpha_k\mathcal{P}_{K(x_k)}\left[u_k - \eta(F(u_k) + \bar{w}_{k,N_k})\right]\|\\ &\leq (1-\alpha_k)\|x_k - \bar{x}_k\|\\ &\quad + \alpha_k\|\mathcal{P}_{K(u_k)}\left[u_k - \eta(F(u_k) + \bar{w}_{k,N_k})\right] - \mathcal{P}_{K(\bar{x}_k)}\left[u_k - \eta(F(u_k) + \bar{w}_{k,N_k})\right]\|\\ &\quad + \alpha_k\|\mathcal{P}_{K(\bar{x}_k)}\left[u_k - \eta(F(u_k) + \bar{w}_{k,N_k})\right] - \mathcal{P}_{K(\bar{x}_k)}\left[\bar{x}_k - \eta F(\bar{x}_k)\right]\| + \alpha_k\|e_k\|\\ &\leq (1-\alpha_k)\|(x_k - \bar{x}_k)\| + \alpha_k\gamma\|u_k - \bar{x}_k\|\\ &\quad + \alpha_k\underbrace{\|u_k - \bar{x}_k - \eta(F(u_k) - F(\bar{x}_k))\|}_{\text{term (a)}} + \alpha_k\eta\|\bar{w}_{k,N_k}\| + \alpha_k\|e_k\|,\end{aligned} \tag{4}$$

where the first inequality follows from the triangle inequality, and in the last inequality, we used Lemma 6-(a) and Assumption 3. Next, we provide an upper bound for the term (a) in (4) by using

Definition 1 and Lipschitz continuity of operator F as follows

$$\begin{aligned}(\text{term (a)})^2 &= \|u_k - \bar{x}_k\|^2 + \eta^2\|F(u_k) - F(\bar{x}_k)\|^2 - 2\eta\langle u_k - \bar{x}_k, F(u_k) - F(\bar{x}_k)\rangle \\ &\leq (1 + L^2\eta^2 - 2\eta\mu_F)\|u_k - \bar{x}_k\|^2 \\ \implies \text{term(a)} &\leq \sqrt{1 + L^2\eta^2 - 2\eta\mu_F}\|u_k - \bar{x}_k\|. \end{aligned} \tag{5}$$

Combining (4) and (5), and defining $\beta \triangleq \gamma + \sqrt{1 + L^2\eta^2 - 2\eta\mu_F}$ we obtain

$$\|x_{k+1} - \bar{x}_k\| \leq (1 - \alpha_k)\|x_k - \bar{x}_k\| + \alpha_k\beta\|u_k - \bar{x}_k\| + \alpha_k\eta\|\bar{w}_{k,N_k}\| + \alpha_k\|e_k\|. \tag{6}$$

Next, we turn our attention to providing an upper bound for $\|u_k - \bar{x}_k\|$. In particular, using the update of u_k in Algorithm 1 by taking similar steps as (4) and (5), one can obtain:

$$\begin{aligned}\|u_k - \bar{x}_k\| &= \|(1 - b_k)x_k + b_k\mathcal{P}_{K(x_k)}\left[x_k - \eta(F(x_k) + \bar{w}'_{k,N_k})\right] \\ &\quad + b_k e'_k - (1 - b_k)\bar{x}_k - b_k\mathcal{P}_{K(\bar{x}_k)}\left[\bar{x}_k - \eta F(\bar{x}_k)\right]\| \\ &\leq (1 - b_k)\|x_k - \bar{x}_k\| + b_k\beta\|x_k - \bar{x}_k\| + b_k\eta\|\bar{w}'_{k,N_k}\| + b_k\|e'_k\| \\ &= (1 - b_k(1 - \beta))\|x_k - \bar{x}_k\| + b_k\eta\|\bar{w}'_{k,N_k}\| + b_k\|e'_k\|. \end{aligned}$$

Replacing the above inequality in (6), and defining $q_i \triangleq \alpha_i(1 - \beta)(1 + \beta b_i)$ we conclude that

$$\begin{aligned}&\|x_{k+1} - \bar{x}_k\| \\ &\leq (1 - \alpha_k)\|x_k - \bar{x}_k\| + \alpha_k\beta((1 - b_k(1 - \beta))\|x_k - \bar{x}_k\| + b_k\eta\|\bar{w}'_{k,N_k}\| + b_k\|e'_k\|) + \alpha_k\eta\|\bar{w}_{k,N_k}\| \\ &\quad + \alpha_k\|e_k\| \\ &= (1 - \alpha_k(1 - \beta)(1 + \beta b_k))\|x_k - \bar{x}_k\| + \alpha_k\eta\beta b_k\|\bar{w}'_{k,N_k}\| + \alpha_k\beta b_k\|e'_k\| + \alpha_k\eta\|\bar{w}_{k,N_k}\| + \alpha_k\|e_k\|. \end{aligned}$$

Now, from the fact that $\bar{x}_{k+1} = \mathcal{P}_{X^*}(x_{k+1})$ one can conclude that $\|x_{k+1} - \bar{x}_{k+1}\| \leq \|x_{k+1} - \bar{x}_k\|$. Therefore, for any $k \geq 0$

$$\begin{aligned}\|x_{k+1} - \bar{x}_{k+1}\| \leq \prod_{i=0}^{k}(1 - q_i)\|x_0 - \bar{x}_0\| &+ \sum_{i=0}^{k}\left(\left(\prod_{j=i}^{k-1}(1 - q_j)\right)\alpha_i\beta b_i\left(\eta\|\bar{w}'_{i,N_i}\| + \|e'_i\|\right)\right) \\ &+ \sum_{i=0}^{k}\left(\left(\prod_{j=i}^{k-1}(1 - q_j)\right)\alpha_i\left(\eta\|\bar{w}_{i,N_i}\| + \|e_i\|\right)\right), \end{aligned}$$

where we assume that the product is 1 when there are no terms in the multiplication, i.e., $\prod_{j=i}^{k-1}(1 - q_{j+1}) = 1$ if $i > k - 1$.

From the condition of η, we have that $\beta < 1$. Moreover, choosing $b_k = \bar{b} < \frac{1}{1-\beta}$ and $\alpha_k = \bar{\alpha} < 1$ one can readily verify that $q_k = q = \bar{\alpha}(1 - \beta)(1 + \beta\bar{b}) < 1$ for all $k \geq 0$. Therefore, the result immediately follows by using the fact that $\prod_{j=i}^{k-1}(1 - q) = (1 - q)^{k-i}$. ■

Proof of Theorem 4. (i) Taking expectation from both sides of (3), choosing $N_k = \lceil\rho^{-2k}\rceil$, and using Assumption 2, one can obtain:

$$\mathbb{E}\left[\|x_T - \bar{x}_T\|\right] \leq (1 - q)^T\|x_0 - \bar{x}_0\| + \bar{\alpha}\beta\bar{b}(1 - q)^{T-1}\sum_{k=0}^{T-1}\left(\eta\nu'(\frac{\rho}{1 - q})^k + \mathbb{E}[\|e'_k\|](1 - q)^{-k}\right)$$

$$+\bar{\alpha}(1-q)^{T-1}\sum_{k=0}^{T-1}\left(\eta\nu(\frac{\rho}{1-q})^k+\mathbb{E}[\|e_k\|](1-q)^{-k}\right).$$

Using the fact that $\sum_{k=0}^{T-1}(\frac{\rho}{1-q})^k=\frac{1-(\frac{\rho}{1-q})^T}{1-\frac{\rho}{1-q}}$, we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\|x_T-\bar{x}_T\|\right]\\
&\leq(1-q)^T\|x_0-\bar{x}_0\|+\bar{\alpha}\beta\bar{b}\eta\nu'\frac{\rho^T-(1-q)^T}{\rho+q-1}+\bar{\alpha}\beta\bar{b}(1-q)^{T-1}\sum_{k=0}^{T-1}\left(\mathbb{E}[\|e'_k\|](1-q)^{-k}\right)\\
&\quad+\bar{\alpha}\eta\nu\frac{\rho^T-(1-q)^T}{\rho+q-1}+\bar{\alpha}(1-q)^{T-1}\sum_{k=0}^{T-1}\left(\mathbb{E}[\|e_k\|](1-q)^{-k}\right).
\end{aligned}$$

Since $\rho\geq 1-q$ and $q\in(0,1)$, one can easily confirm that $-\frac{(1-q)^T}{\rho+q-1}<0$, hence the following holds.

$$\begin{aligned}
\mathbb{E}\left[\|x_T-\bar{x}_T\|\right]&\leq(1-q)^T\|x_0-\bar{x}_0\|+\frac{\bar{\alpha}\beta\bar{b}\eta\nu'\rho^T}{\rho+q-1}+\bar{\alpha}\beta\bar{b}\sum_{k=0}^{T-1}\left(\mathbb{E}[\|e'_k\|](1-q)^{T-1-k}\right)\\
&\quad+\frac{\bar{\alpha}\eta\nu\rho^T}{\rho+q-1}+\bar{\alpha}\sum_{k=0}^{T-1}\left(\mathbb{E}[\|e_k\|](1-q)^{T-1-k}\right)\\
&\leq\rho^T\|x_0-\bar{x}_0\|+\frac{\bar{\alpha}\beta\bar{b}\eta\nu'\rho^T}{\rho+q-1}+\bar{\alpha}\beta\bar{b}\sum_{k=0}^{T-1}\left(\mathbb{E}[\|e'_k\|]\rho^{T-1-k}\right)+\frac{\bar{\alpha}\eta\nu\rho^T}{\rho+q-1}\\
&\quad+\bar{\alpha}\sum_{k=0}^{T-1}\left(\mathbb{E}[\|e_k\|]\rho^{T-1-k}\right).
\end{aligned}$$

According to the Assumption 4, Algorithm $\mathcal{M}$ has a convergence rate of C/t_k^2 within t_k inner steps. By selecting $t_k=\frac{(k+1)\log^2(k+2)}{\rho^k}$, we have that $\mathbb{E}[\|e_k\|]\leq\frac{C}{t_k}=\frac{C\rho^k}{(k+1)\log^2(k+2)}$ and $\mathbb{E}[\|e'_k\|]\leq\frac{C'\rho^k}{(k+1)\log^2(k+2)}$. These upper bounds are independent of x_k, so by using the tower property of expectation in the previous inequality, we obtain the following.

$$\begin{aligned}
\mathbb{E}\left[\|x_T-\bar{x}_T\|\right]&\leq\rho^T\|x_0-\bar{x}_0\|+\frac{\bar{\alpha}\beta\bar{b}\eta\nu'\rho^T}{\rho+q-1}+\bar{\alpha}\beta\bar{b}C'\rho^{T-1}\sum_{k=0}^{T-1}\frac{1}{(k+1)\log^2(k+2)}\\
&\quad+\frac{\bar{\alpha}\eta\nu\rho^T}{\rho+q-1}+\bar{\alpha}C\rho^{T-1}\sum_{k=0}^{T-1}\frac{1}{(k+1)\log^2(k+2)}.
\end{aligned}$$

By applying the Cauchy-Schwarz inequality, using the fact that $D\triangleq\sum_{k=0}^{\infty}\frac{1}{(k+1)\log^2(k+2)}\leq 3.39$ and rearranging the terms, the desired result is obtained:

$$\mathbb{E}\left[\|x_T-\bar{x}_T\|\right]\leq\rho^T\|x_0-\bar{x}_0\|+\rho^T\bar{\alpha}\eta(\tfrac{\beta\bar{b}\nu'+\nu}{\rho+q-1})+\rho^{T-1}\bar{\alpha}D(\beta\bar{b}C'+C). \tag{7}$$

(ii) To compute an ϵ-solution, i.e., $\mathbb{E}[\|x_T - \bar{x}_T\|] \leq \epsilon$, it follows from (7) that $T = \log_{1/\rho}(\bar{D}/\epsilon)$ iterations is required, where $\bar{D} = \|x_0 - \bar{x}_0\| + \bar{\alpha}\eta(\frac{\beta\bar{b}\nu'+\nu}{\rho+q-1}) + \frac{\bar{\alpha}D(\beta\bar{b}C'+C)}{\rho}$. Moreover, in Algorithm 1, each iteration requires taking $t_k = \frac{(k+1)\log^2(k+2)}{\rho^k}$ inner steps of Algorithm $\mathcal{M}$. Therefore, the total number of inner iterations is

$$\sum_{k=0}^{T-1} t_k = \sum_{k=0}^{T-1} \frac{(k+1)\log^2(k+2)}{\rho^k} \leq T\log^2(T+1)\frac{(1/\rho)^T}{1/\rho-1} = \log_{1/\rho}\bar{D}/\epsilon.$$

Furthermore, the total number of sample operator evaluations can be obtained as follows:

$$\sum_{k=0}^{T-1} N_k = \sum_{k=0}^{T-1} \lceil \rho^{-2k} \rceil \geq \frac{\rho^2}{1-\rho^2}\left(\frac{\bar{D}^2}{\epsilon^2} - 1\right). \square$$

Remark 9 *The error bound derived in Theorem 8 and Theorem 4 signify convergence rates concerning the error associated with the projection operator. To be specific, in Theorem 4, we characterized how quickly this error must decrease to ensure linear convergence. Conversely, in cases where the projection onto the constraint set is straightforward to compute, i.e., when $e_k = e'_k = 0$ for all $k \geq 0$, and under the assumptions of Theorem 8 the expectation of solution error will be bound as follows:*

$$\begin{aligned}&\|x_T - \bar{x}_T\| \\ &\leq (1-q)^T\|x_0 - \bar{x}_0\| + \bar{\alpha}\beta\bar{b}\eta\sum_{k=0}^{T-1}(1-q)^{T-k-1}\|\bar{w}'_{k,N_k}\| + \bar{\alpha}\eta\sum_{k=0}^{T-1}(1-q)^{T-k-1}\|\bar{w}_{k,N_k}\|.\end{aligned}$$

By choosing $N_k = \lceil \rho^{-2k} \rceil$ where $\rho > 1 - q$, Algorithm 1 achieves a linear convergence rate, i.e., $\mathbb{E}[\|x_T - \bar{x}_T\|] \leq \mathcal{O}(\rho^T)$.

Proof of Theorem 5. (i) By taking expectation from both sides of 3, choosing $N_k = N$, and using Assumption 2, the following holds.

$$\begin{aligned}\mathbb{E}\left[\|x_T - \bar{x}_T\|\right] \leq (1-q)^T\|x_0 - \bar{x}_0\| &+ \bar{\alpha}\beta\bar{b}(1-q)^{T-1}\sum_{k=0}^{T-1}\left(\frac{\eta\nu'}{(1-q)^k\sqrt{N}} + \mathbb{E}[\|e'_k\|](1-q)^{-k}\right) \\ &+ \bar{\alpha}(1-q)^{T-1}\sum_{k=0}^{T-1}\left(\frac{\eta\nu}{(1-q)^k\sqrt{N}} + \mathbb{E}[\|e_k\|](1-q)^{-k}\right).\end{aligned}$$

Following the similar steps as in the proof of Theorem 4, and defining $D \triangleq \sum_{k=0}^{\infty}\frac{1}{(k+1)\log^2(k+2)} \leq 3.39$, the following can be obtained.

$$\begin{aligned}&\mathbb{E}\left[\|x_T - \bar{x}_T\|\right] \\ &\leq (1-q)^T\|x_0 - \bar{x}_0\| + \frac{\bar{\alpha}\beta\bar{b}\eta\nu'}{q\sqrt{N}} + \frac{\bar{\alpha}\eta\nu}{q\sqrt{N}} + \bar{\alpha}\beta\bar{b}C'D(1-q)^{T-1} + \bar{\alpha}CD(1-q)^{T-1}.\end{aligned}$$

Now by rearranging the terms, we obtain the desired result:

$$\mathbb{E}\left[\|x_T - \bar{x}_T\|\right] \leq (1-q)^T \|x_0 - \bar{x}_0\| + (1-q)^{T-1}\bar{\alpha}D(\beta\bar{b}C' + C) + \frac{\bar{\alpha}\eta}{q\sqrt{N}}(\beta\bar{b}\nu' + \nu). \quad (8)$$

(ii) Let $T = \log_{1/(1-q)}(2\bar{D}/\epsilon)$, $N = \frac{4\bar{C}^2}{q^2\epsilon^2}$, and define $\bar{D} \triangleq \|x_0 - \bar{x}_0\| + \frac{\bar{\alpha}D(\beta\bar{b}C'+C)}{(1-q)}$ and $\bar{C} \triangleq \bar{\alpha}\eta(\beta\bar{b}\nu' + \nu)$, then from 8 we have that:

$$\begin{aligned}\mathbb{E}\left[\|x_T - \bar{x}_T\|\right] &\leq (1-q)^T\left(\|x_0 - \bar{x}_0\| + (1-q)\bar{\alpha}D(\beta\bar{b}C' + C)\right) + \frac{\bar{\alpha}\eta}{q\sqrt{N}}(\beta\bar{b}\nu' + \nu) \\ &\leq (1-q)^T\bar{D} + \frac{\bar{C}}{q\sqrt{N}} \leq \epsilon,\end{aligned}$$

where in the last inequality we used the definition of T and N.□

Numerical Experiment. We run our experiment on different datasets and compare the inexact extra gradient approach with its gradient-based variant, i.e., letting retraction parameter $b_k = 0$ in Algorithm 1. In particular, we propose an inexact Gradient SQVI (iG-SQVI) method in Algorithm 2.

Algorithm 2 inexact Gradient SQVI (iG-SQVI)

Input: $x_0 \in X$, $\eta > 0$, $\{N_k\}_k$, $\{t_k\}_k$, $\{\alpha_k\}_k$ and Algorithm $\mathcal{M}$ satisfying Assumption 4;
for $k = 0, \ldots T-1$ **do**
(1) Use Algorithm $\mathcal{M}$ with t_k iterations and find an approximated solution d_k of

$$\min_{x \in K(x_k)} \left\| x - \left(x_k - \eta \frac{\sum_{j=1}^{N_k} G(x_k, \xi_{j,k})}{N_k} \right) \right\|^2;$$

(2) $x_{k+1} = (1-\alpha_k)x_k + \alpha_k d_k$;
end for
Output: x_{k+1};

Table 2: Parameter settings after fine-tuning for the algorithms across all datasets

	`triazines`	`eunite2001`	`synthetic`
Stepsize η	5e-2	3e-1	1e-2
$\bar{\alpha}$	1e-1	5e-1	9e-1
$\bar{b}$	1e-1	5e-1	12e-1
Regularizer	1e0	1e-1	1e-2

In Figure 2, we present a performance comparison of our proposed methods. For the `triazines` dataset [6], we set the number of players $N = 6$, for the `eunite2001` dataset [6] $N = 4$, and for the `synthetic` dataset $N = 10$. In all cases, we utilized 80% of the data points for training and allocated the remaining 20% for validation. To solve the projection inexactly, observe that the sub-problem is a simple bilevel optimization problem. This type of problem has been explored in

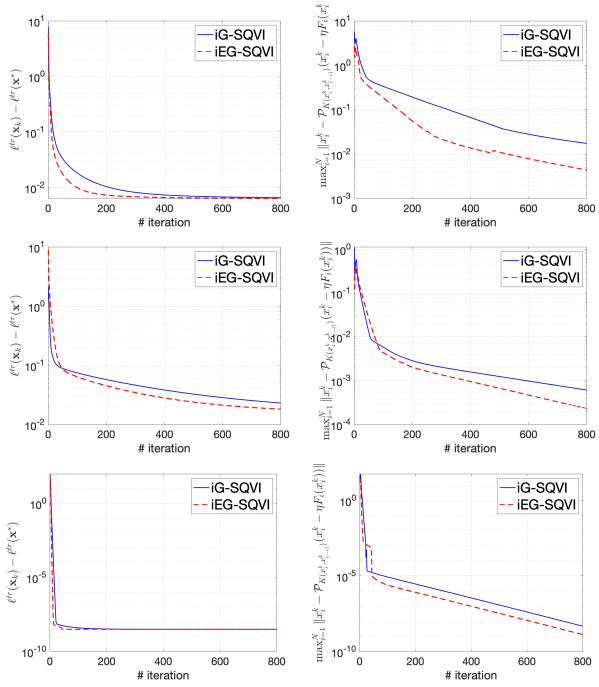

Figure 2: Comparison of iG-SQVI and iEG-SQVI: (Top) for the `triazines` dataset with 180 data points and 60 features, (Middle) for the `eunite2001` dataset with 320 data points and 16 features, (Bottom) for a `synthetic` dataset with 250 data points and 25 features.

the literature [16, 38]. Here, following [38], we employed the FISTA algorithm [4] to solve the corresponding regularized problem satisfying Assumption 4. For all the experiments, we execute the inner algorithm for $k \log^2(k+1)(1-\text{1e-3})^k$ iterations, and the remaining parameters are selected according to the following table after fine-tuning.

In Figure 2, on the left, we compared the suboptimality of the lower-level problem, and on the right, we compared the gap function based on the optimality condition (2). It is evident that both methods converge to the optimal solution. Notably, iEG-SQVI demonstrates a slightly better performance due to a smaller convergence rate factor.