Noname manuscript No.
(will be inserted by the editor)

Inexact Stochastic Subgradient Projection Method for Stochastic Equilibrium Problems with Nonmonotone Bifunctions: Application to Expected Risk Minimization in Machine Learning

Hideaki Iiduka

Received: date / Accepted: date

Abstract This paper discusses a stochastic equilibrium problem for which the function is in the form of the expectation of nonmonotone bifunctions and the constraint set is closed and convex. This problem includes various applications such as stochastic variational inequalities, stochastic Nash equilibrium problems, and nonconvex stochastic optimization problems. For solving this stochastic equilibrium problem, we propose an inexact stochastic subgradient projection method. The proposed method sets a random realization of the bifunction and then updates its approximation by using both its stochastic subgradient and the projection onto the constraint set. The main contribution of this paper is to present a convergence analysis showing that, under certain assumptions, any accumulation point of the sequence generated by the proposed method using a constant step size almost surely belongs to the solution set of the stochastic equilibrium problem. A convergence rate analysis of the method is also provided to illustrate the method's efficiency. Another contribution of this paper is to show that a machine learning algorithm based on the proposed method achieves the expected risk minimization for a class of least absolute selection and shrinkage operator (lasso) problems in statistical learning with sparsity. Numerical comparisons of the proposed machine learning algorithm with existing machine learning algorithms for the expected risk minimization using LIBSVM datasets demonstrate the effectiveness and superior classification accuracy of the proposed algorithm.

Keywords expected risk minimization · least absolute selection and shrinkage operator · nonconvex stochastic optimization · nonmonotone bifunction ·

This work was supported by JSPS KAKENHI Grant Number JP18K11184.

H. Iiduka
Department of Computer Science, Meiji University 1-1-1 Higashimita, Tama-ku, Kawasaki-shi, Kanagawa 214-8571, Japan
E-mail: iiduka@cs.meiji.ac.jp

stochastic equilibrium problem · stochastic variational inequality · stochastic subgradient projection method

Mathematics Subject Classification (2000) 65K05 · 65K15 · 90C15 · 90C26

1 Introduction

Equilibrium problems [8,9,17,36,37] are known as central topics for continuous optimization from the fact that they include such important problems as complementarity problems [21,22], Nash equilibrium problems [20,46,47], fixed point problems [29,30,64], and variational inequalities [21,22] (see [14, 25,43,45,67] for mathematical programming with (stochastic) equilibrium constraints).

In keeping with the importance of equilibrium problems, many iterative methods have been proposed for solving them. For example, projected subgradient methods [36] were proposed to solve a deterministic equilibrium problem with one nonmonotone bifunction over a simple closed convex constraint set. Proximal point algorithms [6,17,63] solve deterministic equilibrium problems with monotone bifunctions (see (5) for the definition of a monotone bifunction). The algorithm in [6] can be applied to an equilibrium problem for a monotone bifunction over the solution set of an equilibrium problem.

Meanwhile, there are useful methods for various types of stochastic programming. The stochastic approximation (SA) methods [10,27,28,54] and their variations [27,48,49,60,61] can solve convex stochastic optimization problems. The Douglas-Rachford splitting method in [13] can solve two-stage stochastic variational inequalities, and the stochastic accelerated mirror-prox method [15] was proposed to solve a class of monotone stochastic variational inequalities. In [42], three methods were presented for computing confidence intervals for components of the solution to a stochastic variational inequality. The regularized smoothed SA method was proposed in [70] to solve stochastic variational inequalities with monotone and non-Lipschitz continuous mappings. In [35], stochastic extragradient methods were proposed to solve stochastic variational inequalities with pseudomonotone and Lipschitz continuous mappings. In [52,58], best-response schemes were proposed to solve stochastic Nash equilibrium problems.

The present paper considers a *stochastic equilibrium problem* for which the function is in the form of the expectation of *nonmonotone* continuous bifunctions and the constraint set is a closed convex set onto which the projection can be efficiently computed (the constraint set is, for example, a closed ball, an affine subspace, a halfspace, or a hyperslab [4, Chapter 28]). Thanks to the useful results in [36,37], we can show that the stochastic equilibrium problem includes practical applications such as nonconvex stochastic optimization problems, stochastic variational inequalities, stochastic complementarity problems, and stochastic Nash equilibrium problems. Each random realization of the bifunction in the equilibrium problem considered here is allowed to be convex in

the second argument. This implies that stochastic subgradients of the convex function can be efficiently used to solve the equilibrium problem. Therefore, in this paper, we propose a *stochastic subgradient projection method* for solving the equilibrium problem and perform a convergence analysis of the method. Through the convergence analysis, we show that, under certain assumptions, any accumulation point of the sequence generated by the method almost surely belongs to the solution set of the equilibrium problem. The convergence analysis shows the almost sure convergence of the proposed method to the solution to the stochastic equilibrium problem with strictly pseudomonotone bifunctions over a bounded closed convex constraint set.

The proposed method is related to the inexact subgradient projection methods [36] for solving the deterministic equilibrium problem. The methods in [36] use projection methods for solving convex feasibility problems [3] and control sequences based on the remotest set control, the approximately remotest set control, and the most violated constraint control [36, p.304, (a)–(c)]. These control sequences require us to use inexact solutions of certain optimization problems. The framework of the proposed method is based on the SA methods [10,27,28,54] using stochastic subgradients to enable us to consider stochastic programming. Blending a stochastic subgradient update with the most violated constraint control [36] leads to the proposed inexact stochastic subgradient method for solving the stochastic equilibrium problem. Therefore, the proposed method enables the solution of practical stochastic problems such as nonconvex stochastic optimization problems, stochastic variational inequalities, stochastic complementarity problems, and stochastic Nash equilibrium problems, which cannot be solved by the deterministic method [36] (see Subsection 3.1 for the detailed differences between the deterministic method [36] and the proposed stochastic method).

When iterative methods with diminishing step sizes are applied to complicated optimization, the step sizes are approximately zero for a number of iterations, which implies that using diminishing step sizes would not be implementable in practice. Even if the methods with diminishing step sizes were made to work, we would need to empirically select suitable step sizes to increase the convergence speed of the methods. However, it is too difficult to select in advance suitable diminishing step sizes that guarantee sufficiently quick convergence. This is because what step sizes are suitable depends on various factors, such as the number of iterations, the number of dimensions, the shapes of objective functions and constraint sets, and the selection of subgradients (see [34] for iterative methods based on line search to resolve the issue of selecting suitable diminishing step sizes). The advantage of the proposed stochastic subgradient method is that it uses a constant step size rather than a diminishing step size. This, together with the results of our convergence analysis, implies that the proposed method can solve the stochastic equilibrium problem from the viewpoints of both theory and practice. We also determine the rate of convergence of the proposed method to establish its performance.

The second contribution of this paper is to show that a machine learning algorithm based on the proposed inexact stochastic subgradient projection

method can solve the expected risk minimization problems in machine learning. In this paper, focusing on the case of statistical learning with sparsity, we apply the proposed machine learning algorithm to the pseudomonotone stochastic variational inequality in the capped-ℓ^1 norm coupled nonconvex overlapping group least absolute selection and shrinkage operator (lasso) [16, (25)], [71, C.3.1]. We are able to demonstrate through our convergence analysis that the proposed machine learning algorithm can solve the pseudomonotone stochastic variational inequality. We also numerically compare the performance of the machine learning algorithm based on each of three existing methods, the SA method [10,27,28,54], the stochastic extragradient method [35, Algorithm 1], and an existing machine learning algorithm [16, Algorithm 2] called IncrePA-ncvx, with that of the proposed machine learning algorithm for concrete classification problems with LIBSVM datasets [12] and show that the proposed machine learning algorithm has higher classification accuracy than the machine learning algorithms based on the three existing methods.

The remainder of this paper is organized as follows. Section 2 gives the mathematical preliminaries and states the main problem with examples. Section 3 presents convergence and convergence rate analyses of the proposed stochastic subgradient projection method under certain assumptions. Section 4 considers concrete classification problems and numerically compares the behaviors of the proposed and existing machine learning algorithms. Section 5 concludes the paper with a brief summary and mentions future directions of research for improving the proposed method.

2 Mathematical preliminaries

2.1 Notation and definitions

Let $\mathbb{N}$ be the set of zero and all positive integers. Let $\mathbb{R}^N$ be an N-dimensional Euclidean space with inner product $\langle \cdot, \cdot \rangle$ and associated norm $\| \cdot \|$. Let $B(r) \subset \mathbb{R}^N$ denote the closed ball with radius $r > 0$ centered at the origin. Let $\mathbb{P}[X]$ and $\mathbb{E}[X]$ denote respectively the probability and the expectation of a random variable X. For the random process $\xi_0, \xi_1, \ldots$, let $\mathbb{E}[X|\xi_{[n]}]$ denote the conditional expectation of X given $\xi_{[n]} = (\xi_0, \xi_1, \ldots, \xi_n)$. Let $(x_n)_{n\in\mathbb{N}}$ and $(y_n)_{n\in\mathbb{N}}$ be positive real sequences. $\mathcal{O}$ and o denote Landau's symbols; i.e., $y_n = \mathcal{O}(x_n)$ if there exist $c > 0$ and $n_0 \in \mathbb{N}$ such that $y_n \leq cx_n$ for all $n \geq n_0$, and $y_n = o(x_n)$ if, for all $\epsilon > 0$, there exists $n_0 \in \mathbb{N}$ such that $y_n \leq \epsilon x_n$ for all $n \geq n_0$.

Given a nonempty closed convex set $C \subset \mathbb{R}^N$, the metric projection onto C [4, Subchapter 4.2, Chapter 28], denoted by P_C, is defined for all $x \in \mathbb{R}^N$ by $P_C(x) \in C$ and $\|x - P_C(x)\| = \mathrm{d}(x, C) := \inf_{y\in C} \|x - y\|$. For example, the metric projection onto an affine subspace, a half-space, or a hyperslab can be easily computed within a finite number of arithmetic operations [4, Chapter 28]. The mapping P_C satisfies the nonexpansivity condition [4, Proposition

4.8, (4.8)], i.e., for all $x, y \in \mathbb{R}^N$,

$$\|P_C(x) - P_C(y)\| \leq \|x - y\|.$$

Let $S\colon \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ be a set-valued mapping and let $\bar{x} \in \mathbb{R}^N$. S is said to be continuous [55, Definition 5.4] at $\bar{x}$ if both $\limsup_{x\to\bar{x}} S(x) \subset S(\bar{x})$ and $\liminf_{x\to\bar{x}} S(x) \supset S(\bar{x})$ hold, where $\limsup_{x\to\bar{x}} S(x) := \bigcup_{x_n\to\bar{x}} \limsup_{n\to+\infty} S(x_n)$ and $\liminf_{x\to\bar{x}} S(x) := \bigcap_{x_n\to\bar{x}} \liminf_{n\to+\infty} S(x_n)$.

Suppose that $f\colon \mathbb{R}^N \to \mathbb{R}$ is continuous and $\bar{x} \in \mathbb{R}^N$ satisfies $f(\bar{x}) < +\infty$. A point $v \in \mathbb{R}^N$ is said to be a *regular subgradient* [55, Definition 8.3(a)] of f at $\bar{x}$ if

$$v \in \hat{\partial} f(\bar{x}) := \left\{ v \in \mathbb{R}^N \colon f(x) \geq f(\bar{x}) + \langle x - \bar{x}, v\rangle + o(\|x - \bar{x}\|) \ \left(x \in \mathbb{R}^N\right) \right\}.$$

A point $v \in \mathbb{R}^N$ is referred to as a *subgradient* [55, Definition 8.3(b)] of f at $\bar{x}$ if

$$\begin{aligned} &\text{there exist } (x_n)_{n\in\mathbb{N}} \subset \mathbb{R}^N \text{ and } (v_n)_{n\in\mathbb{N}} \subset \hat{\partial} f(x_n) \\ &\text{such that } \lim_{n\to+\infty} x_n = \bar{x} \text{ and } \lim_{n\to+\infty} v_n = v. \end{aligned} \tag{1}$$

A point v being a subgradient of f at $\bar{x}$ (defined by (1)) is denoted by

$$v \in \partial f(\bar{x}). \tag{2}$$

Proposition 8.12 in [55] means that, for any convex function $f\colon \mathbb{R}^N \to \mathbb{R}$ and for any point $x \in \mathbb{R}^N$,

$$\partial f(\bar{x}) = \left\{ v \in \mathbb{R}^N \colon f(x) \geq f(\bar{x}) + \langle x - \bar{x}, v\rangle \ \left(x \in \mathbb{R}^N\right) \right\} = \hat{\partial} f(\bar{x}).$$

If f is differentiable at $\bar{x}$, then $\hat{\partial} f(\bar{x}) = \{\nabla f(\bar{x})\}$ and $\nabla f(\bar{x}) \in \partial f(\bar{x})$ [55, Exercise 8.8(a)], where ∇f is the gradient of f. When f is continuous on $\mathrm{dom}(f) := \{x \in \mathbb{R}^N \colon f(x) < +\infty\}$, $\partial f\colon \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ is continuous [55, Exercise 13.29].

A bifunction $F\colon \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is said to be *pseudomonotone* if, for all $x, y \in \mathbb{R}^N$, $F(x, y) \geq 0$ implies $F(y, x) \leq 0$. F is said to be *strictly pseudomonotone* if, for all $x, y \in \mathbb{R}^N$ with $x \neq y$, $F(x, y) \geq 0$ implies $F(y, x) < 0$. $A\colon \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ is *pseudomonotone* [21, Definition 2.3.1], [38, Definition 3.2], [56, Definition 3.4] if, for all $x, y \in \mathbb{R}^N$, all $u \in A(x)$, and all $v \in A(y)$, $\langle y - x, u\rangle \geq 0$ implies $\langle x - y, v\rangle \leq 0$. $A\colon \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ is *strictly pseudomonotone* [23, Definition 4.2], [69, Definition 3.1] if, for all $x, y \in \mathbb{R}^N$ with $x \neq y$, all $u \in A(x)$, and all $v \in A(y)$, $\langle y - x, u\rangle \geq 0$ implies $\langle x - y, v\rangle < 0$. Suppose that $A\colon \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ is pseudomonotone (resp. strictly pseudomonotone) and $F\colon \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is defined for all $(x, y) \in \mathbb{R}^N \times \mathbb{R}^N$ by

$$F(x, y) := \max_{a(x)\in A(x)} \langle y - x, a(x)\rangle.$$

Then F is pseudomonotone (resp. strictly pseudomonotone).

A functional $f\colon \mathbb{R}^N \to \mathbb{R}$ is said to be *pseudoconvex* on a convex set $C \subset \mathbb{R}^N$ [51, NR 4.2-4] if, for all $x, y \in C$ with $f(x) > f(y)$, there exist $\alpha > 0$ and $\tau \in (0, 1]$ such that, for all $t \in [0, \tau]$,

$$f((1-t)x + ty) \leq f(x) - t\alpha. \tag{3}$$

f is called a *strictly pseudoconvex* functional [51, NR 4.2-4] if (3) holds whenever $f(x) \geq f(y)$ with $x \neq y$. Any convex functional is pseudoconvex. Let $f\colon \mathbb{R}^N \to \mathbb{R}$ be locally Lipschitz continuous [55, p.350] on a nonempty, open, convex set $C \subset \mathbb{R}^N$. Theorem 4.1 in [23] and Subchapter 8.J in [55] ensure that f is strictly pseudoconvex on C if and only if ∂f is strictly pseudomonotone on C, where ∂f is defined by (1) and (2).

2.2 Stochastic equilibrium problem with examples

The constraint set and bifunction considered in this paper satisfy the following.

Assumption 2.1

(A1) *$C \subset \mathbb{R}^N$ is a nonempty closed convex set onto which the metric projection can be efficiently computed.*

(A2) *Let $\Xi \subset \mathbb{R}^M$ and let $F\colon \mathbb{R}^N \times \mathbb{R}^N \times \Xi \to \mathbb{R}$ satisfy the following conditions:*
- (i) *$F((x,x);\xi) = 0$ for all $x \in \mathbb{R}^N$ and all $\xi \in \Xi$;*
- (ii) *$F((\cdot, y);\xi)$ is continuous for all $y \in \mathbb{R}^N$ and all $\xi \in \Xi$;*
- (iii) *$F((x,\cdot);\xi)$ is convex for all $x \in \mathbb{R}^N$ and all $\xi \in \Xi$.*

(A3) *$f\colon \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ is a bifunction defined for all $(x, y) \in \mathbb{R}^N \times \mathbb{R}^N$ by*

$$f(x,y) := \mathbb{E}\left[F((x,y);\xi)\right],$$

where ξ is a random vector whose probability distribution P is supported on Ξ and $\mathbb{E}[F((x,y);\xi)] \in \mathbb{R}$ is well defined for all $x, y \in \mathbb{R}^N$.

Assumption (A1) is satisfied when $C \subset \mathbb{R}^N$ is, for example, a closed ball, an affine subspace, a half-space, or a hyperslab onto which the metric projection can be computed within a finite number of arithmetic operations [4, Chapter 28]. Assumptions (A2) and (A3) indicate that f is the expectation of the function $F((\cdot,\cdot);\xi)$ satisfying standard conditions (i)–(iii) [36, p.301] (see, e.g., [36, Examples (a)–(f)] for examples of $F((\cdot,\cdot);\xi)$ satisfying (i)–(iii)). Subsection 4.2 provides an example of $F((\cdot,\cdot);\xi)$ satisfying (i)–(iii) and $\mathbb{E}[F((x,y);\xi)] \in \mathbb{R}$ ($(x,y) \in \mathbb{R}^N \times \mathbb{R}^N$) is well defined.

The main objective of this paper is to solve the following *stochastic equilibrium problem.*[1]

Problem 2.1 *Under Assumption 2.1, find*

$$x^\star \in \mathrm{EP}(C, f) := \left\{x^\star \in C\colon f(x^\star, y) \geq 0 \text{ for all } y \in C\right\}.$$

[1] The existence of a solution to Problem 2.1 is guaranteed under the assumptions in Theorem 3.1 (see the proof of Theorem 3.1 for details).

Problem 2.1 when F does not depend on ξ, i.e., the deterministic equilibrium problem of finding

$$x^\star \in \mathrm{EP}(C,F) := \{x^\star \in C\colon F(x^\star,y) \geq 0 \text{ for all } y \in C\}\,, \tag{4}$$

was examined in [6,9,36,37,63]. The proximal point algorithm [63] was presented for solving problem (4) when F is monotone [63, (A1)], i.e., F is defined for all $x, y \in \mathbb{R}^N$ by

$$F\,(x,y) + F\,(y,x) \leq 0. \tag{5}$$

The proximal point algorithm [6, Section 6] was proposed for solving problem (4) when C is replaced with $\mathrm{EP}(D,\hat{F})$, where D is closed and convex, and F and $\hat{F}$ are monotone. Meanwhile, problem (4) for the case where F is not always monotone was examined in [9,36,37]. Examples (a)–(f) in [36] indicate that problem (4) with a nonmonotone bifunction includes important nonlinear problems, such as complementarity problems, nonmonotone variational inequalities, fixed point problems, and vector minimization problems (see [37] for examples of problem (4) on a real Hausdorff topological vector space). The present paper clarifies how to consider stochastic equilibrium problem 2.1 when $F((\cdot,\cdot);\xi)$ $(\xi \in \Xi)$ is not always monotone.

A particularly interesting application of Problem 2.1 is the case that, for $x, y \in \mathbb{R}^N$ and almost every $\xi \in \Xi$,

$$F((x,y);\xi) := \max_{a(x;\xi)\in A(x;\xi)} \langle y-x, a(x;\xi)\rangle\,, \tag{6}$$

where $A\colon \mathbb{R}^N \times \Xi \rightrightarrows \mathbb{R}^N$ is continuous in the first argument. Problem 2.1 with F defined by (6) is equivalent to finding a solution to the *stochastic variational inequality* [13,40,42,53] (see also [11, Subchapter 8.3], [21, Chapter 1], [39, Chapters I and II]) for A over C, i.e., to find

$$x^\star \in C \text{ and } u^\star \in \mathbb{E}\,[A(x^\star;\xi)] \text{ such that } \langle y-x^\star, u^\star\rangle \geq 0 \text{ for all } y \in C. \tag{7}$$

We will show that any solution of Problem 2.1 with F defined by (6) satisfies (7). F defined by (6) satisfies Assumption (A2). Let $x^\star \in C$ be a solution of Problem 2.1 with F defined by (6) and let $y \in C$ be fixed arbitrarily. From the definition of (6), there exists $u(x^\star;\xi) \in A(x^\star;\xi)$ such that

$$\mathbb{E}\,[\langle y-x^\star, u(x^\star;\xi)\rangle] = \mathbb{E}\left[\max_{a(x^\star;\xi)\in A(x^\star;\xi)} \langle y-x^\star, a(x^\star;\xi)\rangle\right] \geq 0,$$

which implies that there exists $u^\star := \mathbb{E}[u(x^\star;\xi)]$ such that

$$\langle y-x^\star, u^\star\rangle \geq 0, \text{ i.e., } x^\star \text{ is a solution of (7).}$$

Problem (7) when $A(\cdot;\xi)$ is the subdifferential of a continuous function $\theta_\xi\colon \mathbb{R}^N \to \mathbb{R}$ $(\xi \in \Xi)$[2] is to find

$$x^\star \in C \text{ and } u^\star \in \mathbb{E}[\partial\theta_\xi(x^\star)] \text{ such that } \langle y-x^\star, u^\star\rangle \geq 0 \text{ for all } y \in C. \tag{8}$$

[2] In this paper, we sometimes write $\theta_\xi(\cdot)$ for a functional $\theta(\cdot;\xi)\colon \mathbb{R}^N \to \mathbb{R}$ $(\xi \in \Xi)$.

Here, let us consider the case where θ_ξ is differentiable for almost every $\xi \in \Xi$. Under the standard assumptions [59, p.423, (A1)–(A4)], we have $\nabla \mathbb{E}[\theta_\xi(x)] = \mathbb{E}[\nabla_x \theta_\xi(x)]$ for all $x \in \mathbb{R}^N$ [59, Theorem 7.49]. Hence, the point $x^\star$ satisfying (8) is a stationary point of the constrained nonconvex optimization problem of minimizing $\mathbb{E}[\theta_\xi(\cdot)]$ over C. Reference [2] presented proximal alternating and projection methods for deterministic nonconvex optimization problems and applied them to a nonconvex optimization problem of minimizing a weighted ℓ^1 norm over a convex constraint set [2, Problem (33)]. Section 4 applies the proposed method to the stochastic variational inequality (8) in machine learning [16, (25)], [71, C.3.1] and provides the performances of the proposed method for the concrete classification problems.

Another interesting application of Problem 2.1 is a *stochastic Nash equilibrium problem* [20,46,47,52] in noncooperative games. Let $\mathcal{I} := \{1, 2, \ldots, I\}$ be the set of players, let $C_i \subset \mathbb{R}^{n_i}$ $(i \in \mathcal{I})$ be the strategy set of player i, which is nonempty, closed, and convex, and let $C := \prod_{i \in \mathcal{I}} C_i$. Suppose that the loss function of player i, denoted by $f_i \colon \prod_{i \in \mathcal{I}} \mathbb{R}^{n_i} \times \Xi \to \mathbb{R}$, is continuous on $\prod_{i \in \mathcal{I}} \mathbb{R}^{n_i}$ and convex on $\mathbb{R}^{n_i}$ $(i \in \mathcal{I})$ for almost every $\xi \in \Xi$. Then the stochastic Nash equilibrium problem in noncooperative games is stated as follows:

$$\begin{aligned} &\text{find } x^\star \in C \text{ such that, for each } i \in \mathcal{I}, \\ &\quad \mathbb{E}\left[f_i(x^\star; \xi)\right] \leq \mathbb{E}\left[f_i\left(\left(y_i, x^\star_{-i}\right); \xi\right)\right] \text{ for all } y_i \in C_i, \end{aligned} \tag{9}$$

where $(y_i, x_{-i}) := (x_1, x_2, \ldots, x_{i-1}, y_i, x_{i+1}, \ldots, x_I)$ for $x := (x_1, x_2, \ldots, x_I)$, $y := (y_1, y_2, \ldots, y_I) \in \prod_{i \in \mathcal{I}} \mathbb{R}^{n_i}$. Here, we define $F \colon \prod_{i \in \mathcal{I}} \mathbb{R}^{n_i} \times \prod_{i \in \mathcal{I}} \mathbb{R}^{n_i} \times \Xi \to \mathbb{R}$ for $x := (x_1, x_2, \ldots, x_I)$, $y := (y_1, y_2, \ldots, y_I) \in \prod_{i \in \mathcal{I}} \mathbb{R}^{n_i}$ and almost every $\xi \in \Xi$ by

$$F((x, y); \xi) := \sum_{i \in \mathcal{I}} \left[f_i\left((y_i, x_{-i}); \xi\right) - f_i(x; \xi)\right]. \tag{10}$$

An observation stated in Example (d) of [36] implies that Problem 2.1 with F defined by (10) is equivalent to the Nash equilibrium problem (9).

3 Inexact Stochastic Subgradient Projection Method

3.1 Algorithm

The following conditions are needed in the presentation of the proposed method.

Assumption 3.1

(C1) *There is an independent identically distributed sample $\xi_0, \xi_1, \ldots$ of realizations of the random vector ξ.*

(C2) *There is an oracle which, for a given input point $((x, y), \xi) \in \mathbb{R}^N \times \mathbb{R}^N \times \Xi$, returns a stochastic subgradient $\mathsf{G}_x(y; \xi) \in \partial F((x, \cdot); \xi)(y)$.*

Conditions (C1) and (C2) are based on the basic assumptions for convex stochastic optimization [49, Assumptions (A1) and (A2)], [66, Assumption 8]. The history of the process $\xi_0, \xi_1, \ldots$ up to time n is denoted by $\xi_{[n]} = (\xi_0, \xi_1, \ldots, \xi_n)$. Unless stated otherwise, all relations between random variables are supported to hold almost surely.

Algorithm 1 is the inexact stochastic subgradient projection method for solving Problem 2.1 under Assumption 3.1.

Algorithm 1 Inexact stochastic subgradient projection method for Problem 2.1

Require: $n \in \mathbb{N}$, $(\lambda_n)_{n\in\mathbb{N}} \subset (0,2)$, $(\epsilon_n)_{n\in\mathbb{N}} \subset [0,+\infty)$

1: $n \leftarrow 0$, $w_0 \in C$, $\rho_0 := \|w_0\|$
2: **repeat**
3: $K_n := C \cap B(\rho_n + 1)$
4: $v_n \in K_n$ such that $F((v_n, w_n); \xi_n) \geq 0$ and
5: $\max_{v\in K_n} F((v, w_n); \xi_n) \leq F((v_n, w_n); \xi_n) + \epsilon_n$ (inexact step)
6: **if** $\mathsf{G}_{v_n}(w_n; \xi_n) \neq 0$ **then**
7: $w_{n+1} := P_C \left[w_n - \lambda_n \dfrac{F((v_n, w_n); \xi_n)}{\|\mathsf{G}_{v_n}(w_n; \xi_n)\|^2} \mathsf{G}_{v_n}(w_n; \xi_n) \right]$
8: **else**
9: $w_{n+1} := w_n$
10: **end if**
11: $\rho_{n+1} := \max\{\rho_n, \|w_{n+1}\|\}$
12: $n \leftarrow n + 1$
13: **until** stopping condition is satisfied

The way in which the point v_n satisfying step 5 in Algorithm 1 is chosen is based on the most violated constraint control [36, (4), (20)] under problem (4): given $w_n \in \mathbb{R}^N$, $v_n \in C$ is chosen so that $F(v_n, w_n) \geq 0$ and

$$\max_{v\in C\cap B(\rho_n+1)} F(v, w_n) \leq F(v_n, w_n) + \epsilon_n. \tag{11}$$

Step 5 in Algorithm 1 can be obtained by replacing F in (11) with a randomly chosen bifunction $F((\cdot,\cdot); \xi_n)$ at each iteration n. Assumption (A2)(i) and the condition $w_n \in K_n$ (from steps 3, 7, 9, and 11) imply that v_n defined by step 5 satisfies $F((v_n, w_n); \xi_n) \geq 0$ (for details, see the proof of Lemma 3.1). Since there are useful iterative methods (see, e.g., [1,2,24,26,32,33,41]) for constrained nonconvex optimization, these can be used to compute maximizers of a nonconcave function $F((\cdot, w_n); \xi_n)$ (see Assumption (A2)(ii)) over the intersection of simple closed convex sets C and $B(\rho_n+1)$, i.e., we can compute $v_n \in K_n$ satisfying step 5. See Subsection 4.3 for the computation method of v_n used in the experiments called Sequential Least SQuares Programming (SLSQP) [41].

For problem (4), the condition $0 \in \partial F(v_n, \cdot)(w_n)$ guarantees that $0 \leq F(v_n, w_n) \leq F(v_n, y)$ $(y \in C)$, which implies that $v_n \in \mathrm{EP}(C, F)$. Accordingly, a stopping condition of algorithms for solving problem (4) is $0 \in \partial F(v_n, \cdot)(w_n)$

[36, (c'), (c")]. For such a stopping condition, the following subgradient projection method for problem (4) was presented in [36, (21), (22)]: given $w_n \in \mathbb{R}^N$ and $v_n \in \mathbb{R}^N$ with (11), compute $g_n \in \partial F(v_n, \cdot)(w_n)$. If $g_n \neq 0$, then compute w_{n+1} by

$$\begin{aligned} z_n &:= w_n - \lambda_n \frac{F(v_n, w_n)}{\|g_n\|^2} g_n, \\ w_{n+1} &:= z_n + \lambda_n \left(P_C(z_n) - z_n\right), \end{aligned} \tag{12}$$

where $(\lambda_n)_{n\in\mathbb{N}} \subset [\alpha, 1] \subset (0, 1]$ for some $\alpha > 0$. In contrast, the condition $\mathsf{G}_{v_n}(w_n; \xi_n) = 0$ cannot be used as a stopping condition for Algorithm 1 since Problem 2.1 is a stochastic equilibrium problem for $f(x, y) := \mathbb{E}[F((x, y); \xi)]$. Hence, if $\mathsf{G}_{v_n}(w_n; \xi_n) = 0$, Algorithm 1 uses step 9. When $\mathsf{G}_{v_n}(w_n; \xi_n) \neq 0$, Algorithm 1 uses step 7. Step 11 is needed to show $w_n \in K_n$ ($n \in \mathbb{N}$) and the convergence of $(w_n)_{n\in\mathbb{N}}$ generated by Algorithm 1 to a solution to Problem 2.1 (see the proofs of Lemma 3.1 and Theorem 3.1 for details of using step 11). The stopping condition in step 13 is, for example, n equals some adequate number.

We rewrite step 7 in Algorithm 1 by using the stochastic subgradient $\mathsf{G}_{v_n}(w_n; \xi_n)$ of a convex functional $F((v_n, \cdot); \xi_n) \colon \mathbb{R}^N \to \mathbb{R}$ at $w_n \in C$,

$$w_{n+1} := P_C \left[w_n - \lambda_n \frac{F((v_n, w_n); \xi_n)}{\|\mathsf{G}_{v_n}(w_n; \xi_n)\|^2} \mathsf{G}_{v_n}(w_n; \xi_n) \right].$$

This step is based on (12) and the following SA method [10,27,28,54] for minimizing $\Theta := \mathbb{E}[\theta_\xi]$ for a convex functional $\theta_\xi \colon \mathbb{R}^N \to \mathbb{R}$ ($\xi \in \Xi$) over a closed convex set $C \subset \mathbb{R}^N$: given $w_0 \in \mathbb{R}^N$ and $(\alpha_n)_{n\in\mathbb{N}} \subset (0, +\infty)$,

$$w_{n+1} := P_C \left[w_n - \alpha_n \mathsf{G}(w_n, \xi_n) \right] \quad (n \in \mathbb{N}), \tag{13}$$

where there is an oracle which, for a given $(x, \xi) \in \mathbb{R}^N \times \Xi$, returns a stochastic subgradient $\mathsf{G}(x, \xi)$ such that $\mathbb{E}[\mathsf{G}(x, \xi)] \in \partial\Theta(x)$. Let $(\nu_t)_{t=1}^n \subset [0, +\infty)$ satisfy $\sum_{t=1}^n \nu_t = 1$ and let us define

$$\tilde{w}_i^n := \sum_{t=i}^n \nu_t w_t, \tag{14}$$

where $i \leq n$ and w_t ($t = 1, 2, \ldots, n$) is defined as in (13). If α_n is constant, then the result in [49, (2.21)] implies that the sequence generated by (14) satisfies

$$\mathbb{E}[\Theta(\tilde{w}_1^n) - \Theta^\star] = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right), \tag{15}$$

where $\Theta^\star$ denotes the optimal value of the minimization problem for $\Theta := \mathbb{E}[\theta_\xi]$ over C. If α_n is a diminishing step size such that $\alpha_n = c/\sqrt{n}$, where $c > 0$, then the results in [49, (2.26), (2.27)] imply that, for $i \leq n$,

$$\mathbb{E}[\Theta(\tilde{w}_i^n) - \Theta^\star] = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right). \tag{16}$$

The next section provides a convergence analysis showing that, under certain assumptions, any accumulation point of the sequence $(w_n)_{n\in\mathbb{N}}$ generated by Algorithm 1 with a constant step size almost surely belongs to $\mathrm{EP}(C, f)$ and a convergence rate analysis showing that, under certain assumptions, Algorithm 1 achieves a convergence rate of $\mathcal{O}(1/\sqrt{n})$.

3.2 Convergence analysis of Algorithm 1

Let us first prove the following lemma.

Lemma 3.1 *Under Assumptions 2.1 and 3.1, Algorithm 1 is well defined.*

Proof From $w_0 \in C$ and $\rho_0 := \|w_0\|$, $w_0 \in K_0 := C \cap B(\rho_0 + 1)$ holds. Assume that almost surely $w_n \in K_n$ for some $n \in \mathbb{N}$ and define $\rho_n := \max\{\rho_{n-1}, \|w_n\|\}$. From the closedness of C (see Assumption (A1)) and the boundedness and closedness of $B(\rho_n + 1)$, K_n is compact. Accordingly, Assumption (A2)(ii) ensures that there exists $\bar{v}_n \in K_n$ such that almost surely

$$F((\bar{v}_n, w_n); \xi_n) = \max_{v\in K_n} F((v, w_n); \xi_n).$$

Assumption (A2)(i) and the condition $w_n \in K_n$ imply that

$$F((\bar{v}_n, w_n); \xi_n) \geq F((w_n, w_n); \xi_n) = 0.$$

Therefore, there exists $v_n \in K_n$ satisfying steps 4 and 5 in Algorithm 1. Furthermore, Assumptions (C2) and (A2)(iii) imply the existence of $\mathsf{G}_{v_n}(w_n; \xi_n) \in \partial F((v_n, \cdot); \xi_n)(w_n)$. Hence, $w_{n+1} \in C$ can be defined by step 7 or step 9 in Algorithm 1. If we define $\rho_{n+1} := \max\{\rho_n, \|w_{n+1}\|\}$ and $K_{n+1} := C \cap B(\rho_{n+1}+1)$, then $w_{n+1} \in K_{n+1}$ almost surely. This implies Algorithm 1 is well defined. □

Suppose that $((w_n, v_n); \xi_n)_{n\in\mathbb{N}}$ is the sequence generated by Algorithm 1. Here, let us define $L(v_n; \xi_n) \subset C$ for $n \in \mathbb{N}$ by

$$L(v_n; \xi_n) := \{u \in C \colon F((v_n, u); \xi_n) \leq 0 \text{ almost surely}\}. \tag{17}$$

The proof of [36, Corollary 2.4] ensures that the nonempty condition of $\bigcap_{n=0}^{+\infty} L(v_n; \xi_n)$ is satisfied if $(v_n)_{n\in\mathbb{N}} \subset \mathbb{R}^N$ is almost surely bounded[3] and if $F((\cdot, \cdot); \xi)$ is pseudomonotone for almost every $\xi \in \Xi$; i.e., for all $x, y \in \mathbb{R}^N$,

$$F((x, y); \xi) \geq 0 \text{ implies } F((y, x); \xi) \leq 0. \tag{18}$$

Suppose that $A(\cdot; \xi) \colon \mathbb{R}^N \rightrightarrows \mathbb{R}^N$ is pseudomonotone for almost every $\xi \in \Xi$, i.e., for all $x, y \in \mathbb{R}^N$, all $a(x; \xi) \in A(x; \xi)$, and all $a(y; \xi) \in A(y; \xi)$,

$$\langle y - x, a(x; \xi)\rangle \geq 0 \text{ implies } \langle x - y, a(y; \xi)\rangle \leq 0.$$

[3] The sequence $(v_n)_{n\in\mathbb{N}} \subset \mathbb{R}^N$ is said to be almost surely bounded if $\sup\{\|v_n\| \colon n \in \mathbb{N}\} < +\infty$ holds almost surely [10, (2.1.4)].

Then, $F((\cdot,\cdot),\xi)$ ($\xi \in \Xi$) defined by (6) obviously satisfies the pseudomonotonicity condition (18). Subchapter 2.3 in [21] describes the detailed properties in variational inequalities for continuous pseudomonotone operators.

Here, let us provide the relationship between the nonempty condition of $\bigcap_{n=0}^{+\infty} L(v_n;\xi_n)$ for $F((\cdot,\cdot),\xi)$ ($\xi \in \Xi$) defined by (6) and the Minty variational inequality (see, e.g., [18] and [19, Definition 2.5]) to find

$$u \in C \text{ such that } \langle u - v, a(v;\xi)\rangle \leq 0 \text{ for all } v \in C \text{ and all } a(v;\xi) \in A(v;\xi). \tag{19}$$

Let $u \in C$ satisfy (19). Then, for all $n \in \mathbb{N}$,

$$F((v_n,u);\xi_n) = \max_{a(v_n;\xi_n)\in A(v_n;\xi_n)} \langle u - v_n, a(v_n;\xi_n)\rangle \leq 0,$$

which implies that $u \in \bigcap_{n=0}^{+\infty} L(v_n;\xi_n)$. Hence, the relationship between the solution set S of the Minty variational inequality (19) and $\bigcap_{n=0}^{+\infty} L(v_n;\xi_n)$ is

$$\mathrm{S} \subset \bigcap_{n=0}^{+\infty} L(v_n;\xi_n). \tag{20}$$

Therefore, we can see that, if S is nonempty, then the nonempty condition of $\bigcap_{n=0}^{+\infty} L(v_n;\xi_n)$ is satisfied.

The following assumption is needed to analyze the convergence of Algorithm 1.

Assumption 3.2

(A4) *There exists a positive real number M such that, for all $n \in \mathbb{N}$, $\|\mathsf{G}_{v_n}(w_n;\xi_n)\| \leq M$ almost surely;*

(A5) *$(\lambda_n)_{n\in\mathbb{N}} \subset [a,b]$, where $a,b \in \mathbb{R}$ with $0 < a \leq b < 2$.*

Assumptions (C2) and (A2)(iii) ensure the existence of the sequence $(\mathsf{G}_{v_n}(w_n;\xi_n))_{n\in\mathbb{N}}$ generated by Algorithm 1 (see also the proof of Lemma 3.1). Assumption (A4) is the stochastic subgradient boundedness (see, e.g., [60, Assumption 3], [61, Lemma 4.1], and [68, Assumption 1(c)]). The discussion in [36, Assumption (A), Proposition 4.3] shows that, if $F\colon \mathbb{R}^N \times \mathbb{R}^N \times \Xi \to \mathbb{R}$ satisfying Assumption (A2) is restricted on $C \times C \times \Xi$, where C is a closed convex set defined as in Assumption (A1), then Assumption (A4) holds. Assumption (A5) allows us to use a constant step size $\lambda_n := \lambda \in (0,2)$.

The following lemma provides the basic properties of Algorithm 1.

Lemma 3.2 *Suppose that Assumptions 2.1, 3.1, and 3.2 hold. Let $(w_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ be the sequences generated by Algorithm 1 and suppose that $L(v_n;\xi_n)$ is defined for $n \in \mathbb{N}$ by (17). Then the following hold:*

(i) *For all $n \in \mathbb{N}$ and all $u \in \bigcap_{n=0}^{+\infty} L(v_n;\xi_n)$, almost surely*

$$\|w_{n+1} - u\|^2 \leq \|w_n - u\|^2 - \frac{a(2-b)}{M^2} F((v_n,w_n);\xi_n)^2,$$

where $a,b \in (0,2)$ and $M > 0$ are defined as in Assumption 3.2.

(ii) *For all* $n \in \mathbb{N}$ *and all* $u \in \bigcap_{n=0}^{+\infty} L(v_n; \xi_n)$, $(\|w_n - u\|)_{n\in\mathbb{N}}$ *converges almost surely.*

(iii) *If* $\bigcap_{n=0}^{+\infty} L(v_n; \xi_n)$ *is nonempty, then* $(w_n)_{n\in\mathbb{N}}$ *and* $(v_n)_{n\in\mathbb{N}}$ *are almost surely bounded. Moreover, if* $(\epsilon_n)_{n\in\mathbb{N}} \subset [0, +\infty)$ *is a sequence converging to zero, then* $\limsup_{n\to+\infty} \mathbb{E}[f(v, w_n)] \leq 0$ *for all* $v \in \hat{C}$, *where* $\hat{C} \subset \mathbb{R}^N$ *is a closed convex set satisfying that there exists* $n_0 \in \mathbb{N}$ *such that, for all* $n \geq n_0$, $\hat{C} \subset K_n$.

Proof (i) Choose $n \in \mathbb{N}$ arbitrarily and let $u \in \bigcap_{n=0}^{+\infty} L(v_n; \xi_n)$. In the case where $\mathsf{G}_{v_n}(w_n; \xi_n) \neq 0$, step 7 in Algorithm 1 and the nonexpansivity condition of P_C with $u = P_C(u)$ ensure that

$$\begin{aligned} \|w_{n+1} - u\|^2 &= \left\| P_C\left[w_n - \lambda_n \frac{F((v_n, w_n); \xi_n)}{\|\mathsf{G}_{v_n}(w_n; \xi_n)\|^2} \mathsf{G}_{v_n}(w_n; \xi_n) \right] - P_C(u) \right\|^2 \\ &\leq \left\| (w_n - u) - \lambda_n \frac{F((v_n, w_n); \xi_n)}{\|\mathsf{G}_{v_n}(w_n; \xi_n)\|^2} \mathsf{G}_{v_n}(w_n; \xi_n) \right\|^2, \end{aligned} \tag{21}$$

which, together with $\|x - y\|^2 = \|x\|^2 - 2\langle x, y\rangle + \|y\|^2$ $(x, y \in \mathbb{R}^N)$, implies that

$$\begin{aligned} \|w_{n+1} - u\|^2 \leq \|w_n - u\|^2 &- 2\lambda_n \frac{F((v_n, w_n); \xi_n)}{\|\mathsf{G}_{v_n}(w_n; \xi_n)\|^2} \langle w_n - u, \mathsf{G}_{v_n}(w_n; \xi_n)\rangle \\ &+ \lambda_n^2 \frac{F((v_n, w_n); \xi_n)^2}{\|\mathsf{G}_{v_n}(w_n; \xi_n)\|^2}. \end{aligned}$$

Assumption (A2)(iii), $u \in L(v_n; \xi_n)$, and $\mathsf{G}_{v_n}(w_n; \xi_n) \in \partial F((v_n, \cdot); \xi_n)(w_n)$ imply that almost surely

$$0 \geq F((v_n, u); \xi_n) \geq F((v_n, w_n); \xi_n) + \langle u - w_n, \mathsf{G}_{v_n}(w_n; \xi_n)\rangle. \tag{22}$$

Accordingly, from $F((v_n, w_n); \xi_n) \geq 0$ and Assumptions (A4) and (A5), almost surely

$$\begin{aligned} \|w_{n+1} - u\|^2 &\leq \|w_n - u\|^2 - 2\lambda_n \frac{F((v_n, w_n); \xi_n)^2}{\|\mathsf{G}_{v_n}(w_n; \xi_n)\|^2} + \lambda_n^2 \frac{F((v_n, w_n); \xi_n)^2}{\|\mathsf{G}_{v_n}(w_n; \xi_n)\|^2} \\ &\leq \|w_n - u\|^2 - a(2 - b) \frac{F((v_n, w_n); \xi_n)^2}{\|\mathsf{G}_{v_n}(w_n; \xi_n)\|^2} \\ &\leq \|w_n - u\|^2 - \frac{a(2 - b)}{M^2} F((v_n, w_n); \xi_n)^2. \end{aligned} \tag{23}$$

In the case where $\mathsf{G}_{v_n}(w_n; \xi_n) = 0$, (22) and the condition $F((v_n, w_n); \xi_n) \geq 0$ guarantee that almost surely $F((v_n, w_n); \xi_n) = 0$. Since step 9 in Algorithm 1 implies that almost surely $\|w_{n+1} - u\|^2 = \|w_n - u\|^2$, (23) holds.

(ii) Taking the expectation in (23) conditioned on $\xi_{[n]}$ guarantees that, for all $n \in \mathbb{N}$ and all $u \in \bigcap_{n=0}^{+\infty} L(v_n; \xi_n)$,

$$\mathbb{E}\left[\|w_{n+1} - u\|^2 \left| \xi_{[n]} \right.\right] \leq \|w_n - u\|^2 - \frac{a(2-b)}{M^2} \mathbb{E}\left[F((v_n, w_n); \xi_n)^2 \middle| \xi_{[n]}\right] \leq \|w_n - u\|^2 - \frac{a(2-b)}{M^2} \mathbb{E}\left[F((v_n, w_n); \xi_n) \middle| \xi_{[n]}\right]^2, \quad (24)$$

where the second inequality comes from Jensen's inequality. The supermartingale convergence theorem [7, Proposition 8.2.10] thus ensures that $(\|w_n - u\|)_{n\in\mathbb{N}}$ converges almost surely.

(iii) Choose an arbitrary $u \in \bigcap_{n=0}^{+\infty} L(v_n; \xi_n)$, which is assumed to be nonempty. Lemma 3.2(ii) implies the existence of $\lim_{n\to+\infty} \|w_n - u\|$, which in turn implies that there exists $\hat{\Xi} \subset \Xi$ with $\mathbb{P}[\hat{\Xi}] = 1$ such that, for all $\xi \in \hat{\Xi}$, $\lim_{n\to+\infty} \|w_n(\xi) - u\| < +\infty$ holds. This implies that there exists $M_1 \in \mathbb{R}$ such that $\|w_n(\xi)\| \leq M_1$ for all $n \in \mathbb{N}$ and almost every $\xi \in \hat{\Xi}$. Hence, $(w_n)_{n\in\mathbb{N}}$ satisfies that almost surely $\|w_n\| \leq M_1$ for all $n \in \mathbb{N}$. Step 11 thus implies that almost surely $\rho_n = \max\{\|w_1\|, \|w_2\|, \ldots, \|w_n\|\} \leq M_1$ for all $n \in \mathbb{N}$. Hence, $v_n \in K_n \subset B(\rho_n + 1) \subset B(M_1 + 1)$, i.e., $(v_n)_{n\in\mathbb{N}}$ is almost surely bounded.

Taking total expectation in (24), together with Jensen's inequality, implies that, for all $n \in \mathbb{N}$ and all $u \in \bigcap_{n=0}^{+\infty} L(v_n; \xi_n)$,

$$\mathbb{E}\left[\|w_{n+1} - u\|^2\right] \leq \mathbb{E}\left[\|w_n - u\|^2\right] - \frac{a(2-b)}{M^2} \mathbb{E}\left[\mathbb{E}\left[F((v_n, w_n); \xi_n) \middle| \xi_{[n]}\right]\right]^2.$$

Accordingly, summing the above inequality from $n = 0$ to $n = m \in \mathbb{N}$ guarantees that, for all $u \in \bigcap_{n=0}^{+\infty} L(v_n; \xi_n)$,

$$\frac{a(2-b)}{M^2} \sum_{n=0}^{m} \mathbb{E}\left[\mathbb{E}\left[F((v_n, w_n); \xi_n) \middle| \xi_{[n]}\right]\right]^2 \leq \mathbb{E}\left[\|w_0 - u\|^2\right] < +\infty, \quad (25)$$

which implies that $\sum_{n=0}^{+\infty} \mathbb{E}[\mathbb{E}[F((v_n, w_n); \xi_n)|\xi_{[n]}]]^2 < +\infty$, and hence,

$$\lim_{n\to+\infty} \mathbb{E}\left[\mathbb{E}\left[F((v_n, w_n); \xi_n) \middle| \xi_{[n]}\right]\right] = 0. \quad (26)$$

The definition of ρ_n (step 11 in Algorithm 1) ensures that $\rho_n = \max\{\|w_1\|, \|w_2\|, \ldots, \|w_n\|\}$, which, together with the almost sure boundedness of $(w_n)_{n\in\mathbb{N}}$, implies that $(\rho_n)_{n\in\mathbb{N}}$ is almost surely bounded. Defining $\rho^\star := \sup_{n\in\mathbb{N}} \rho_n < +\infty$, the monotone increasing condition of $(\rho_n)_{n\in\mathbb{N}}$ implies that, for all $\delta \in (0, 1)$, there exists $n_0 \in \mathbb{N}$ such that $\rho_n \geq \rho^\star - \delta$ for all $n \geq n_0$. Accordingly, $C \cap B(\rho^\star + 1 - \delta) \subset K_n := C \cap B(\rho_n + 1)$ for all $n \geq n_0$ almost surely. Step 5 in Algorithm 1 ensures that, for all $n \geq n_0$ and all $v \in C \cap B(\rho^\star + 1 - \delta)$, almost surely

$$F((v, w_n); \xi_n) \leq F((v_n, w_n); \xi_n) + \epsilon_n.$$

Hence, we have that, for all $n \geq n_0$ and all $v \in C \cap B(\rho^\star + 1 - \delta)$,

$$\mathbb{E}\left[\mathbb{E}\left[F((v, w_n); \xi_n) \middle| \xi_{[n]}\right]\right] \leq \mathbb{E}\left[\mathbb{E}\left[F((v_n, w_n); \xi_n) \middle| \xi_{[n]}\right]\right] + \epsilon_n.$$

The definition of f with Assumption (A3) and the condition $w_n = w_n(\xi_{[n-1]})$ guarantee that, for all $n \geq n_0$ and all $v \in C \cap B(\rho^\star + 1 - \delta)$,

$$\mathbb{E}[f(v, w_n)] \leq \mathbb{E}\left[\mathbb{E}\left[F((v_n, w_n); \xi_n)\middle|\xi_{[n]}\right]\right] + \epsilon_n. \tag{27}$$

Accordingly, (26) and $\lim_{n\to+\infty} \epsilon_n = 0$ imply that, for all $v \in C \cap B(\rho^\star + 1 - \delta)$,

$$\limsup_{n\to+\infty} \mathbb{E}[f(v, w_n)] \leq \lim_{n\to+\infty} \mathbb{E}\left[\mathbb{E}\left[F((v_n, w_n); \xi_n)\middle|\xi_{[n]}\right]\right] + \lim_{n\to+\infty} \epsilon_n \leq 0. \tag{28}$$

This completes the proof. □

The following theorem establishes the almost sure convergence of Algorithm 1.

Theorem 3.1 *Suppose that $(w_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ are the sequences generated by Algorithm 1 under Assumptions 2.1, 3.1, and 3.2 and that $(\epsilon_n)_{n\in\mathbb{N}} \subset [0, +\infty)$ is a sequence converging to zero. Then the following hold:*

(i) *If $\bigcap_{n=0}^{+\infty} L(v_n; \xi_n)$ is nonempty, then any accumulation point of $(w_n)_{n\in\mathbb{N}}$ almost surely belongs to $\mathrm{EP}(C, f)$.*
(ii) *If $C \subset \mathbb{R}^N$ is bounded and $F((\cdot, \cdot); \xi)$ is pseudomonotone for almost every $\xi \in \Xi$, then any accumulation point of $(w_n)_{n\in\mathbb{N}}$ almost surely belongs to $\mathrm{EP}(C, f)$.*
(iii) *If $C \subset \mathbb{R}^N$ is bounded and $F((\cdot, \cdot); \xi)$ is strictly pseudomonotone for almost every $\xi \in \Xi$, then $(w_n)_{n\in\mathbb{N}}$ converges almost surely to the unique point in $\mathrm{EP}(C, f)$.*

Proof (i) Let $\mathcal{A}(w_n)_{n\in\mathbb{N}}$ be the set of accumulation points of $(w_n)_{n\in\mathbb{N}}$. The almost sure boundedness of $(w_n)_{n\in\mathbb{N}}$ implies that $\mathcal{A}(w_n)_{n\in\mathbb{N}} \neq \emptyset$. We shall prove that $\mathcal{A}(w_n)_{n\in\mathbb{N}} \subset \mathrm{EP}(C, f)$ almost surely. The closedness of C and $(w_n)_{n\in\mathbb{N}} \subset C$ ensure that any accumulation point $x^\star$ of $(w_n)_{n\in\mathbb{N}}$ is in C almost surely. The proof of Lemma 3.2(iii) ensures that there exists $n_0 \in \mathbb{N}$ such that, for all $n \geq n_0$, $\|w_n\| \leq \rho^\star < \rho^\star + 1 - \delta$ almost surely. Since there exists $(w_{n_k})_{k\in\mathbb{N}}$ which converges almost surely to $x^\star$, the continuity of $\|\cdot\|$ guarantees that

$$\|x^\star\| \leq \liminf_{k\to+\infty} \|w_{n_k}\| \leq \rho^\star + 1 - \delta;\ \text{i.e., } x^\star \in C \cap B\left(\rho^\star + 1 - \delta\right).$$

Since Assumption (A2)(iii) implies that $F((v, \cdot); \xi)$ is continuous for all $v \in C$ and all $\xi \in \Xi$, (28) ensures that, for all $v \in C \cap B(\rho^\star + 1 - \delta)$,

$$\begin{aligned} f(v, x^\star) &\leq \liminf_{k\to+\infty} \mathbb{E}\left[f(v, w_{n_k})\right] \leq \limsup_{k\to+\infty} \mathbb{E}\left[f(v, w_{n_k})\right] \\ &\leq \limsup_{n\to+\infty} \mathbb{E}\left[f(v, w_n)\right] \leq 0. \end{aligned} \tag{29}$$

Let us prove that $f(x^\star, y) \geq 0$ for all $y \in C \cap B(\rho^\star + 1 - \delta)$. Fix $y \in C \cap B(\rho^\star + 1 - \delta)$ arbitrarily and define $w_t := ty + (1 - t)x^\star$ for $t \in (0, 1)$. From $x^\star \in C \cap B(\rho^\star + 1 - \delta)$ and the convexity of C (see Assumption (A1)), $w_t \in$

$C \cap B(\rho^\star+1-\delta)$ holds. Accordingly, $f(w_t, x^\star) \leq 0$ for all $t \in (0,1)$. Assumptions (A2)(i) and (iii) thus imply that, for all $n \in \mathbb{N}$ and all $t \in (0,1)$, almost surely

$$0 = F((w_t, w_t); \xi_n) \leq tF((w_t, y); \xi_n) + (1-t)F((w_t, x^\star); \xi_n),$$

which implies that, for all $t \in (0,1)$,

$$0 \leq tf(w_t, y) + (1-t)f(w_t, x^\star) \leq tf(w_t, y); \text{ i.e., } 0 \leq f(w_t, y).$$

Hence, Assumptions (A2)(ii) and (A3) imply that $0 \leq f(x^\star, y)$. Let us prove that $x^\star \in \mathrm{EP}(C, f)$. Define $\hat{f}\colon \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ for $y \in \mathbb{R}^N$ by $\hat{f}(y) := f(x^\star, y) + I_C(y)$, where I_C denotes the indicator function of C. Assumptions (A2)(i) and (iii) imply that $\hat{f}$ is convex and $\hat{f}(x^\star) = 0$. Furthermore, from $f(x^\star, y) \geq 0$ for all $y \in C \cap B(\rho^\star+1-\delta)$, $\hat{f}(x^\star) \leq \hat{f}(y)$ for all $y \in B(\rho^\star+1-\delta)$, which implies that $x^\star$ is a local minimizer of $\hat{f}$. The convexity of $\hat{f}$ thus ensures that $x^\star$ is a global minimizer of $\hat{f}$, which implies that $0 = f(x^\star, x^\star) \leq f(x^\star, y)$ for all $y \in C$, i.e., $x^\star \in \mathrm{EP}(C, f)$. Therefore, $\mathcal{A}(w_n)_{n\in\mathbb{N}} \subset \mathrm{EP}(C, f)$ almost surely.

(ii) The boundedness of both C and $(w_n)_{n\in\mathbb{N}} \subset C$ implies that $(w_n)_{n\in\mathbb{N}}$ is almost surely bounded. Accordingly, the proof of Lemma 3.2(iii) implies the almost sure boundedness of $(v_n)_{n\in\mathbb{N}}$. Since $F((\cdot,\cdot),\xi)$ $(\xi \in \Xi)$ is pseudomonotone, the proof of Corollary 2.4 in [36] guarantees that $\bigcap_{n=0}^{+\infty} L(v_n; \xi_n) \neq \emptyset$. Therefore, Theorem 3.1(i) implies the assertion in Theorem 3.1(ii).

(iii) The strict pseudomonotonicity of $F((\cdot,\cdot),\xi)$ $(\xi \in \Xi)$ ensures the existence and uniqueness of a solution to Problem 2.1. Theorem 3.1(ii) thus guarantees that any accumulation point of $(w_n)_{n\in\mathbb{N}}$ is equal to the solution to Problem 2.1, which implies that $(w_n)_{n\in\mathbb{N}}$ converges almost surely to the solution. This completes the proof. □

Regarding the application of Algorithm 1 to the deterministic equilibrium problem, we remark as follows.

Remark 3.1 Let us consider the deterministic equilibrium problem (4) and prove that Theorem 3.1(i) implies that the whole sequence $(w_n)_{n\in\mathbb{N}}$ converges to a point in $\mathrm{EP}(C, F)$. Let $x^\star$ be an accumulation point of $(w_n)_{n\in\mathbb{N}}$. Then there exists $(w_{n_k})_{k\in\mathbb{N}} \subset (w_n)_{n\in\mathbb{N}} \subset C$ converging to $x^\star \in C$. It can be deduced from (29) that

$$F(v, x^\star) \leq 0 \text{ for all } v \in C \cap B(\rho^\star + 1 - \delta),$$

where $\rho^\star := \sup_{n\in\mathbb{N}} \rho_n < +\infty$. The arbitrariness of δ and the continuity of F ensure that

$$F(v, x^\star) \leq 0 \text{ for all } v \in C \cap B(\rho^\star + 1).$$

Condition $v_n \in K_n$ $(n \in \mathbb{N})$ implies that $F(v_n, x^\star) \leq 0$ for all $n \in \mathbb{N}$, i.e., $x^\star \in \bigcap_{n=0}^{+\infty} \{u \in C\colon F(v_n, u) \leq 0\}$. Hence, Lemma 3.2(i) and Theorem 3.1(i) guarantee that

$$0 = \lim_{k\to+\infty} \|w_{n_k} - x^\star\| = \lim_{n\to+\infty} \|w_n - x^\star\|,$$

which implies that the whole sequence $(w_n)_{n\in\mathbb{N}}$ converges to a point in $\mathrm{EP}(C, F)$.

An example of Problem 2.1 for which the if-condition in Theorem 3.1(iii) holds is the stochastic variational inequality problem (38) (see also (7) and (8)) in expected risk minimization in machine learning. See Subsection 4.2 for the proof that the variational inequality problem (38) is an example of Problem 2.1.

3.3 Convergence rate analysis of Algorithm 1

Theorem 3.1 and its proof (see (28) and (29)) guarantee that any accumulation point, denoted by $x^\star$, of $(w_n)_{n\in\mathbb{N}}$ generated by Algorithm 1 satisfies that, for all $v \in C \cap B(\rho^\star + 1 - \delta)$,

$$f(v, x^\star) \leq \limsup_{n\to+\infty} \mathbb{E}[f(v, w_n)] \leq 0 \text{ and } x^\star \in \mathrm{EP}(C, f).$$

Accordingly, we shall evaluate the rate of convergence of $(\mathbb{E}[f(v, w_n)])_{n\in\mathbb{N}}$ ($v \in C \cap B(\rho^\star + 1 - \delta)$). We can show the following proposition by referring to the proof of [35, Theorem 3.22].

Proposition 3.1 *Suppose that the assumptions in Theorem 3.1 are satisfied and that $\bigcap_{n=0}^{+\infty} L(v_n; \xi_n)$ is nonempty.*[4] *Then, for all $\epsilon > 0$, there exists $N_\epsilon \in \mathbb{N}$ such that, for all $v \in C \cap B(\rho^\star + 1 - \delta)$,*

$$\mathbb{E}\left[f(v, w_{N_\epsilon})\right] < \frac{K_1}{\sqrt{N_\epsilon}} + \epsilon_{N_\epsilon},$$

where $B(\rho^\star + 1 - \delta)$ is defined as in the proof of Lemma 3.2(iii) and $K_1 := M\sqrt{\mathbb{E}[\|w_0 - u\|^2]/(a(2-b))}$ for some $u \in \bigcap_{n=0}^{+\infty} L(v_n; \xi_n)$.

Proof Inequality (25) guarantees that, for all $n \in \mathbb{N}$,

$$\sum_{k=0}^{n} \mathbb{E}\left[\mathbb{E}\left[F((v_k, w_k); \xi_k)\middle|\xi_{[k]}\right]\right]^2 \leq \underbrace{\frac{M^2\mathbb{E}[\|w_0 - u\|^2]}{a(2-b)}}_{=:K_1^2}. \tag{30}$$

Given $\epsilon > 0$, we define $N_\epsilon \in \mathbb{N} \cup \{+\infty\}$ by

$$N_\epsilon := \inf\left\{n \in \mathbb{N}\colon \mathbb{E}\left[\mathbb{E}\left[F((v_n, w_n); \xi_n)\middle|\xi_{[n]}\right]\right]^2 \leq \epsilon\right\}. \tag{31}$$

Then, for all $n < N_\epsilon$, $\mathbb{E}[\mathbb{E}[F((v_n, w_n); \xi_n)|\xi_{[n]}]]^2 > \epsilon$, which implies that, for all $n < N_\epsilon$,

$$\sum_{k=0}^{n} \mathbb{E}\left[\mathbb{E}\left[F((v_k, w_k); \xi_k)\middle|\xi_{[k]}\right]\right]^2 > (n+1)\epsilon. \tag{32}$$

[4] This condition holds when C is bounded and $F((\cdot,\cdot);\xi)$ is pseudomonotone (see the proof of Theorem 3.1(ii)) or when $F((\cdot,\cdot);\xi)$ is defined by (6) and the solution set of the Minty variational inequality (19) is nonempty (see (20)).

We shall show that $N_\epsilon < +\infty$. Assume that $N_\epsilon = +\infty$. Then, (30) and (32) hold for all $n \in \mathbb{N}$, which implies that

$$+\infty = \sum_{k=0}^{+\infty} \mathbb{E}\left[\mathbb{E}\left[F((v_k, w_k); \xi_k)\middle|\xi_{[k]}\right]\right]^2 \leq K_1^2 < +\infty.$$

Since we have a contradiction, the condition $N_\epsilon < +\infty$ holds. Inequalities (30) and (32) with $n = N_\epsilon - 1$ guarantee that

$$\epsilon < \frac{K_1^2}{N_\epsilon}.$$

Accordingly, (31) ensures that

$$\mathbb{E}\left[\mathbb{E}\left[F((v_{N_\epsilon}, w_{N_\epsilon}); \xi_{N_\epsilon})\middle|\xi_{[N_\epsilon]}\right]\right]^2 \leq \epsilon < \frac{K_1^2}{N_\epsilon}.$$

Inequality (27), together with $F((v_n, w_n); \xi_n) \geq 0$ ($n \in \mathbb{N}$), thus implies that, for all $v \in C \cap B(\rho^\star + 1 - \delta)$,

$$\mathbb{E}\left[f(v, w_{N_\epsilon})\right] \leq \mathbb{E}\left[\mathbb{E}\left[F((v_{N_\epsilon}, w_{N_\epsilon}); \xi_{N_\epsilon})\middle|\xi_{[N_\epsilon]}\right]\right] + \epsilon_{N_\epsilon} < \frac{K_1}{\sqrt{N_\epsilon}} + \epsilon_{N_\epsilon}.$$

This completes the proof. □

Let us consider the convex stochastic optimization problem of minimizing $\Theta := \mathbb{E}[\theta_\xi]$ for a convex functional $\theta_\xi \colon \mathbb{R}^N \to \mathbb{R}$ ($\xi \in \Xi$) over a closed convex set $C \subset \mathbb{R}^N$ which satisfies Assumption (A1). Define $F \colon \mathbb{R}^N \times \mathbb{R}^N \times \Xi \to \mathbb{R}$ for $x, y \in \mathbb{R}^N$ and almost every $\xi \in \Xi$ by

$$F((x, y); \xi) := \theta_\xi(y) - \theta_\xi(x). \tag{33}$$

Then F defined by (33) satisfies Assumptions (A2) and (A3) and the monotonicity condition (5). Accordingly, when C is bounded [49, p.1574], we can check that Assumption (A4) and the if-condition of Theorem 3.1(ii) hold. Proposition 3.1, together with $w_n \in C$ ($n \in \mathbb{N}$) and $v := x^\star \in C \cap B(\rho^\star + 1 - \delta)$, implies that Algorithm 1 with F defined by (33) satisfies that

$$\mathbb{E}\left[\Theta(w_{N_\epsilon}) - \Theta^\star\right] \leq \frac{K_1}{\sqrt{N_\epsilon}},$$

where $\Theta^\star$ denotes the optimal value of the convex stochastic optimization problem and ϵ_n ($n \in \mathbb{N}$) simply equals zero (see (15) and (16) for the rate of convergence of the SA method defined by (13) and (14)).

Regarding the application of Algorithm 1 to the deterministic equilibrium problem, we remark as follows.

Remark 3.2 While algorithms were presented in [36,63] for solving the deterministic equilibrium problem (4), to the best of our knowledge, no convergence rate analyses have been performed for these algorithms. Since problem (4) is a special case of Problem 2.1, Proposition 3.1 indicates that Algorithm 1 for problem (4) satisfies that

$$F(v, w_{N_\epsilon}) < \frac{K_1}{\sqrt{N_\epsilon}} + \epsilon_{N_\epsilon}.$$

4 Application to expected risk minimization in machine learning

4.1 Capped-ℓ^1 norm coupled nonconvex overlapping group lasso and an existing machine learning algorithm called IncrePA-ncvx

Let $\{(x_i, y_i)\}_{i=1}^M \subset \mathbb{R}^N \times \mathbb{R}$ be a training set and let X be a matrix with rows x_i. The *lasso* problem [31, Chapters 2 and 4] in statistical learning [65] with sparsity is to

$$\text{find } w^\star \in \operatorname*{argmin}_{w \in \mathbb{R}^N} \frac{1}{2M} \|y - Xw\|^2 + \lambda \|w\|_1,$$

where $\lambda \geq 0$ and $\|\cdot\|_1$ denotes the ℓ^1 norm. In this section, we consider a capped-ℓ^1 norm coupled nonconvex overlapping group lasso [16, (25)], [71, C.3.1] defined as follows:

$$\text{find } w^\star \in \operatorname*{argmin}_{w \in \mathbb{R}^N} \frac{1}{2M} \underbrace{\|y - Xw\|^2}_{\substack{=\sum_{i \in \mathcal{M}} (y_i - \langle x_i, w \rangle)^2 \\ =: \sum_{i \in \mathcal{M}} l_i(w)}} + \underbrace{\sum_{k=1}^K \omega_k \underbrace{\min\{\|w_{g_k}\|, c\}}_{=: r_k(w)}}_{=: r(w)}, \tag{34}$$

where, for $g_k \subset \{1, 2, \ldots, N\}$ $(k = 1, 2, \ldots, K)$ and any $w \in \mathbb{R}^N$, w_{g_k} denotes the vector whose entries are the same as those of w for the elements in g_k and 0 for other elements, c is a constant defining the ℓ^1 norm, $\mathcal{M} := \{1, 2, \ldots, M\}$, and $\omega_k \subset [0, +\infty)$ $(k = 1, 2, \ldots, K)$ satisfies $\sum_{k=1}^K \omega_k = 1$.

The IncrePA-ncvx algorithm [16, Algorithm 2], which uses the proximal average (PA) [5] and incremental gradient methods, can be applied to the following surrogate problem in which $r(w) := \sum_{k=1}^K \omega_k r_k(w)$ is approximated by its PA $\hat{r}(w)$:

$$\text{find } w^\star \in \operatorname*{argmin}_{w \in \mathbb{R}^N} \frac{1}{2M} \sum_{i \in \mathcal{M}} l_i(w) + \hat{r}(w), \tag{35}$$

where, for all $w \in \mathbb{R}^N$,

$$M_{r_k}^\eta(w) := \min_{y \in \mathbb{R}^N} \left\{ \frac{1}{2\eta} \|w - y\|^2 + r_k(y) \right\} \ (k = 1, 2, \ldots, K, \eta > 0)$$

and $\hat{r}$ satisfies that, for all $w \in \mathbb{R}^N$,

$$M_{\hat{r}}^\eta(w) := \sum_{k=1}^K \alpha_k M_{r_k}^\eta(w),$$

where $(\alpha_k)_{k=1}^K \subset [0, +\infty)$ with $\sum_{k=1}^K \alpha_k = 1$. The proximal map of PA $\hat{r}$ is defined for all $w \in \mathbb{R}^N$ by

$$P_{\hat{r}}^\eta(w) := \sum_{k=1}^K \alpha_k P_{r_k}^\eta(w),$$

where

$$P^{\eta}_{r_k}(w) := \operatorname*{argmin}_{y\in\mathbb{R}^N}\left\{\frac{1}{2\eta}\|w-y\|^2 + r_k(y)\right\} \ (k=1,2,\ldots,K,\eta>0).$$

IncrePA-ncvx is shown in Algorithm 2. Under certain assumptions, IncrePA-ncvx converges almost surely to the asymptotic stationary point of the surrogate problem (35) [16, Theorem 3].

Algorithm 2 IncrePA-ncvx for problem (35) [16, Algorithm 2]

Require: $n\in\mathbb{N}$, $\eta>0$, $(\alpha_k)_{k=1}^K \subset [0,+\infty)$ with $\sum_{k=1}^K \alpha_k = 1$, $\mathcal{M} := \{1,2,\ldots,M\}$

$n \leftarrow 0$, $w_0 \in \mathbb{R}^N$, $\phi_{0,i} \in \mathbb{R}^N$, $\nabla l_i(\phi_{0,i}) \in \mathbb{R}^N$ $(i\in\mathcal{M})$

repeat

  $i(n) \in \mathcal{M}$ (randomly chosen)

  $\nabla l_i(\phi_{n+1,i}) := \begin{cases} \nabla l_i(w_n) & \text{if } i = i(n) \\ \nabla l_i(\phi_{n,i}) & \text{if } i \neq i(n) \end{cases} \quad (i\in\mathcal{M})$

  $\phi_{n+1,i} := \begin{cases} w_n & \text{if } i = i(n) \\ \phi_{n,i} & \text{if } i \neq i(n) \end{cases} \quad (i\in\mathcal{M})$

  $G_n := \frac{1}{M}\sum_{i\in\mathcal{M}} \nabla l_i(\phi_{n,i})$

  $v_{n+1} := \frac{1}{M}\sum_{i\in\mathcal{M}} \phi_{n,i} - \eta G_n$

  $w_{n+1} := P^{\eta}_{\hat{r}}(v_{n+1}) = \sum_{k=1}^K \alpha_k P^{\eta}_{r_k}(v_{n+1})$

  $n \leftarrow n+1$

until stopping condition is satisfied

4.2 Stochastic variational inequality in expected risk minimization and the existing and proposed machine learning algorithms

To satisfy the boundedness condition on C in Theorem 3.1(iii), we set C in Algorithm 1 as a sufficiently large closed ball (see [57, Fig.1] for the existing machine learning algorithm using the projection onto a bounded set). Since P_C can be easily computed, (A1) holds. For all $i\in\mathcal{M}$, $\theta_i\colon \mathbb{R}^N \to \mathbb{R}$ is defined for $w\in\mathbb{R}^N$ by

$$\theta_i(w) := \frac{1}{M}\left\{\frac{1}{2}l_i(w) + \underbrace{\sum_{k=1}^K \omega_k \min\left\{\|w_{g_k}\|, s\|w_{g_k}\| + (1-s)c\right\}}_{=:\tilde{r}(w)}\right\}, \tag{36}$$

where $s>0$ is sufficiently small. The functional θ_i $(i\in\mathcal{M})$ defined by (36) is an approximation function[5] of $(1/M)((1/2)l_i(w)+r(w))$. Since $\tilde{r}$ is strictly

[5] When s is sufficiently small, $\tilde{r}(w) := \sum_{k=1}^K \omega_k \min\{\|w_{g_k}\|, s\|w_{g_k}\|+(1-s)c\} \approx r(w) := \sum_{k=1}^K \omega_k \min\{\|w_{g_k}\|, c\}$ in the sense of the norm of $\mathbb{R}$.

pseudoconvex[6] [23, Definition 4.1(ii)], [51, NR 4.2-4] and l_i ($i \in \mathcal{M}$) is convex, θ_i ($i \in \mathcal{M}$) is strictly pseudoconvex. Here, we define $F\colon \mathbb{R}^N \times \mathbb{R}^N \times \mathcal{M} \to \mathbb{R}$ by (6) with $A(\cdot;\xi) := \partial\theta_\xi(\cdot)$; i.e., for $w, y \in \mathbb{R}^N$ and for $\xi \in \mathcal{M}$,

$$F((w,y);\xi) := \max_{a(w;\xi)\in\partial\theta_\xi(w)} \langle y - w, a(w;\xi)\rangle\,. \tag{37}$$

It is obvious that $F((w,w);\xi) = 0$ ($\xi \in \mathcal{M}, w \in \mathbb{R}^N$) and $F((w,\cdot);\xi)$ ($\xi \in \mathcal{M}, w \in \mathbb{R}^N$) is convex. The continuity of $\partial\theta_\xi$ ($\xi \in \mathcal{M}$) implies that $F((\cdot,y);\xi)$ ($\xi \in \mathcal{M}, y \in \mathbb{R}^N$) is continuous. Hence, F defined by (37) satisfies (A2). Since θ_ξ ($\xi \in \mathcal{M}$) is strictly pseudoconvex, $A(\cdot;\xi) := \partial\theta_\xi(\cdot)$ ($i \in \mathcal{M}$) is strictly pseudomonotone [23, Theorem 4.1], which implies that $F((\cdot,\cdot);\xi)$ ($\xi \in \mathcal{M}$) is strictly pseudomonotone (see also the if-condition of Theorem 3.1(iii)). Theorem 7.47 in [59] and the continuity of $F((\cdot,\cdot);\xi)$ ($\xi \in \mathcal{M}$) guarantee that $f(w,y) := \mathbb{E}[F((w,y);\xi)] \in \mathbb{R}$ is well defined for all $w, y \in \mathbb{R}^N$, which implies that (A3) holds. Therefore, Assumption 2.1 and the if-condition of Theorem 3.1(iii) are satisfied for C and F considered in this subsection.

The useful machine learning algorithms [57] randomly choose training examples and improve their approximations by using (sub)gradients of objective functions corresponding to the chosen examples. Moreover, the convexity of $F((w,\cdot);\xi)$ ($\xi \in \mathcal{M}, w \in \mathbb{R}^N$) defined by (37) ensures that $\mathsf{G}_w(y;\xi) \in \partial F((w,\cdot);\xi)(y)$ can be determined when $((w,y),\xi)$ is given. Accordingly, Assumption 3.1 holds.

The discussion in Subsection 2.2 implies that Problem 2.1 with F defined by (37) is the stochastic variational inequality problem for $\partial\theta_\xi$ with θ_ξ defined by (36) over C to find

$$w^\star \in C \text{ and } u^\star \in \mathbb{E}\left[\partial\theta_\xi(w^\star)\right] \text{ such that } \langle y - w^\star, u^\star\rangle \geq 0 \text{ for all } y \in C. \tag{38}$$

The sequences $(w_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ generated by Algorithm 1 with $w_0 \in C$ and F defined by (37) satisfy that $(w_n)_{n\in\mathbb{N}}, (v_n)_{n\in\mathbb{N}} \subset C$ almost surely. Accordingly, we may assume without loss of generality that F is defined on $C \times C \times \mathcal{M}$. This implies that Algorithm 1 satisfies (A4). Therefore, the above discussion and Theorem 3.1(iii), together with (A5), imply the following result:

- The sequence $(w_n)_{n\in\mathbb{N}}$ generated by Algorithm 1 with F defined by (37) and $\lambda_n := \lambda \in (0,2)$ (Algorithm 3) converges almost surely to the solution to the stochastic variational inequality (38) in the capped-ℓ^1 norm coupled nonconvex overlapping group lasso problem (34).

The stochastic extragradient (SE) method [35, Algorithm 1] can solve stochastic pseudomonotone variational inequalities. This implies that the SE method can be applied to problem (38). The SE method is shown in Algorithm 4. Under certain assumptions, every accumulation point of $(w_n)_{n\in\mathbb{N}}$

[6] The function r defined by (34) is quasiconvex [62, Theorems 4.1 and 4.3] rather than pseudoconvex. Accordingly, we modify $(1/M)((1/2)l_i(w) + r(w))$ with θ_i so that F defined by (37) can satisfy the strict pseudomonotonicity condition that is needed to guarantee the almost sure convergence of Algorithm 1 to the solution to Problem 2.1 with F defined by (37) (see Theorem 3.1(iii)).

Algorithm 3 ISSP (Algorithm 1 with $F((w,y);\xi) := \max_{a(w;\xi)\in\partial\theta_\xi(w)} \langle y-w, a(w;\xi)\rangle$)

Require: $n \in \mathbb{N}$, $(\lambda_n)_{n\in\mathbb{N}} \subset (0,2)$, $(\epsilon_n)_{n\in\mathbb{N}} \subset [0,+\infty)$
$n \leftarrow 0$, $w_0 \in C$, $\rho_0 := \|w_0\|$
repeat
  $K_n := C \cap B(\rho_n + 1)$
  $v_n \in K_n$ such that $F((v_n,w_n);\xi_n) \geq 0$ and
  $\max_{v\in K_n} F((v,w_n);\xi_n) \leq F((v_n,w_n);\xi_n) + \epsilon_n$ (inexact step)
  if $\mathsf{G}_{v_n}(w_n;\xi_n) \neq 0$ **then**
    $w_{n+1} := P_C\left[w_n - \lambda_n \frac{F((v_n,w_n);\xi_n)}{\|\mathsf{G}_{v_n}(w_n;\xi_n)\|^2}\mathsf{G}_{v_n}(w_n;\xi_n)\right]$
  else
    $w_{n+1} := w_n$
  end if
  $\rho_{n+1} := \max\{\rho_n, \|w_{n+1}\|\}$
  $n \leftarrow n+1$
until stopping condition is satisfied

generated by the SE method almost surely belongs to the solution set of problem (38) [35, Theorem 3.18]. Moreover, the SE method satisfies that, for all $\epsilon > 0$, there exists $N_\epsilon \in \mathbb{N}$ such that $\mathbb{E}[r_\alpha(w_{N_\epsilon})^2] \leq K/N_\epsilon$, where $r_\alpha(w) := \|w - P_C[w - \mathbb{E}[\partial\theta_\xi(w)]]\|^2$ ($w \in \mathbb{R}^N$) and $K < +\infty$ [35, Theorem 3.22] (see Proposition 3.1 for the convergence rate of Algorithm 3).

Algorithm 4 Stochastic extragradient (SE) method for problem (38) [35, Algorithm 1]

Require: $n \in \mathbb{N}$, $(\alpha_n)_{n\in\mathbb{N}} \subset (0,+\infty)$, $(N_n)_{n\in\mathbb{N}} \subset (0,+\infty)$, $(\xi_{n,j})_{j=1}^{N_n}, (\eta_{n,j})_{j=1}^{N_n} \subset \mathcal{M}$
$n \leftarrow 0$, $w_0 \in \mathbb{R}^N$
repeat
  $\mathsf{G}(w_n,\xi_{n,j}) \in \partial\theta_{\xi_{n,j}}(w_n)$ $(j = 1,2,\ldots,N_n)$
  $z_n := P_C\left[w_n - \frac{\alpha_n}{N_n}\sum_{j=1}^{N_n}\mathsf{G}(w_n,\xi_{n,j})\right]$
  $\mathsf{G}(z_n,\eta_{n,j}) \in \partial\theta_{\eta_{n,j}}(z_n)$ $(j = 1,2,\ldots,N_n)$
  $w_{n+1} := P_C\left[w_n - \frac{\alpha_n}{N_n}\sum_{j=1}^{N_n}\mathsf{G}(z_n,\xi_{n,j})\right]$
  $n \leftarrow n+1$
until stopping condition is satisfied

4.3 Numerical comparisons of existing machine learning algorithms with the proposed machine learning algorithm

This subsection numerically compares the performance of the existing machine learning algorithm based on each of Algorithm 2 (IncrePA), Algorithm 4 (SE), and the SA method (SA) defined by (13) and (14) with $\mathsf{G}(w_n,\xi_n) \in \partial\theta_{\xi_n}(w_n)$

with that of the proposed machine learning algorithm based on Algorithm 3 (inexact stochastic subgradient projection method, ISSP) for the capped-ℓ^1 norm coupled nonconvex overlapping group lasso (see Subsections 4.1 and 4.2 for details of the methods). In the experiments, we used the following parameters, which are based on those in [16, Subsection 6.3]: $w_0 = \phi_{0,i} := 0$ $(i \in \mathcal{M})$, $K \in \{1, 5, 10, 15, 20\}$, $\omega_k := 1/K$ $(k = 1, 2, \ldots, K)$, $c := 0.1$, and $s := 10^{-8}$. The machine learning algorithms used in the experiments are as follows:

- **IncrePA**: Machine learning algorithm based on Algorithm 2 [16, Algorithm 2] with η and α_k satisfying [16, (2), (29)]
- **SE**: Machine learning algorithm based on Algorithm 4 [35, Algorithm 1] with α_n satisfying [35, Assumption 3.7] and $N_n := \Theta(n + \mu)^{1+a}(\ln(n + \mu))^{1+b}$ [35, p.696], where $\Theta, \mu, a > 0$ and $b \geq -1$
- **SA**: Machine learning algorithm based on the SA method (13) and (14) with α_n and ν_t satisfying [49, (2.25), p.1579] and $\mathsf{G}(w_n, \xi_n) \in \partial\theta_{\xi_n}(w_n)$
- **ISSP(C1)**: Machine learning algorithm based on Algorithm 1 with F defined by (37) (Algorithm 3) and $\lambda_n := 0.5$
- **ISSP(C2)**: Machine learning algorithm based on Algorithm 1 with F defined by (37) (Algorithm 3) and $\lambda_n := 1.0$
- **ISSP(C3)**: Machine learning algorithm based on Algorithm 1 with F defined by (37) (Algorithm 3) and $\lambda_n := 1.5$

The computer used in the experiments was a MacPro (Late 2013) computer with a 3 GHz 8-Core Intel Xeon E5 CPU, 32 GB 1,866 MHz DDR3 memory, and 500 GB flash storage. The operating system was MacOS Sierra (version 10.14.6). The evaluation programs were run in Python 3.7.4 with NumPy 1.17.0, SciPy 1.3.1, and scikit-learn 0.21.3. The experiments used the datasets from the LIBSVM [12], for which information is shown in Table 1. The index group of features $g_k \subset \{1, 2, \ldots, N\}$ $(k = 1, 2, \ldots, K)$ was set by using the `numpy.random.randint` function in NumPy 1.17.0, which returns random integers 0 and 1 from the discrete uniform distribution. For example, the following $K \times N$ matrix for the "breast-cancer" dataset ($N = 10$) with $K = 5$ was set by using the `numpy.random.randint` function:

$$\begin{bmatrix} 0\,1\,1\,0\,1\,1\,1\,1\,1\,0 \\ 1\,1\,0\,0\,1\,0\,0\,0\,0\,1 \\ 0\,1\,0\,1\,1\,0\,0\,1\,1\,0 \\ 1\,1\,0\,1\,0\,1\,0\,1\,1\,0 \\ 0\,1\,1\,0\,0\,1\,0\,1\,1\,1 \end{bmatrix} = \begin{bmatrix} \tilde{g}_1 \\ \tilde{g}_2 \\ \tilde{g}_3 \\ \tilde{g}_4 \\ \tilde{g}_5 \end{bmatrix},$$

which implies that

$$g_1 = \{2, 3, 5, 6, 7, 8, 9\},\ g_2 = \{1, 2, 5, 10\},\ g_3 = \{2, 4, 5, 8, 9\},$$
$$g_4 = \{1, 2, 4, 6, 8, 9\},\ g_5 = \{2, 3, 6, 8, 9, 10\}.$$

The point v_n satisfying steps 4 and 5 in Algorithm 1 (Algorithm 3) was computed by using the SLSQP [41] optimization solver with $\mathsf{max\_iter} = 5$ and the

initial point v_{n-1} in the SciPy 1.3.1 package. The point $P^{\eta}_{r_k}(v_{n+1})$ defined in step 8 in Algorithm 2 was computed by using BFGS [50, Chapter 6] in the SciPy 1.3.1 package. The stopping condition for the algorithms was $n = 100$. In the experiments, 10-fold cross-validation for the datasets was performed using the `sklearn.model_selection.StratifiedKFold` class. Multiclass classifiers were conducted using the `sklearn.multiclass.OneVsRestClassifier` class, which provides a construction of one-versus-the-rest (OvR) multiclass classifiers.

Table 1 Datasets used for classification [12]

Dataset	Classes	Data points	Features
breast-cancer	2	683	10
german.numer	2	1,000	24
ionosphere	2	351	34
iris	3	150	4
wine	3	178	13
vehicle	4	846	18
covtype	7	581,012	54
pendigits	10	7,494 + 3,498	16

Table 2 Classification accuracies (%) and elapsed times (s) for the machine learning algorithms applied to the datasets in Table 1 (highest accuracy score is indicated by bold type)

Dataset	K	**IncrePA**		**SE**		**SA**		**ISSP(C1)**		**ISSP(C2)**		**ISSP(C3)**	
		acc.	time	acc.	time	acc.	time	acc.	time	acc.	time	acc.	time
breast-cancer	1	65.4	0.04	37.4	0.04	37.2	0.09	92.0	0.24	**93.5**	0.04	88.3	0.04
	5	41.0	0.04	37.2	0.04	37.2	0.09	91.7	0.06	**92.0**	0.04	91.7	0.04
	10	53.7	0.04	37.2	0.04	37.2	0.10	**92.7**	0.04	**92.7**	0.04	91.0	0.04
	15	43.1	0.04	37.2	0.04	37.2	0.09	**92.0**	0.05	**92.0**	0.04	88.9	0.04
	20	49.4	0.04	37.2	0.04	37.2	0.11	**91.9**	0.04	89.2	0.04	90.2	0.04
german.numer	1	70.0	0.04	**75.1**	0.04	61.3	0.09	69.9	0.04	68.6	0.04	67.6	0.04
	5	**70.0**	0.04	68.8	0.04	57.0	0.09	63.4	0.04	68.2	0.04	66.0	0.04
	10	70.0	0.04	70.0	0.04	64.8	0.09	63.0	0.04	**70.9**	0.03	69.3	0.04
	15	70.0	0.04	70.0	0.04	59.6	0.10	61.7	0.04	**72.6**	0.04	66.5	0.04
	20	**70.0**	0.04	**70.0**	0.04	64.2	0.04	61.8	0.04	**70.0**	0.04	**70.0**	0.04
ionosphere	1	58.4	0.04	77.7	0.04	69.7	0.09	**81.7**	0.04	76.3	0.04	80.0	0.04
	5	49.2	0.04	79.1	0.04	70.9	0.09	58.6	0.05	76.9	0.03	**82.3**	0.03
	10	47.2	0.04	79.1	0.04	68.3	0.09	75.5	0.05	**81.2**	0.04	78.6	0.03
	15	39.8	0.04	79.1	0.04	67.7	0.10	62.7	0.04	**84.3**	0.03	82.9	0.04
	20	39.5	0.04	79.1	0.04	71.5	0.09	78.0	0.04	**79.8**	0.03	75.8	0.03
iris	1	69.3	0.03	33.3	0.03	**72.6**	0.09	**72.6**	0.03	69.3	0.03	71.3	0.03
	5	48.6	0.03	33.3	0.03	64.6	0.09	**79.3**	0.03	64.0	0.03	62.6	0.03
	10	42.0	0.03	33.3	0.03	43.3	0.09	**75.3**	0.03	74.6	0.03	61.3	0.03
	15	36.0	0.03	33.3	0.03	43.3	0.09	66.6	0.03	60.0	0.03	**70.0**	0.03
	20	35.3	0.03	33.3	0.03	25.3	0.09	60.6	0.03	**64.6**	0.03	49.3	0.03
wine	1	89.4	0.03	32.7	0.04	85.5	0.10	**95.5**	0.03	94.4	0.03	91.5	0.03
	5	69.4	0.03	32.7	0.03	82.2	0.09	**97.2**	0.03	93.8	0.04	95.0	0.03
	10	63.0	0.03	32.7	0.04	78.0	0.09	**96.1**	0.03	94.4	0.03	94.9	0.03
	15	58.3	0.03	32.7	0.04	75.4	0.09	94.4	0.03	94.4	0.03	**95.0**	0.04
	20	62.8	0.03	32.7	0.04	80.5	0.09	91.1	0.04	96.0	0.04	**97.1**	0.04
vehicle	1	38.5	0.05	25.0	0.04	37.8	0.09	**59.2**	0.04	55.8	0.04	55.0	0.04
	5	23.5	0.05	25.0	0.04	37.7	0.10	57.4	0.06	54.3	0.04	**57.5**	0.04
	10	23.5	0.05	25.0	0.04	29.1	0.10	**59.0**	0.05	58.1	0.04	55.7	0.04
	15	23.5	0.05	25.0	0.05	36.9	0.10	**57.7**	0.04	57.2	0.04	56.3	0.04
	20	23.5	0.05	25.0	0.05	34.2	0.10	56.2	0.04	**57.6**	0.04	54.5	0.04
covtype	1	51.2	0.22	51.2	0.51	49.0	0.28	**58.8**	0.22	58.7	0.21	58.5	0.21
	5	58.9	0.20	51.2	0.87	56.0	0.26	**60.5**	0.28	56.4	0.19	60.0	0.20
	10	**58.9**	0.20	51.2	1.34	50.5	0.26	58.0	0.32	55.7	0.19	55.7	0.20
	15	**58.9**	0.20	51.2	1.78	49.5	0.26	55.6	0.34	58.0	0.20	54.4	0.19
	20	58.9	0.20	51.2	2.16	48.9	0.27	55.5	0.40	**59.7**	0.20	57.9	0.20
pendigits	1	10.0	0.06	11.5	0.08	51.9	0.17	**81.1**	0.18	80.7	0.06	75.5	0.06
	5	10.0	0.06	11.6	0.08	43.1	0.12	66.8	0.06	**67.3**	0.06	59.6	0.06
	10	10.0	0.07	11.5	0.08	43.6	0.14	**65.7**	0.06	56.8	0.06	54.6	0.09
	15	10.0	0.11	11.6	0.08	41.2	0.16	**62.3**	0.06	55.6	0.11	56.6	0.06
	20	10.0	0.11	11.4	0.07	39.1	0.17	**64.4**	0.06	56.5	0.11	51.9	0.10
Average		47.0	0.06	42.6	0.21	53.5	0.12	72.1	0.09	72.5	0.06	71.0	0.06

Table 2 shows the classification accuracies and the elapsed times for the machine learning algorithms used in the experiments. We first consider the results of the binary classification. For the "breast-cancer" dataset with a fixed K, ISSP(C1), ISSP(C2), and ISSP(C3) performed better than IncrePA, SE, and SA from the viewpoint of accuracy. The elapsed times for IncrePA and SE were almost the same as those for ISSP(C2) and ISSP(C3). For the "german.numer" dataset with $K = 1$, the accuracy of SE was higher than those of other algorithms. For the "german.numer" dataset with $K = 5, 10, 15, 20$, the accuracies of IncrePA, SE, and ISSP(C2) were about 70%. For the "ionosphere" dataset with a fixed K, SE and ISSP(C3) had high accuracies on average.

We next consider the results of the multiclass classification. For the "iris", "wine", "vehicle", and "pendigits" datasets with a fixed K, the accuracy of each of ISSP(C1), ISSP(C2), and ISSP(C3) was higher than those of other algorithms. In particular, for the "vehicle" dataset with a fixed K, the accuracies of ISSP(C1), ISSP(C2), and ISSP(C3) were about 60%, while the accuracies of other algorithms were less than 40%. The elapsed times of IncrePA and SE were almost the same as those of ISSP(C1), ISSP(C2), and ISSP(C3). For the "covtype" dataset, IncrePA, ISSP(C1), ISSP(C2), and ISSP(C3) had high accuracies. The elapsed time of SE was longer than that of other algorithms. This is because the computation of $\sum_{j=1}^{N_n} \mathsf{G}(\cdot, \xi_{n,j})$ in SE (Algorithm 4) became more time-consuming with the increase of the value of $N_n := \Theta(n+\mu)^{1+a}(\ln(n+\mu))^{1+b}$.

The average accuracies of IncrePA, SE, SA, ISSP(C1), ISSP(C2), and ISSP(C3) were respectively 47.0%, 42.6%, 53.5%, 72.1%, 72.5%, and 71.0%. The average elapsed times of IncrePA, SE, SA, ISSP(C1), ISSP(C2), and ISSP(C3) were respectively 0.06 s, 0.21 s, 0.12 s, 0.09 s, 0.06 s, and 0.06 s. Therefore, we can see that the average performances of the different settings of the proposed machine learning algorithm were almost the same.

The average accuracies and elapsed times of the existing algorithms (IncrePA, SE, and SA) were compared to the average accuracies and elapsed times of the proposed algorithms (ISSP(C1), ISSP(C2), and ISSP(C3)) by using an analysis of variance (ANOVA) test and Tukey-Kramer's honestly significant difference (HSD) test. The `scipy.stats.f_oneway` method in the SciPy library was used as the implementation of the ANOVA test, and the `statsmodels.stats.multicomp.pairwise_tukeyhsd` method in the StatsModels package was used as the implementation of Tukey-Kramer's HSD test. The ANOVA test examines whether the hypothesis that the given groups have the same population mean is rejected. Tukey-Kramer's HSD test can be used to find specifically which pair has a significant difference in groups. The significance level was 5% (0.05) for the ANOVA and Tukey-Kramer's HSD tests.

Let us evaluate the accuracies on the datasets in Table 1. The p-value computed by the ANOVA test was about 4.50×10^{-20} (< 0.05). This implies that there is a significant difference in terms of accuracy between the algorithms used in the experiments for every dataset. Here, let us check the results of the Tukey-Kramer's HSD test, as shown in Table 3. Table 3 indicates that the ad-

justed p-value between each of the proposed algorithms (ISSP(C1), ISSP(C2), and ISSP(C3)) and each of the existing algorithms (IncrePA, SE, and SA) was 0.001 (< 0.05), which implies that the accuracies of the proposed algorithms were significantly better than those of the existing algorithms. It also shows that the adjusted p-value between the proposed algorithms was 0.9 (> 0.05). Accordingly, the proposed algorithms had almost the same performances in the sense of accuracy. We can also check that there was not a significant difference in accuracy between the existing algorithms.

Finally, let us evaluate the elapsed time on the datasets in Table 1. The p-value computed by the ANOVA test was about 0.01 (< 0.05). Table 4 indicates that there is a significant difference in the sense of the elapsed time between each of IncrePA, ISSP(C2), and ISSP(C3) and SE. Specifically, it shows that IncrePA, ISSP(C2), and ISSP(C3) ran significantly faster than SE.

Table 3 Multiple comparison for accuracies for the machine learning algorithms applied to the datasets in Table 1 using Tukey-Kramer's HSD test with the 5% significant level ("meandiffs" indicates the pairwise mean differences between Groups 1 and 2, "p-adj" indicates the adjusted p-value, and "Lower" (resp. "Upper") indicates the lower (resp. upper) value of the confidence interval for the pairwise mean differences.)

Group 1	Group 2	meandiffs	p-adj	Lower	Upper	Reject
ISSP(C1)	ISSP(C2)	0.4673	0.9	-10.7066	11.6412	False
ISSP(C1)	ISSP(C3)	-1.0669	0.9	-12.2408	10.107	False
ISSP(C1)	IncrePA	-25.0972	0.001	-36.2711	-13.9233	**True**
ISSP(C1)	SE	-29.5122	0.001	-40.6861	-18.3383	**True**
ISSP(C1)	SA	-18.5863	0.001	-29.7601	-7.4124	**True**
ISSP(C2)	ISSP(C3)	-1.5341	0.9	-12.708	9.6398	False
ISSP(C2)	IncrePA	-25.5645	0.001	-36.7384	-14.3906	**True**
ISSP(C2)	SE	-29.9795	0.001	-41.1534	-18.8056	**True**
ISSP(C2)	SA	-19.0535	0.001	-30.2274	-7.8796	**True**
ISSP(C3)	IncrePA	-24.0304	0.001	-35.2042	-12.8565	**True**
ISSP(C3)	SE	-28.4453	0.001	-39.6192	-17.2715	**True**
ISSP(C3)	SA	-17.5194	0.001	-28.6933	-6.3455	**True**
IncrePA	SE	-4.415	0.8524	-15.5889	6.7589	False
IncrePA	SA	6.511	0.5431	-4.6629	17.6849	False
SE	SA	-10.926	0.0595	-22.0998	0.2479	False

From the above discussion, we can conclude that the proposed machine learning algorithm is superior for solving the capped-ℓ^1 norm coupled nonconvex overlapping group lasso.

5 Conclusion and future work

This paper presented an inexact stochastic subgradient projection method for solving the stochastic equilibrium problem with nonmonotone bifunctions. A convergence analysis showed that, under certain assumptions, any accumulation point of the sequence generated by the proposed method almost surely belongs to the solution set of the stochastic equilibrium problem. A convergence

Table 4 Multiple comparison for elapsed time for the machine learning algorithms applied to the datasets in Table 1 using Tukey-Kramer's HSD test with the 5% significant level ("meandiffs" indicates the pairwise mean differences between Groups 1 and 2, "p-adj" indicates the adjusted p-value, and "Lower" (resp. "Upper") indicates the lower (resp. upper) value of the confidence interval for the pairwise mean differences.)

Group 1	Group 2	meandiffs	p-adj	Lower	Upper	Reject
ISSP(C1)	ISSP(C2)	-0.0236	0.9	-0.1558	0.1085	False
ISSP(C1)	ISSP(C3)	-0.0241	0.9	-0.1562	0.1081	False
ISSP(C1)	IncrePA	-0.0209	0.9	-0.1531	0.1112	False
ISSP(C1)	SE	0.1207	0.0954	-0.0115	0.2528	False
ISSP(C1)	SA	0.0374	0.9	-0.0948	0.1695	False
ISSP(C2)	ISSP(C3)	-0.0004	0.9	-0.1326	0.1317	False
ISSP(C2)	IncrePA	0.0027	0.9	-0.1294	0.1349	False
ISSP(C2)	SE	0.1443	0.0233	0.0122	0.2765	**True**
ISSP(C2)	SA	0.061	0.7428	-0.0711	0.1932	False
ISSP(C3)	IncrePA	0.0031	0.9	-0.129	0.1353	False
ISSP(C3)	SE	0.1447	0.0227	0.0126	0.2769	**True**
ISSP(C3)	SA	0.0614	0.7376	-0.0707	0.1936	False
IncrePA	SE	0.1416	0.0278	0.0094	0.2737	**True**
IncrePA	SA	0.0583	0.7766	-0.0739	0.1904	False
SE	SA	0.0833	0.4618	-0.0488	0.2155	False

rate analysis was also shown that supported the efficiency of the proposed method. The machine learning algorithm based on the proposed method was numerically compared to existing machine learning algorithms with respect to the capped-ℓ^1 norm coupled nonconvex overlapping group lasso. The numerical results using LIBSVM datasets demonstrated that the average performances of the existing machine learning algorithms are significantly different from the average performance of the proposed machine learning algorithm and that the proposed machine learning algorithm is useful for solving the capped-ℓ^1 norm coupled nonconvex overlapping group lasso.

However, the numerical results also showed that, for a dataset with many data points, the proposed machine learning algorithm did not always have a high accuracy. This is because the machine learning algorithms could not use much training data before the stopping condition was reached. In the future, we should consider developing stochastic optimization methods based on other useful learning methods, such as ensemble leaning [72], to improve accuracy. Moreover, the numerical results showed that, in the case where the stopping condition was $n = 100$, the classification accuracies seemed not to be closely related to the parameter K values. As a result, the numerical results did not directly suggest an explanation about the effects of parameter K. In the future, we should investigate the relationship between the grouping information for datasets and parameter K, e.g., the relationship between the classification accuracies and K in the case that the stopping conditions are, for example, $n = \hat{M}/2, \hat{M}, 2\hat{M}$, where $\hat{M}$ is the number of training data needed for each classifier to learn the weights. Additionally, problem (34) is related to partial sparse optimization [44]. When our focus is sparse optimization, we should

implement not only stochastic optimization algorithms but also the algorithms in [44, Section 5] and evaluate their performances.

Acknowledgments

I am sincerely grateful to Editor-in-Chief Sergiy Butenko and the two anonymous reviewers for helping me improve the original manuscript. I also thank Kazuhiro Hishinuma for his input on the numerical examples.

Funding

This work was supported by JSPS KAKENHI Grant Number JP18K11184.

Conflicts of interest

The author declares that he has no conflict of interest.

Availability of data and material

Not applicable.

Code availability

Not applicable.

References

1. Androulakis, I.P., Maranasand, C.D., Flouda, C.A.: αBB: A global optimization method for general constrained nonconvex problems. Journal of Global Optimization **7**, 337–363 (1995)
2. Attouch, H., Bolte, J., Redont, P., Soubeyran, A.: Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality. Mathematics of Operations Research **35**, 438–457 (2010)
3. Bauschke, H.H., Borwein, J.M.: On projection algorithms for solving convex feasibility problems. SIAM Review **38**, 367–426 (1996)
4. Bauschke, H.H., Combettes, P.L.: Convex Analysis and Monotone Operator Theory in Hilbert Spaces. Springer, New York (2011)
5. Bauschke, H.H., Goebel, R., Lucet, Y., Wang, X.: The proximal average: Basic theory. SIAM Journal on Optimization **19**, 766–785 (2008)
6. Bento, G.C., Cruz Neto, J.X., Lopes, J.O., Soares Jr., P.A., Soubeyran, A.: Generalized proximal distances for bilevel equilibrium problems. SIAM Journal on Optimization **26**, 810–830 (2016)
7. Bertsekas, D.P., Nedić, A., Ozdaglar, A.E.: Convex Analysis and Optimization. Athena Scientific (2003)

8. Bianchi, M., Schaible, S.: Generalized monotone bifunctions and equilibrium problems. Journal of Mathematical Analysis and Applications **90**, 31–43 (1996)
9. Blum, E., Oettli, W.: From optimization and variational inequalities to equilibrium problems. The Mathematics Student **63**, 123–145 (1994)
10. Borkar, V.S.: Stochastic Approximation: A Dynamical Systems Viewpoint. Cambridge University Press, Cambridge, New York (2008)
11. Borwein, J.M., Lewis, A.S.: Convex Analysis and Nonlinear Optimization: Theory and Examples, 2nd edn. Springer, New York (2000)
12. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/ (2011)
13. Chen, X., Pong, T.K., Wets, R.J.B.: Two-stage stochastic variational inequalities: an ERM-solution procedure. Mathematical Programming **165**, 71–111 (2017)
14. Chen, X., Sun, H., Wets, R.J.B.: Regularized mathematical programs with stochastic equilibrium constraints: Estimating structural demand models. SIAM Journal on Optimization **25**, 53–75 (2015)
15. Chen, Y., Lan, G., Ouyang, Y.: Accelerated schemes for a class of variational inequalities. Mathematical Programming **165**, 113–149 (2017)
16. Cheung, Y., Lou, J.: Proximal average approximated incremental gradient descent for composite penalty regularized empirical risk minimization. Machine Learning **106**, 595–622 (2017)
17. Combettes, P.L., Hirstoaga, S.A.: Equilibrium programming in Hilbert spaces. Journal of Nonlinear and Convex Analysis **6**, 117–136 (2005)
18. Crespi, G.P., Guerraggio, A., Rocca, M.: Minty Variational Inequality and Optimization: Scalar and Vector Case. In: Eberhard A., Hadjisavvas N., Luc D.T. (eds) Generalized Convexity, Generalized Monotonicity and Applications. Nonconvex Optimization and Its Applications, **77**, Springer, Boston, MA (2005)
19. Crespi, G.P., Rocca, M.: Minty variational inequalities and monotone trajectories of differential inclusions. Journal of Inequalities in Pure and Applied Mathematics **5**, 48 (2004)
20. Facchinei, F., Kanzow, C.: Penalty methods for the solution of generalized Nash equilibrium problems. SIAM Journal on Optimization **20**, 2228–2253 (2010)
21. Facchinei, F., Pang, J.S.: Finite-Dimensional Variational Inequalities and Complementarity Problems I. Springer, New York (2003)
22. Facchinei, F., Pang, J.S.: Finite-Dimensional Variational Inequalities and Complementarity Problems II. Springer, New York (2003)
23. Fan, L., Liu, S., Gao, S.: Generalized monotonicity and convexity of non-differentiable functions. Journal of Mathematical Analysis and Applications **279**, 276–289 (2003)
24. Floudas, C.A., Visweswaran, V.: A global optimization algorithm (GOP) for certain classes of nonconvex NLPs: I. Theory. Computers and Chemical Engineering **14**, 1397–1417 (1990)
25. Fukushima, M., Pang, J.S.: Some feasibility issues in mathematical programs with equilibrium constraints. SIAM Journal on Optimization **8**, 673–681 (1998)
26. Gao, D.Y.: Canonical duality theory and solutions to constrained nonconvex quadratic programming. Journal of Global Optimization **29**, 377–399 (2004)
27. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization I: A generic algorithmic framework. SIAM Journal on Optimization **22**, 1469–1492 (2012)
28. Ghadimi, S., Lan, G.: Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization II: Shrinking procedures and optimal algorithms. SIAM Journal on Optimization **23**, 2061–2089 (2013)
29. Goebel, K., Kirk, W.A.: Topics in Metric Fixed Point Theory. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge (1990)
30. Goebel, K., Reich, S.: Uniform Convexity, Hyperbolic Geometry, and Nonexpansive Mappings. Dekker, New York (1984)
31. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. Chapman and Hall/CRC, Boca Raton (2015)
32. Iiduka, H.: Fixed point optimization algorithm and its application to power control in CDMA data networks. Mathematical Programming **133**, 227–242 (2012)

33. Iiduka, H.: Iterative algorithm for triple-hierarchical constrained nonconvex optimization problem and its application to network bandwidth allocation. SIAM Journal on Optimization **22**, 862–878 (2012)
34. Iiduka, H.: Stochastic fixed point optimization algorithm for classifier ensemble. IEEE Transactions on Cybernetics **50**, 4370–4380 (2020)
35. Iusem, A.N., Jofré, A., Oliveira, R.I., Thompson, P.: Extragradient method with variance reduction for stochastic variational inequalities. SIAM Journal on Optimization **27**, 686–724 (2017)
36. Iusem, A.N., Sosa, W.: Iterative algorithms for equilibrium problems. Optimization **52**, 301–316 (2003)
37. Iusem, A.N., Sosa, W.: New existence results for equilibrium problems. Nonlinear Analysis: Theory, Methods and Applications **52**, 621–635 (2003)
38. Karamardian, S.: Complementarity problems over cones with monotone and pseudomonotone maps. Journal of Optimization Theory and Applications **18**, 445–454 (1976)
39. Kinderlehrer, D., Stampacchia, G.: An Introduction to Variational Inequalities and Their Applications. Classics Appl. Math. 31. SIAM, Philadelphia (2000)
40. King, A., Rockafellar, R.: Asymptotic theory for solutions in statistical estimation and stochastic programming. Mathematics of Operations Research **18**, 148–162 (1993)
41. Kraft, D.: A software package for sequential quadratic programming. Tech. rep., DFVLR-FB 88-28, DLR German Aerospace Center–Institute for Flight Mechanics, Koln, Germany (1988)
42. Lamm, M., Lu, S., Budhiraja, A.: Individual confidence intervals for solutions to expected value formulations of stochastic variational inequalities. Mathematical Programming **165**, 151–196 (2017)
43. Liu, Y., Xu, H.: Entropic approximation for mathematical programs with robust equilibrium constraints. SIAM Journal on Optimization **24**, 933–958 (2014)
44. Lu, Z., Li, X.: Sparse recovery via partial regularization: Models, theory, and algorithms. Mathematics of Operations Research **43**, 1290–1316 (2018)
45. Luo, Z.Q., Pang, J.S., Ralph, D.: Mathematical Programs with Equilibrium Constraints. Cambridge University Press, Cambridge (1996)
46. Nash, J.F.: Equilibrium points in n-person games. Proceedings of the National Academy of Sciences **36**, 48–49 (1950)
47. Nash, J.F.: Non-cooperative games. Annals of Mathematics **54**, 286–295 (1951)
48. Nedić, A., Lee, S.: On stochastic subgradient mirror-descent algorithm with weighted averaging. SIAM Journal on Optimization **24**, 84–107 (2014)
49. Nemirovski, A., Juditsky, A., Lan, G., Shapiro, A.: Robust stochastic approximation approach to stochastic programming. SIAM Journal on Optimization **19**, 1574–1609 (2009)
50. Nocedal, J., Wright, S.J.: Numerical Optimization, 2nd edn. Springer Series in Operations Research and Financial Engineering. Springer, New York (2006)
51. Ortega, J.M., Rheinboldt, W.C.: Iterative Solution of Nonlinear Equations in Several Variables. Academic Press, Elsevier, Amsterdam (1970)
52. Pang, J.S., Sen, S.E., Shanbhag, U.V.: Two-stage non-cooperative games with risk-averse players. Mathematical Programming **165**, 235–290 (2017)
53. Ravat, U., Shanbhag, U.V.: On the existence of solutions to stochastic quasi-variational inequality and complementarity problems. Mathematical Programming **165**, 291–330 (2017)
54. Robbins, H., Monro, S.: A stochastic approximation method. The Annals of Mathematical Statistics **22**, 400–407 (1951)
55. Rockafellar, R.T., Wets, R.J.B.: Variational Analysis. Springer, Berlin (2010)
56. Saigal, R.: Extension of the generalized complementarity problem. Mathematics of Operations Research **1**, 260–266 (1976)
57. Shalev-Shwartz, S., Singer, Y., Srebro, N., Cotter, A.: Pegasos: Primal estimated subgradient solver for SVM. Mathematical Programming **127**, 3–30 (2011)
58. Shanbhag, U.V., Pang, J.S., Sen, S.: Inexact best-response schemes for stochastic Nash games: Linear convergence and iteration complexity analysis. In: 2016 IEEE 55th Conference on Decision and Control (CDC), pp. 3591–3596 (2016)

59. Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on Stochastic Programming: Modeling and Theory, 2nd edn. MOS-SIAM Series on Optimization. SIAM, Philadelphia (2014)
60. Sundhar Ram, S., Nedić, A., Veeravalli, V.V.: Incremental stochastic subgradient algorithms for convex optimization. SIAM Journal on Optimization **20**, 691–717 (2009)
61. Sundhar Ram, S., Nedić, A., Veeravalli, V.V.: Distributed stochastic subgradient projection algorithms for convex optimization. Journal of Optimization Theory and Applications **147**, 516–545 (2010)
62. Suzuki, S.: Quasiconvexity of sum of quasiconvex functions. Linear and Nonlinear Analysis **3**, 287–295 (2017)
63. Tada, A., Takahashi, W.: Weak and strong convergence theorems for a nonexpansive mapping and an equilibrium problem. Journal of Optimization Theory and Applications **133**, 359–370 (2007)
64. Takahashi, W.: Nonlinear Functional Analysis. Yokohama Publishers, Yokohama (2000)
65. Vapnik, V.N.: Statistical Learning Theory. John Wiley & Sons Inc, Canada (1998)
66. Wang, M., Bertsekas, D.P.: Stochastic first-order methods with random constraint projection. SIAM Journal on Optimization **26**, 681–717 (2016)
67. Xu, H., Ye, J.J.: Necessary optimality conditions for two-stage stochastic programs with equilibrium constraints. SIAM Journal on Optimization **20**, 1685–1715 (2010)
68. Xu, Y., Lin, Q., Yang, T.: Accelerated stochastic subgradient methods under local error bound condition. arXiv, https://arxiv.org/abs/1607.01027
69. Yao, J.C.: Multi-valued variational inequalities with K-pseudomonotone operators. Journal of Optimization Theory and Applications **83**, 391–403 (1994)
70. Yousefian, F., Nedić, A., Shanbhag, U.V.: On smoothing, regularization, and averaging in stochastic approximation methods for stochastic variational inequality problems. Mathematical Programming **165**, 391–431 (2017)
71. Zhang, T.: Analysis of multi-stage convex relaxation for sparse regularization. The Journal of Machine Learning Research **11**, 1081–1107 (2010)
72. Zhou, Z.H.: Ensemble Methods: Foundations and Algorithms. Chapman and Hall/CRC, Boca Raton (2012)