A Single-Projection Proximal Algorithm for Stochastic Mixed Variational Inequalities

Applications to Breast Cancer Screening

Francis O. Nwawuru

Joint work with Mohammad Dilshad, Ibrahim Al-Dayel, and Praveen Agarwal

Department of Mathematics
Chukwuemeka Odumegwu Ojukwu University, Nigeria

Mathematical Optimization and Control for Industrial and Societal Challenges in Sub-Saharan Africa
Pretoria, South Africa — 7–10 September 2026

Talk roadmap

1. Why stochastic mixed variational inequalities?
2. The cost bottleneck and our SiPPA method
3. Convergence analysis
4. Breast-cancer classification study
5. Practical conclusions and open questions

Why this problem matters

Many real-world decision problems combine three important features:

- **Uncertain or sampled information:** The exact operator or expected objective is often unavailable and must be estimated from observations, random samples, or minibatches.
- **Constraints and nonsmooth regularization:** The solution may be required to satisfy feasibility conditions, while terms such as ℓ_1 regularization promote sparsity or control model complexity.
- **Equilibrium or optimality conditions:** The desired solution must balance competing effects or satisfy the first-order condition of an underlying optimization problem.

Why use an SMVIP?

A stochastic mixed variational inequality provides a unified formulation for uncertainty, constraints, nonsmooth terms, and equilibrium conditions.

From stochastic optimization to the SMVIP

The stochastic optimization problem is

$$\min_{x \in X} F(x) := \mathbb{E}[f(x, \xi)], \tag{1}$$

where $X \subseteq \mathbb{R}^n$ is nonempty, closed and convex, ξ is a random vector representing uncertainty, $f(x, \xi)$ is the cost incurred when the decision variable x is chosen and $\mathbb{E}$ denotes expectation over the probability distribution of ξ.

When the objective function F is convex and continuously differentiable on X, the stochastic optimization problem (1) admits a variational characterization via its first-order optimality condition. Specifically, $x \in X$ solves (1) if and only if

$$\langle \nabla F(x), y - x \rangle \geq 0, \ \forall y \in X. \tag{2}$$

Setting $T = \nabla F$ gives a variational inequality.

Problem formulation: stochastic mixed variational inequality

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\xi : \Omega \to \Xi$ be a random variable representing the uncertainty. Let $X \subseteq \mathbb{R}^n$ be nonempty, closed, and convex.
The expected operator $T : X \to \mathbb{R}^n$ is defined by

$$T(x) := \mathbb{E}[T(x, \xi)] = \int_{\Omega} T(x, \xi(\omega))\, d\mathbb{P}(\omega), \tag{3}$$

where $T(x, \cdot)$ is integrable for every $x \in X$. Let $g : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ be proper, lower semicontinuous, and convex. The SMVIP is:

Problem

Find $x^\star \in X$ such that

$$\langle T(x^\star), y - x^\star \rangle + g(y) - g(x^\star) \geq 0, \qquad \forall y \in X. \tag{4}$$

Dilshad et al. (2025); Zhang et al. (2019)

How the SMVIP unifies related models

The general stochastic mixed variational inequality is

$$\langle \mathbb{E}[T(x^\star, \xi)], y - x^\star \rangle + g(y) - g(x^\star) \geq 0, \quad \forall y \in X.$$

1. Stochastic composite optimization
If $T(x, \xi) = \nabla f(x, \xi)$, then the SMVIP is the optimality condition for

$$\min_{x \in X} \mathbb{E}[f(x, \xi)] + g(x).$$

2. Stochastic variational inequality
If $g = \delta_X$, then

$$\langle \mathbb{E}[T(x^\star, \xi)], y - x^\star \rangle \geq 0, \quad \forall y \in X.$$

3. Deterministic mixed VI
If T has no random dependence, then

$$\langle T(x^\star), y - x^\star \rangle + g(y) - g(x^\star) \geq 0, \quad \forall y \in X.$$

4. Stochastic monotone inclusion
When the constraint is absorbed into g, the problem becomes

$$0 \in \mathbb{E}[T(x^\star, \xi)] + \partial g(x^\star).$$

Methods for solving SMVIPs

The expected operator $T(x) = \mathbb{E}[T(x, \xi)]$ is generally unavailable in closed form. Two common approaches are:

Sample average approximation (SAA)

Draw one fixed sample $\{\xi_j\}_{j=1}^N$ and define

$$T_N(x) = \frac{1}{N} \sum_{j=1}^{N} T(x, \xi_j).$$

Then solve the resulting deterministic approximation.

Stochastic approximation (SA)

At each iteration k, draw a fresh sample or minibatch and use

$$T(x_k, \xi_k)$$

to update the current iterate progressively.

Position of SiPPA

SiPPA is SA-based: one fresh minibatch is drawn per iteration, reused at x_k and x_{k-1}, followed by one proximal update.

Robbins and Monro (1951); Jiang and Xu (2008); Liang et al. (2021)

Classical extragradient: prediction and correction

Starting from $x_k \in X$, the stochastic extragradient method performs:

Prediction step: first projection

$$y_k = P_X\big(x_k - \lambda_k T(x_k, \xi_k)\big).$$

The point y_k predicts the next direction using the sampled operator at x_k.

Correction step: second projection

$$x_{k+1} = P_X\big(y_k - \lambda_k T(y_k, \eta_k)\big).$$

Evaluating the operator at y_k corrects and stabilizes the update.

Why seek a single-projection method?

Each iteration requires two projections and two stochastic stages. This motivates the golden-ratio approach used by SiPPA.

Korpelevich (1976); Yang and Lin (2021).

SiPPA: one iteration

For $r \in (1,2)$, let $\theta = \dfrac{1+\sqrt{1+4r}}{2r}$. Initialize $x_{-1} = x_0 = z_0 \in X$ and $\lambda_{-1} = \lambda_0 > 0$.

A. Minibatch and adaptive stepsize

Draw one minibatch $\xi_k = \{\xi_k^j\}_{j=1}^{N_k}$ and compute

$$T(x_i, \xi_k) = \frac{1}{N_k} \sum_{j=1}^{N_k} T(x_i, \xi_k^j), \quad i = k, k-1.$$

Define

$$\Delta x_k = x_k - x_{k-1}, \quad \Delta T_k = T(x_k, \xi_k) - T(x_{k-1}, \xi_k).$$

Choose

$$\lambda_k = \begin{cases} a_1 \dfrac{\|\Delta x_k\|}{\|\Delta T_k\|}, & \|\Delta T_k\| > \dfrac{a_0}{\lambda_{k-1}} \|\Delta x_k\|, \\ (1+\beta_{k-1})\lambda_{k-1}, & \text{otherwise.} \end{cases}$$

B. Averaging and proximal update

Form the golden-ratio averaging point

$$z_k = \frac{(\theta-1)x_k + z_{k-1}}{\theta}.$$

Then perform one proximal update

$$x_{k+1} = \operatorname{prox}_{\lambda_k g}\big(z_k - \lambda_k T(x_k, \xi_k)\big).$$

Work in one iteration

One fresh minibatch, evaluated at two points, and one proximal operation.

Dilshad et al. (2026), Algorithm 1

Main assumptions

- The sampled operator $T(\cdot, \xi)$ is measurable, integrable, and Lipschitz continuous almost surely.
- The expected operator $T(x) = \mathbb{E}[T(x, \xi)]$ is monotone.
- The function g is proper, convex, and lower semicontinuous, with $\mathrm{dom}(g) \cap X \neq \varnothing$.
- The solution set Γ is nonempty.
- Each minibatch contains i.i.d. samples that are independent of the information available before iteration k.
- The oracle error $\varepsilon_k = T(x_k, \xi_k) - T(x_k)$ satisfies

$$\mathbb{E}[\varepsilon_k \mid \mathcal{F}_{k-1}] = 0, \qquad \mathbb{E}[\|\varepsilon_k\|^2 \mid \mathcal{F}_{k-1}] \leq \frac{\sigma^2}{N_k}.$$

- The batch sizes satisfy $\sum_{k=1}^{\infty} N_k^{-1} < \infty$; for example, $N_k = \lceil k^{1+\delta} \rceil$, where $\delta > 0$.

Dilshad et al. (2026)

Convergence guarantees

Almost-sure convergence

Under monotonicity and the stated stochastic assumptions, the sequence $\{x_k\}$ converges almost surely to a solution $x^\star \in \Gamma$ of the SMVIP.

Linear convergence under strong monotonicity

If T is strongly monotone and the stochastic perturbations decay geometrically, then the solution is unique and, for some $C > 0$ and $\bar{\rho} \in (0, 1)$,

$$\mathbb{E}[\|x_k - x^\star\|] \leq C\bar{\rho}^k.$$

The summability condition $\sum_{k=1}^{\infty} N_k^{-1} < \infty$ guarantees sufficient noise control for almost-sure convergence; a geometric rate requires a correspondingly stronger decay of the oracle error.

Dilshad et al. (2026)

Proof strategy: what drives convergence?

1. Proximal optimality produces a one-step Lyapunov inequality.
2. The identity $r\theta^2 = \theta + 1$ reorganizes the cross terms generated by the golden-ratio averaging step.
3. The adaptive stepsize controls the local operator-difference terms.
4. Conditional unbiasedness and the variance estimate lead to a recursion of the form

$$\mathbb{E}[\Psi_{k+1} \mid \mathcal{F}_{k-1}] \leq \Psi_k - D_k + \frac{C}{N_k},$$

   where $\Psi_k \geq 0$ is a Lyapunov function and $D_k \geq 0$ represents the descent terms.
5. Since $\sum_{k=1}^{\infty} N_k^{-1} < \infty$, the Robbins–Siegmund theorem yields almost-sure convergence.
6. Under strong monotonicity, the deterministic part becomes contractive; a geometric stochastic-error decay then gives a linear rate.

Robbins and Siegmund (1972); Dilshad et al. (2026)

Logistic regression as an SMVIP

Let $\xi = (a, b)$, where $a \in \mathbb{R}^n$ is a feature vector and $b \in \{-1, 1\}$ is the corresponding class label. Consider the regularized stochastic logistic-regression problem

$$\min_{x \in X} \mathbb{E}\big[\log\big(1 + \exp(-b\, a^\top x)\big)\big] + g(x),$$

where, for example,

$$g(x) = \tau \|x\|_1 \qquad \text{or} \qquad g(x) = \tau \|x\|_2^2, \qquad \tau > 0.$$

The sampled gradient is

$$T(x, \xi) = -\frac{b\, a}{1 + \exp(b\, a^\top x)}.$$

Therefore, the first-order optimality condition is

$$0 \in \mathbb{E}[T(x^\star, \xi)] + \partial g(x^\star) + N_X(x^\star),$$

which is equivalent to the SMVIP formulation.

Bottou (2010); Sarnmeta et al. (2021)

Breast-cancer outcome-classification study

Dataset and preprocessing

- Public retrospective Kaggle dataset
- 4,024 raw records and 16 original variables
- 4,023 records after deduplication
- No missing values reported
- 20 modelling features after encoding and feature engineering
- 70% training and 30% testing

Prediction target

Binary outcome status: **Alive versus Dead**.

Class imbalance

- Alive: 84.7%
- Dead: 15.3%

Accuracy alone is insufficient; recall, F1-score and AUC must also be examined.

Source: Dilshad et al. (2026); data available from Kaggle.com.

Predictive performance and its interpretation

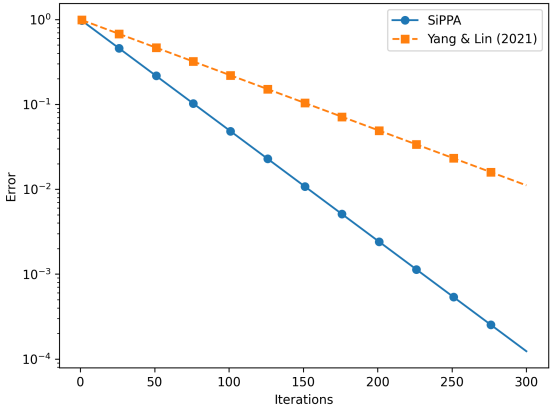
Model	Acc.	Prec.	Recall	F1	AUC
Logistic regression	95.1	94.3	96.0	95.1	0.982
SVM	96.3	95.8	96.9	96.3	0.987
Random forest	96.8	96.5	97.1	96.8	0.989
Yang–Lin	96.9	96.4	97.3	96.8	0.990
SiPPA	**97.6**	**97.1**	**98.4**	**97.7**	**0.995**

Interpretation of the SiPPA results

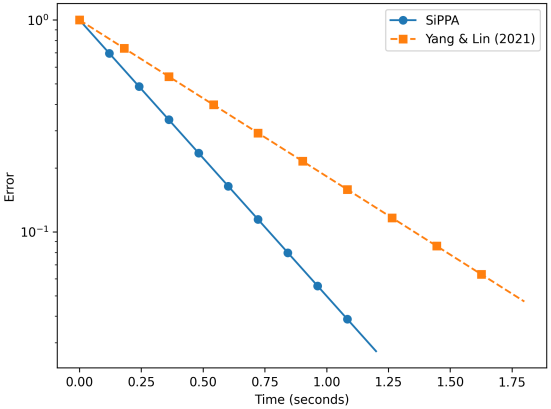
- **Accuracy** $= 97.6\%$**:** 97.6% of all test records were classified correctly as Alive or Dead.
- **Precision** $= 97.1\%$**:** among records predicted as the positive class, 97.1% actually belonged to that class.
- **Recall** $= 98.4\%$**:** among records truly belonging to the positive class, 98.4% were identified.
- **F1-score** $= 97.7\%$**:** precision and recall were jointly high, giving a strong balance between missed and incorrect positive predictions.
- **AUC** $= 0.995$**:** the classifier showed excellent discrimination between the recorded Alive and Dead outcomes across thresholds.

Values are percentages except AUC. Precision and recall depend on which outcome was encoded as the positive class; the published paper does not state this explicitly.

Faster error decay



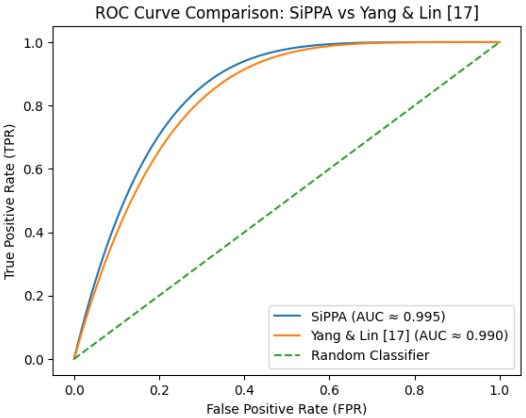
Error versus iterations



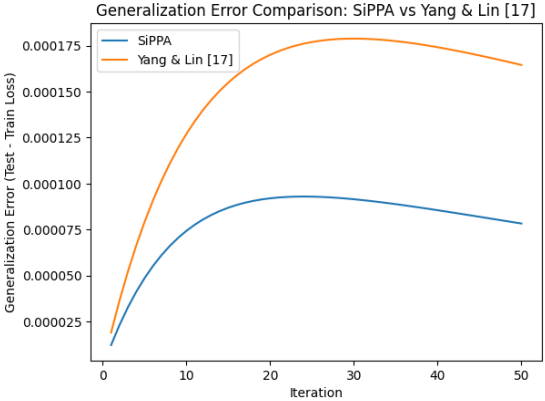
Error versus time

Yang and Lin (2021); Dilshad et al. (2026)

Classification and generalization



ROC comparison: SiPPA AUC $= 0.995$



Smaller, stable generalization gap

Yang and Lin (2021); Dilshad et al. (2026)

Efficiency under ℓ_1 and ℓ_2 regularization

Method	Regularization	F1 (%)	Iterations	Time (s)
Yang–Lin	ℓ_1	96.5	420	1.84
Yang–Lin	ℓ_2	96.7	395	1.72
SiPPA	ℓ_1	**97.7**	**280**	**1.12**
SiPPA	ℓ_2	**97.3**	**260**	**1.05**

SiPPA combines improved predictive performance with fewer iterations and lower runtime in these experiments.

Yang and Lin (2021); Dilshad et al. (2026)

Conclusion

- SMVIPs unify stochastic optimization, variational inequalities, and monotone inclusions.
- SiPPA uses one proximal operation and one fresh minibatch per iteration, reusing that minibatch at two points in the adaptive test.
- The stepsize adapts without a line-search loop or prior knowledge of a global Lipschitz constant.
- Almost-sure convergence holds under monotonicity; under strong monotonicity, a linear rate additionally requires sufficiently fast decay of the stochastic error.
- The breast-cancer study shows promising computational and predictive performance.

Discussion and next steps

- Compare SiPPA with variance-reduced stochastic VI methods.
- Test the method on additional, prospectively collected and externally validated clinical datasets.
- Study class-weighted or cost-sensitive variants for clinically important minority outcomes.
- Extend the analysis beyond monotonicity and investigate distributed implementations.

Thank you

Correspondence: fo.nwawuru@coou.edu.ng
DOI: 10.3934/math.2026434

Selected references

- Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. *Proceedings of COMPSTAT 2010*, 177–186.
- Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics*, 22, 400–407.
- Jiang, H. and Xu, H. (2008). Stochastic approximation approaches to the stochastic variational inequality problem. *IEEE Transactions on Automatic Control*, 53, 1462–1475.
- Korpelevich, G. M. (1976). The extragradient method for finding saddle points and other problems. *Ekon. Mat. Metody*, 12, 747–756.
- Malitsky, Y. (2019). Golden ratio algorithms for variational inequalities. *Mathematical Programming*, 184, 383–410.
- Yang, Z. P. and Lin, G. H. (2021). Variance-based single-call proximal extragradient algorithms for stochastic mixed variational inequalities. *Journal of Optimization Theory and Applications*, 190, 393–427.
- Dilshad, M., Al-Dayel, I., Nwawuru, F. O., and Agarwal, P. (2026). A single-projection proximal algorithm for stochastic mixed variational inequalities with applications to breast cancer screening. *AIMS Mathematics*, 11(4), 10533–10565.