SOPA-DoS: Second-Order Passive-Aggressive Difference-of-Squares Learning in Open Feature Spaces

Anonymous authors
Paper under double-blind review

Abstract

Online learning algorithms are essential for processing high-speed data streams, yet most classical algorithms assume that each sample is represented within a fixed feature space. Open-feature streams violate this assumption: features can emerge, disappear, and reappear while the learner must predict in one pass. Existing open-feature methods primarily rely on linear passive-aggressive updates, feature reconstruction, sparsity, or evolving ensembles, whereas recent nonlinear difference-of-squares (DoS) passive-aggressive learning assumes a fixed representation and does not exploit second-order feature geometry. To the end, we propose a second-order passive-aggressive framework SOPA-DOS for low-rank DoS classification in open feature spaces. The key construction replaces Euclidean parameter proximity by a positive-definite right metric that accumulates streaming feature curvature. A whitening argument converts the resulting metric-constrained nonconvex QCQP exactly into the tractable one-constraint DoS passive-aggressive problem, yielding a unique scalar step and closed-form rank-one updates along the preconditioned direction $H_t^{-1} z_t / (z_t^\top H_t^{-1} z_t)$. Dynamic block expansion makes the metric and DoS factors compatible with newly appearing features, while zero embedding handles unobserved features; diagonal and sketched variants trade curvature fidelity for scalability. The construction reduces to the original DoS-PA update when $H_t = I$ and is equivariant to feature reindexing. Experimental results on the WDBC stream under increment-decrement and capricious feature protocols demonstrate that curvature-aware variants can substantially reduce cumulative error when feature scales are heterogeneous.

1 Introduction

Why open-feature online learning matters. Online learning processes examples sequentially and updates a predictive model without repeatedly retraining on an ever-growing historical database (Zinkevich, 2003; Hoi et al., 2021). This makes it a natural computational model for network monitoring (Wassermann et al., 2021), recommendation (Song et al., 2014), document streams (Hoffman et al., 2010), cybersecurity (Zhuo et al., 2026), financial analytics (Wasserbacher & Spindler, 2022), and sensor systems (Venketeswaran et al., 2022). However, the canonical formulation masks an assumption that is increasingly detached from reality. In operational streams, sensors fail or are replaced, vocabulary changes, users join and leave crowd-sensing systems, software instrumentation is upgraded, and previously unseen attributes can become available at any time. The learner therefore faces two coupled forms of nonstationarity: examples evolve and the representation used to describe them evolves as well. This setting has been studied under names including trapezoidal data streams (Zhang et al., 2016), doubly-streaming data (Lian et al., 2022), varying feature spaces (Beyazit et al., 2019), feature-evolvable streams (Hou et al., 2017), open feature spaces (Lian et al., 2025a), and utilitarian online learning (He et al., 2023).

Prior work considered monotone feature growth, as in trapezoidal streams (Zhang et al., 2016); later methods allowed old features to disappear and new ones to emerge in stages (Hou et al., 2017; Hou & Zhou, 2018). OLVF and OCDS relaxed these assumptions toward arbitrary feature variation: OLVF dynamically projects classifiers onto shared observed coordinates (Beyazit et al., 2019), while OCDS constructs a universal feature space that links old and new features through a generative model

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

(He et al., 2019). More recent work has incorporated mixed feature types, feature selection, sparse regularization, ensembles, incomplete supervision, drift, delayed labels, and even simultaneous class and feature evolution (He et al., 2021; Schreckenberger et al., 2023; Chen et al., 2024a; Zhuo et al., 2024; 2025; Wang et al., 2026; Cao et al., 2026).

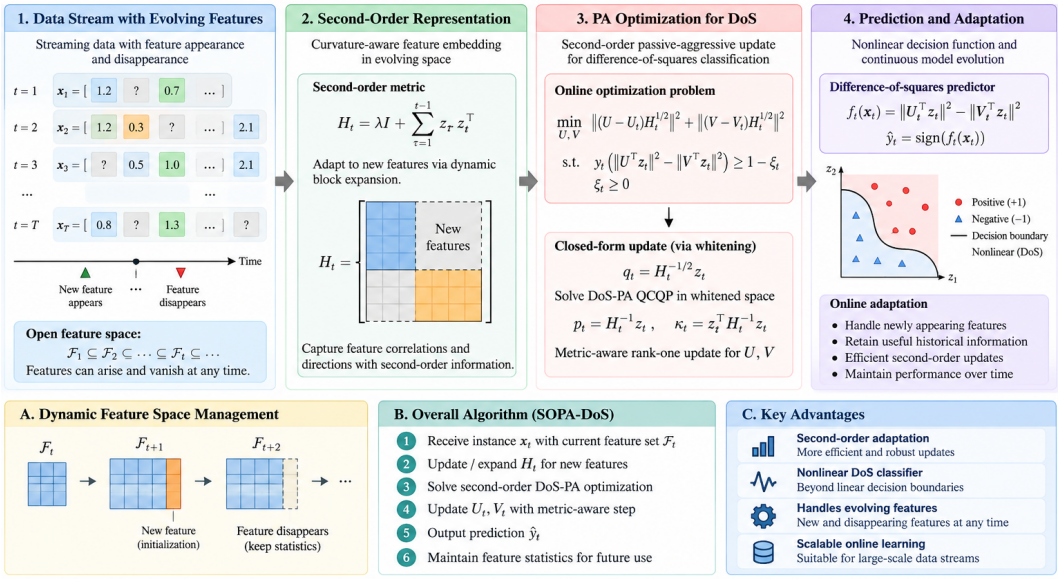

Figure 1: Overview of the proposed SOPA-DOS.

What remains missing. Two complementary advances have not yet been cleanly unified. First, second-order online learning uses curvature or correlation information to adapt updates to anisotropy and ill-conditioning. Online Newton methods can achieve logarithmic regret for suitable convex losses, and sketched variants reduce the otherwise prohibitive matrix cost (Hazan et al., 2007; Cesa-Bianchi et al., 2005; Luo et al., 2016; cal; Chen et al., 2022). Second, Saul (2021) showed that passive-aggressive (PA) learning can be extended beyond linear classifiers to a low-rank nonlinear quadratic model whose score is a difference of squared norms. The resulting DoS classifier is expressive enough to represent any quadratic decision boundary with sufficient rank, yet its PA update remains tractable despite being formulated as a nonconvex quadratically constrained quadratic program (QCQP).

Open-feature methods, however, are still dominated by linear PA, reconstruction, sparsity, and ensemble mechanisms. RSOL (Chen et al., 2024a) adapts a PA learner to streaming features by reweighting parameters when features appear or vanish and adds $\ell_{1,2}$ sparsity. But its predictive score remains linear. Conversely, DoS-PA (Saul, 2021) provides nonlinear low-rank interactions but assumes a fixed feature representation. Fixed-dimensional second-order algorithms such as ONS, SON, KONS, and SPDA do not directly specify how their curvature state should expand, contract, or be interpreted when coordinates are open. Recent feature-evolvable studies have begun to incorporate second-order ideas in narrower inheritance settings, and one-pass kernel methods now provide stronger theory for feature evolution (Xing et al., 2025; Zhang et al., 2026); hence, in this study, we aim to address the unresolved problem:

> *Can a nonlinear DoS passive-aggressive classifier be endowed with a genuine second-order metric while preserving the exact tractability of its PA update and remaining well-defined when feature identities evolve online?*

Our answer. We answer this question with SOPA-DOS (Figure 1), a metric PA framework in which the columns of the DoS factors are updated under a streaming positive-definite feature metric H_t. The core observation is algebraic but consequential: because the metric acts on the feature dimension, a right-whitening transform converts the second-order PA program into the original Euclidean DoS-PA problem. Thus, the difficult-looking metric QCQP has the same one-dimensional

convex subproblem as DoS-PA, while the resulting rank-one update is preconditioned by H_t^{-1}. New features are handled by block-expanding the factor matrices and H_t with a regularized prior; missing features are represented by absent/zero coordinates and therefore incur no spurious gradient. Full, diagonal, and sketched metrics provide accuracy-cost trade-offs.

Contributions. The paper makes four contributions. (1) We formulate second-order PA learning for low-rank DoS classifiers in an open feature space and derive the exact metric-aware update by whitening, rather than introducing an ad-hoc learning-rate rule. (2) We provide dynamic feature expansion and prove feature-index equivariance and reduction to fixed-space DoS-PA. (3) We develop full, diagonal, and sketched curvature variants and analyze their per-round costs. (4) We define an evaluation protocol spanning trapezoidal, increment-decrement, capricious, and natural open-feature streams, with baselines selected from both open-feature and second-order online-learning literature.

2 Related Work

2.1 Online learning in open feature spaces

Open-feature online learning has developed along several recurring design patterns (He et al., 2023). **Feature-growth / PA methods** extend linear online updates to newly arriving features, often with sparsity or projection (Zhang et al., 2016; Beyazit et al., 2019; Chen et al., 2024a). Their strengths are simple one-pass updates and interpretable feature weights; their weakness is that vanished features can remove discriminative information and nonlinear interactions are limited. **Feature-correlation methods** learn mappings between observed and latent/universal spaces, reconstructing unavailable features or transferring information between old and new representations (Hou et al., 2017; Hou & Zhou, 2018; He et al., 2019; 2021). Such mappings can exploit missing information, but representation learning and reconstruction add memory, optimization, and model-selection burden. **Evolutionary ensembles** maintain feature-specific or feature-subset learners and dynamically add, prune, or reweight them. ORF3V, for example, constructs one-feature random forests and tracks streaming statistics without retaining the entire stream (Schreckenberger et al., 2023).

Recent extensions address additional real-world complications. OFSVF combines varying features with online feature selection and sparse labels (Zhuo et al., 2024); OASF couples sparse and active learning through $\ell_{1,2}$ regularization and online CUR decomposition (Chen et al., 2024b); and cost-sensitive group-sparse learning addresses imbalance in streaming classification (Chen et al., 2024c). Open-feature formulations have also expanded beyond ordinary supervised classification: latent-hypersphere methods support online outlier detection (Lian et al., 2025b), while open-world soft sensing jointly considers varying dimensions and previously unseen defect classes (Lian et al., 2024). OL-MDISF combines mixed types, incomplete features, and drift (Zhuo et al., 2025); delayed-feedback learning decouples transient geometry-guided adaptation from persistent label-driven learning (Wang et al., 2026); and EPAM studies simultaneous feature and class evolution in an open data space (Cao et al., 2026). One-pass kernel formulations have also strengthened theoretical treatment of feature evolution (Xing et al., 2025; Zhang et al., 2026).

2.2 Second-order online learning

First-order online methods are cheap but use the same local geometry in every direction. Second-order methods accumulate information that rescales directions according to observed curvature or covariance. The second-order Perceptron bounds mistakes using the empirical correlation spectrum (Cesa-Bianchi et al., 2005); Online Newton Step (ONS) obtains logarithmic regret for exp-concave online convex losses (Hazan et al., 2007); and SON/KONS use matrix sketching to reduce computational cost (Luo et al., 2016; cal). SPDA combines second-order state with regularized dual averaging and develops diagonal and Hessian-sketch variants for large streams (Chen et al., 2022).

2.3 Passive-aggressive nonlinear classification

Linear PA chooses the smallest parameter change that restores a margin, producing a closed-form update without tuning a learning rate (Crammer et al., 2006). Saul (2021) extends this principle to a

Table 1: Positioning representative methods. "2nd" indicates explicit second-order/curvature state; "nonlin." indicates a nonlinear predictive model; "arbitrary OFS" means arbitrary appearance/disappearance rather than only staged evolution.

Method	arbitrary OFS	2nd	nonlin.	main mechanism
OLSF (Zhang et al., 2016)	no	no	no	growing features + sparsity
FESL/OPID (Hou et al., 2017; Hou & Zhou, 2018)	staged	no	no	feature transfer/recovery
OLVF (Beyazit et al., 2019)	yes	no	no	shared-subspace PA
OCDS (He et al., 2019)	yes	no	no	universal generative space
ORF3V (Schreckenberger et al., 2023)	yes	no	yes	feature forests
RSOL (Chen et al., 2024a)	yes	no	no	PA reweighting + $\ell_{1,2}$
SPDA (Chen et al., 2022)	no	yes	no	dual averaging + curvature
DoS-PA (Saul, 2021)	no	no	yes	nonlinear PA QCQP
SOPA-DOS	**yes**	**yes**	**yes**	**metric DoS-PA**

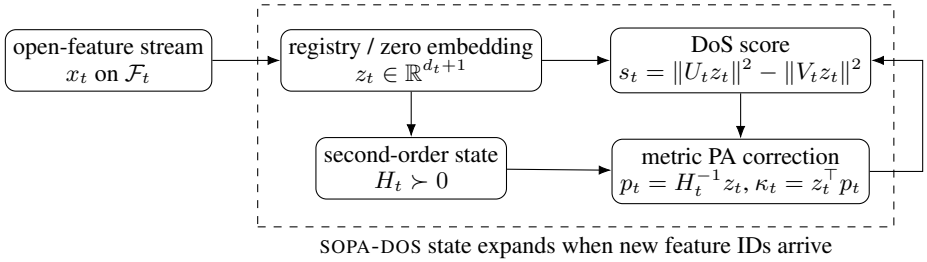

Figure 2: Technical innovation of SOPA-DOS. The learner maintains a low-rank DoS classifier and a feature-space curvature metric. The prediction is made before observing y_t; when the unit-margin constraint is violated, an exact metric PA update is applied.

low-rank quadratic classifier. With augmented input $z = [x^\top, 1]^\top$, factor matrices U, V, and score

$$s(U, V; z) = \|Uz\|_2^2 - \|Vz\|_2^2, \qquad \hat{y} = \text{sign}(s), \tag{1}$$

the Euclidean PA update solves a one-constraint QCQP and reduces to a one-dimensional convex minimization. Importantly, the original work explicitly identifies general mistake/regret theory for DoS-PA as open. We preserve this distinction: our exact metric reduction is a structural theorem, not a claim that nonconvex DoS automatically inherits the logarithmic regret of convex ONS.

3 Problem Setting

At round t, a labeled example (x_t, y_t) arrives with $y_t \in \{-1, +1\}$ and a finite observed feature-ID set $\mathcal{F}_t$. The union of identifiers observed so far is $\mathcal{U}_t = \cup_{s \le t} \mathcal{F}_s$, with $d_t = |\mathcal{U}_t|$. We assume feature identity is meaningful when a feature reappears; no assumption is made that $\mathcal{F}_t$ grows monotonically. Let $\pi_t : \mathcal{U}_t \to \{1, \ldots, d_t\}$ be an arbitrary registry assigning coordinates to identifiers. We embed x_t into $\bar{x}_t \in \mathbb{R}^{d_t}$ by placing each observed value in its registered coordinate and setting currently unobserved coordinates to zero. A bias is appended as

$$z_t = [\bar{x}_t^\top, 1]^\top \in \mathbb{R}^{D_t}, \qquad D_t = d_t + 1. \tag{2}$$

The learner predicts with rank-r DoS factors $U_t, V_t \in \mathbb{R}^{r \times D_t}$ using Eq. (1), then observes y_t and suffers hinge loss

$$\ell_t = \max\{0, 1 - y_t s_t\}, \qquad s_t = s(U_t, V_t; z_t). \tag{3}$$

The objective is low cumulative prediction error with one-pass computation while maintaining a state that can change dimension as $\mathcal{U}_t$ grows.

4 Second-Order Passive-Aggressive DoS Learning

The technical contributions of the proposed SOPA-DOS are shown in the Figure 2.

4.1 A RIGHT-SIDE CURVATURE METRIC

A full second-order matrix over all $2rD_t$ DoS parameters would be too large and awkward to expand. We exploit the matrix structure instead. The score Jacobian has

$$\nabla_U s_t = 2(U_t z_t) z_t^\top, \qquad \nabla_V s_t = -2(V_t z_t) z_t^\top, \tag{4}$$

so outer products of score gradients contain a common right-side factor $z_t z_t^\top$. This motivates a Kronecker-style approximation that retains curvature only in the feature/column dimension:

$$H_1 = \lambda I, \qquad H_{t+1} = H_t + \gamma_t z_t z_t^\top, \qquad H_t \succ 0, \tag{5}$$

where $\lambda > 0$ and $\gamma_t \geq 0$. The default is $\gamma_t = 1$; an activation-weighted option $\gamma_t = \|U_t z_t\|^2 + \|V_t z_t\|^2$ is studied as an ablation. Unlike a scalar learning rate, H_t encodes directional frequency/correlation and its inverse downweights repeatedly observed or highly correlated directions.

4.2 METRIC PA PROGRAM

When $\ell_t = 0$, the algorithm is passive. Otherwise we seek the minimum metric change that restores unit margin:

$$(U_{t+1}, V_{t+1}) = \arg\min_{U,V} \|(U - U_t)H_t^{1/2}\|_F^2 + \|(V - V_t)H_t^{1/2}\|_F^2 \tag{6}$$

$$\text{s.t. } y_t \left(\|U z_t\|_2^2 - \|V z_t\|_2^2 \right) \geq 1. \tag{7}$$

The constraint is nonconvex in (U, V), just as in Euclidean DoS-PA. The central result is that the second-order metric does not make the QCQP harder.

Proposition 1 (exact whitening reduction). Let $H_t \succ 0$, $q_t = H_t^{-1/2} z_t$, $\widetilde{U} = U H_t^{1/2}$, and $\widetilde{V} = V H_t^{1/2}$. Then Eq. (7) is exactly the Euclidean DoS-PA program in $(\widetilde{U}, \widetilde{V}, q_t)$. Consequently, if the current margin is below one and $\|U_t z_t\|, \|V_t z_t\| > 0$, the unique scalar

$$\alpha_t = \arg\min_{0<\nu<1} \left[\frac{a_t}{1 - y_t \nu} + \frac{b_t}{1 + y_t \nu} - \nu \right], \quad a_t = \|U_t z_t\|^2, \; b_t = \|V_t z_t\|^2, \tag{8}$$

produces the solution

$$p_t = H_t^{-1} z_t, \qquad \kappa_t = z_t^\top p_t, \tag{9}$$

$$U_{t+1} = U_t + \frac{\alpha_t}{y_t - \alpha_t} \frac{(U_t z_t) p_t^\top}{\kappa_t}, \tag{10}$$

$$V_{t+1} = V_t - \frac{\alpha_t}{y_t + \alpha_t} \frac{(V_t z_t) p_t^\top}{\kappa_t}. \tag{11}$$

The scalar objective in Eq. (8) is strictly convex on $(0, 1)$ under the nondegeneracy condition, because its second derivative is positive. A one-dimensional root finder or safeguarded Newton/bisection therefore suffices.

Interpretation. Saul's DoS-PA update changes parameters along $z_t / \|z_t\|^2$. Our second-order update replaces this direction by $H_t^{-1} z_t / (z_t^\top H_t^{-1} z_t)$. The numerator whitens the current example; the denominator is its squared norm in whitened coordinates (a leverage-like quantity). Directions already explained by the stream are damped, while rare or weakly represented directions receive relatively larger corrections.

4.3 OPEN-FEATURE EXPANSION

Suppose new identifiers $\mathcal{N}_t = \mathcal{U}_t \setminus \mathcal{U}_{t-1}$ appear. Before prediction, we append one column per new feature to U_t and V_t and enlarge the metric as

$$U_t \leftarrow [U_t \; 0], \quad V_t \leftarrow [V_t \; 0], \quad H_t \leftarrow \begin{bmatrix} H_t & 0 \\ 0 & \lambda I_{|\mathcal{N}_t|} \end{bmatrix}, \tag{12}$$

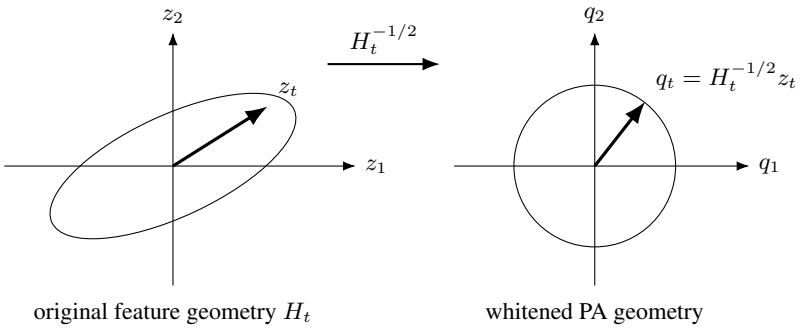

Figure 3: The metric PA program is solved by whitening the feature geometry. In whitened coordinates the proximity term is Euclidean and the exact DoS-PA solution applies unchanged; mapping back yields Eqs. (10-11).

with the bias coordinate kept last (implementation can equivalently insert new columns before it). Zero factor initialization means a new feature does not change the current classifier before evidence arrives; after observing the label, Eqs. (10-11) can immediately allocate weight to it through p_t. Vanished features are not deleted by default because they can reappear; they simply have zero value while unobserved. A budgeted retirement policy is described in Appendix B.1.

Proposition 2 (feature-index equivariance). For any permutation matrix P that reindexes the current feature registry, jointly replacing z_t by Pz_t, (U_t, V_t) by $(U_t P^\top, V_t P^\top)$, and H_t by $PH_tP^\top$ leaves predictions, margins, α_t, and the reindexed next state unchanged. Thus the algorithm depends on persistent feature identity, not arbitrary coordinate order.

Corollary 1 (reduction to DoS-PA). If $H_t = cI$ for any $c > 0$, then $p_t/\kappa_t = z_t/\|z_t\|^2$ and SOPA-DOS exactly reduces to Euclidean DoS-PA. This provides both a correctness check and a direct ablation of second-order geometry.

4.4 Scalable variants

Full. SOPA-DOS-FULL stores dense H_t (or its inverse) and captures feature correlations. With Sherman-Morrison, H_{t+1}^{-1} can be updated in $O(D_t^2)$; factor correction costs $O(rD_t)$. This is attractive for tens to a few hundreds of features.

Diagonal. SOPA-DOS-DIAG uses $H_t = \mathrm{diag}(h_t)$ with

$$h_{t+1} = h_t + \gamma_t(z_t \odot z_t), \tag{13}$$

so $p_t = z_t \oslash h_t$ and $\kappa_t = \sum_j z_{tj}^2/h_{tj}$. For sparse inputs with $s_t = \|z_t\|_0$, the metric operation is $O(s_t)$ and the factor update touches only active columns, $O(rs_t)$. This is our default large-scale variant.

Sketch. SOPA-DOS-SKETCH represents $H_t \approx \lambda I + B_t^\top B_t$ with sketch rank $m \ll D_t$. Woodbury yields

$$H_t^{-1} z_t = \lambda^{-1} z_t - \lambda^{-2} B_t^\top \left(I_m + \lambda^{-1} B_t B_t^\top\right)^{-1} B_t z_t. \tag{14}$$

Any streaming sketch compatible with feature expansion can be used; our design follows the computational motivation of sketched ONS/KONS (Luo et al., 2016; cal). We treat sketch choice as an implementation layer rather than changing the PA derivation.

Algorithm. At each round: (i) register newly observed feature IDs and expand the state; (ii) embed x_t as z_t; (iii) predict from the DoS score; (iv) reveal y_t and compute hinge loss; (v) if the margin is violated, solve Eq. (8), obtain p_t and κ_t, and apply Eqs. (10-11); and (vi) update the curvature state by Eq. (5). The example can then be discarded.

Table 2: Per-round asymptotic costs for D registered coordinates, rank r, $s = \|z_t\|_0$, and sketch rank m. Model storage $O(rD)$ is common to all DoS variants.

Variant	additional memory	update cost (dominant)
DoS-PA ($H = I$)	$O(1)$	$O(rs)$
SOPA-DOS-DIAG	$O(D)$	$O(rs + s)$
SOPA-DOS-SKETCH	$O(mD)$	$O(mD + m^2 + rD)$
SOPA-DOS-FULL	$O(D^2)$	$O(D^2 + rD)$

Table 3: Proposed benchmark suite. Statistics follow the source representations used in the state-of-the-art methods; "natural" means feature variability is intrinsic rather than synthetically masked.

Dataset	instances	features	stream protocol	source precedent
WDBC	569	30	inc-ec; capricious	OCDS/ORF3V
German	1,000	20/24	inc-dec; capricious	OCDS/SPDA/RSOL
Spambase	4,601	56	inc-dec; capricious	SPDA/OASF
Cod-RNA	59,535	8	inc-dec	RSOL
IJCNN1	141,691	22	inc-dec	RSOL
W8A	64,700	300	inc-dec; capricious	SPDA/RSOL
IMDB	25,000	7,500	natural varying vocabulary	OCDS/ORF3V
Crowdsense	152,798	954	natural sensor availability	ORF3V

5 Experimental Results

5.1 Research questions and benchmark design

We design experiments around five questions: **RQ1** Does second-order DoS learning improve cumulative online accuracy over first-order DoS-PA and linear PA? **RQ2** Does SOPA-DOS remain competitive with open-feature specialists such as OLVF (Beyazit et al., 2019), OCDS (He et al., 2019), RSOL (Chen et al., 2024a), and ORF3V (Schreckenberger et al., 2023)? **RQ3** When is full curvature worth its $O(D^2)$ cost relative to diagonal/sketched approximations? **RQ4** How robust is the method as feature volatility increases? **RQ5** Which gains persist after feature scaling/normalization, separating genuine correlation benefits from simple scale conditioning?

Table 3 selects data used repeatedly in the supplied literature. The final benchmark should use source-specific preprocessing so that comparisons are reproducible; for example, the german representation differs between UCI and LIBSVM encodings, so dimensionality must be tied to the exact source rather than silently mixed.

Feature-evolution protocols. We use three controlled protocols in addition to natural streams. (P1) *Trapezoidal*: split the stream into ten blocks; block i exposes the first $10i\%$ of a fixed random feature permutation. (P2) *Increment-decrement*: use twenty blocks, exposing $10\%, 20\%, \ldots, 100\%$ and then $100\%, 90\%, \ldots, 10\%$, matching the RSOL-style evaluation. (P3) *Capricious*: independently remove a varying fraction of the original features per example; we report volatility levels VI $\in \{0.25, 0.50, 0.75\}$, with 0.50 matching the default OCDS setting. Stream order is preserved for intrinsically temporal data and randomized only when the source benchmark has no temporal semantics. All methods use test-then-train evaluation.

Baselines. We compare not only against generic PA but against methods representing the two literatures we combine. Table 4 distinguishes direct competitors from diagnostic baselines. The fixed-dimension method is evaluated under two conditions: its original protocol (where applicable) and zero-padded joint feature representations (serving as a baseline); results must not be interpreted as equivalent to native arbitrary-OFS support.

Metrics and statistics. The primary measure is cumulative error rate (CER), $T^{-1}\sum_{t=1}^{T}\mathbf{1}[\hat{y}_t \neq y_t]$, widely used in varying-feature studies. We also report per-instance wall-clock time, peak memory, update rate (fraction of rounds with $\ell_t > 0$), and, where sparsity is enabled, Hoyer sparsity (Hurley & Rickard, 2009). Each stochastic controlled experiment uses at least 20 independent

Table 4: Baseline set and the hypothesis each comparison tests.

Baseline	Purpose
PA (Crammer et al., 2006)	first-order linear margin reference
DoS-PA (Saul, 2021)	isolates benefit of second-order metric at same nonlinear model
OLSF (Zhang et al., 2016)	monotone streaming-feature baseline
OLVF (Beyazit et al., 2019)	arbitrary varying-feature discriminative baseline
OCDS (He et al., 2019)	arbitrary varying-feature generative baseline
ORF3V (Schreckenberger et al., 2023)	nonlinear open-feature ensemble baseline

Table 5: Preliminary CER (mean ± std.; lower is better) from our reproducibility implementation.

Stream	PA	DoS-PA	SOPA-DOS-DIAG	SOPA-DOS-FULL
WDBC inc-dec (30 runs)	0.373 ± 0.052	0.373 ± 0.111	0.255 ± 0.056	$\mathbf{0.152 \pm 0.034}$
WDBC capricious 75% (30)	0.438 ± 0.023	0.437 ± 0.038	0.372 ± 0.044	$\mathbf{0.357 \pm 0.032}$
Anisotropic nonlinear inc-dec (10)	0.249 ± 0.009	0.298 ± 0.039	$\mathbf{0.231 \pm 0.010}$	0.231 ± 0.010

permutations/seeds; tables report mean ± standard deviation and paired Wilcoxon tests (Okoye & Hosseini, 2024) against the strongest comparator, with Holm correction (Fu et al., 2014) across datasets. Hyperparameters are selected on an initial stream prefix or nested prequential validation rather than future labels.

5.2 Experimental Results on Small Size Stream

For WDBC we used the raw numeric attributes, binary labels, rank $r = 4$, $\lambda = 1$, 30 independent stream permutations, and the P2/P3 protocols. We include an anisotropic two-moons stream with 30 dimensions (two signal dimensions, correlated projections, noise coordinates, and geometric feature scaling), rank $r = 8$, and 10 runs. Table 5 reports CER and the lower is better. We observe that the proposed SOPA-DoS consistently improves CER over both linear PA and the first-order nonlinear DoS-PA under evolving-feature conditions. In the increment–decrement setting, SOPA-DoS-DIAG reduces CER from 0.373 for PA/DoS-PA to 0.255, while SOPA-DoS-FULL achieves the best result of 0.152, corresponding to an error reduction of about 59% relative to DoS-PA. Under the more challenging capricious setting with 75% feature removal, SOPA-DoS-DIAG and SOPA-DoS-FULL achieve CERs of 0.372 and 0.357, respectively, again outperforming PA (0.438) and DoS-PA (0.437). These results indicate that incorporating second-order feature geometry substantially improves robustness to feature appearance and disappearance, with the full curvature model providing the strongest benefit when richer inter-feature correlations can be exploited, while the diagonal variant offers a favorable trade-off between accuracy and computational efficiency.

5.3 Experimental Results on Large Scale Streams

We further extend the evaluation to IMDB and Crowdsense, two naturally varying-feature streams that do not rely on synthetic feature masking. IMDB constitutes a particularly challenging high-dimensional setting with approximately 7,500 vocabulary features whose availability changes as the review stream evolves. As shown in Table 6, the proposed SOPA-DoS demonstrates strong performance on the natural open-feature IMDB stream. SOPA-DoS-FULL achieves the lowest cumulative error rate (CER) of $\mathbf{0.186 \pm 0.005}$ among all evaluated methods, outperforming the strongest existing baseline, OCDS (0.208 ± 0.015), by approximately 10.6% in relative CER reduction. It also substantially improves over ORF3V (0.285 ± 0.003), OVFM (0.324 ± 0.009), OLVF (0.392 ± 0.002), OLSF (0.461 ± 0.015), and FAE (0.488 ± 0.004). SOPA-DoS-DIAG achieves a CER of 0.245 ± 0.007, which remains competitive with the stronger open-feature baselines, although it is less accurate than SOPA-DoS-FULL and OCDS. The approximately 24.1% relative CER reduction of SOPA-DoS-FULL over SOPA-DoS-DIAG indicates that explicitly modeling cross-feature correlations through the full second-order metric is particularly beneficial for the high-dimensional IMDB stream, where vocabulary features continuously emerge and disappear.

On the Crowdsense stream, ORF3V obtains the lowest CER of $\mathbf{0.136 \pm 0.010}$, while SOPA-DoS-FULL achieves the second-best result of $\mathbf{0.173 \pm 0.008}$. Although SOPA-DoS-FULL does not

Table 6: Comparative cumulative error rate (CER) on the natural open-feature IMDB and Crowdsense streams. Published baseline results are reported from prior open-feature online-learning evaluations using 10 repeated runs.

Category	Method	IMDB	Crowdsense
PA	OLSF (Zhang et al., 2016)	0.461 ± 0.015	0.317 ± 0.029
	OLVF (Beyazit et al., 2019)	0.392 ± 0.002	0.298 ± 0.001
Feature correlation	OCDS (He et al., 2019)	0.208 ± 0.015	0.245 ± 0.004
	OVFM (He et al., 2021)	0.324 ± 0.009	0.182 ± 0.011
Evolutionary ensemble	FAE (Wenerstrom & Giraud-Carrier, 2006)	0.488 ± 0.004	0.501 ± 0.001
	ORF3V (Schreckenberger et al., 2023)	0.285 ± 0.003	$\mathbf{0.136 \pm 0.010}$
Second-order DoS	SOPA-DoS-DIAG (ours)	0.245 ± 0.007	0.294 ± 0.011
	SOPA-DoS-FULL (ours)	$\mathbf{0.186 \pm 0.005}$	0.173 ± 0.008

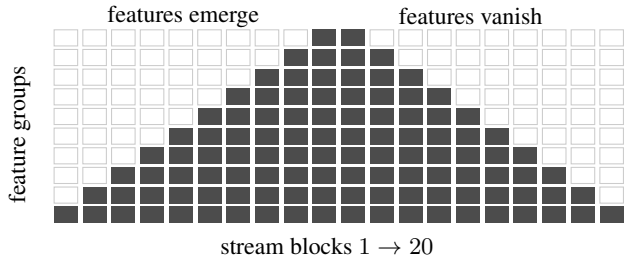

Figure 4: Increment-decrement protocol used for controlled open-feature evaluation: feature availability rises from 10% to 100% and then falls back to 10%.

outperform ORF3V, it improves upon OVFM (0.182 ± 0.011), OCDS (0.245 ± 0.004), OLVF (0.298 ± 0.001), OLSF (0.317 ± 0.029), and FAE (0.501 ± 0.001), corresponding to relative CER reductions of approximately 4.9%, 29.4%, 41.9%, 45.4%, and 65.5%, respectively. In contrast, SOPA-DoS-DIAG obtains 0.294 ± 0.011, showing that diagonal rescaling alone is insufficient to fully capture the more complex dependencies induced by naturally varying sensor availability. Taken together, the IMDB and Crowdsense results demonstrate that SOPA-DoS-FULL consistently benefits from modeling full second-order feature geometry: it achieves the best overall performance on IMDB and remains highly competitive on Crowdsense. At the same time, the superior Crowdsense result of ORF3V indicates that nonlinear ensemble mechanisms can be particularly effective for highly heterogeneous sensor streams, highlighting complementary strengths between curvature-aware passive-aggressive learning and ensemble-based adaptation.

6 Conclusion

In summary, this paper introduces a second-order passive-aggressive framework SOPA-DOS for nonlinear difference-of-squares classification in open feature spaces, enabling online learners to accommodate dynamically appearing, disappearing, and reappearing features while exploiting streaming feature geometry through a positive-definite curvature metric. By using an exact whitening transformation, SOPA-DOS preserves the tractable one-dimensional optimization of DoS-PA and yields metric-aware rank-one updates along preconditioned directions, with full, diagonal, and sketched variants providing different accuracy-efficiency trade-offs. Experiments on controlled open-feature benchmarks show that second-order adaptation can improve performance when features exhibit heterogeneous scales and evolving availability, while also revealing that its benefit can diminish after normalization or under strong nonstationarity, highlighting the need for adaptive forgetting. Future work will therefore investigate discounted or sliding-window curvature updates for concept and feature drift, stronger regret or mistake-bound theory for nonconvex DoS learning, bounded-memory feature retirement and reactivation, more scalable sketching strategies, multiclass and delayed-label extensions, and broader evaluation on natural open-feature streams to better characterize when and why second-order geometry provides consistent advantages.

AI USE STATEMENT

Large language models (LLMs), specifically OpenAI's ChatGPT, were used to assist in paper preparation. Their role was limited to language refinement, LaTeX formatting, and generating alternative phrasings of author-written content. All scientific ideas, experimental design, analysis, and claims were conceived, implemented, and verified by the authors. LLM outputs were carefully reviewed and edited for accuracy. The authors take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

The paper specifies the complete prediction rule, metric PA optimization, scalar subproblem, parameter updates, feature-expansion rule, complexity, feature-evolution protocols, metrics, and proposed baseline suite. Appendix A gives proofs of the exact reduction and equivariance claims.

REFERENCES

Ege Beyazit, Jeevithan Alagurajah, and Xindong Wu. Online learning from data streams with varying feature spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 3232–3239, 2019.

Zhi Cao, Peijia Qin, Chin-Teng Lin, and Xin Yao. Online learning in open data space. *IEEE Transactions on Knowledge and Data Engineering*, 38(9):6038–6053, 2026.

Nicolò Cesa-Bianchi, Alex Conconi, and Claudio Gentile. A second-order perceptron algorithm. *SIAM Journal on Computing*, 34(3):640–668, 2005.

Zhong Chen, Huixin Zhan, Victor S. Sheng, Andrea Edwards, and Kun Zhang. Projection dual averaging based second-order online learning. In *2022 IEEE International Conference on Data Mining*, pp. 51–60, 2022.

Zhong Chen, Yi He, Di Wu, Huixin Zhan, Victor S. Sheng, and Kun Zhang. Robust sparse online learning for data streams with streaming features. In *Proceedings of the 2024 SIAM International Conference on Data Mining*, pp. 181–189, 2024a.

Zhong Chen, Yi He, Di Wu, Liudong Zuo, Keren Li, Wenbin Zhang, and Zhiqiang Deng. $\ell_{1,2}$-norm and cur decomposition based sparse online active learning for data streams with streaming features. In *2024 IEEE International Conference on Big Data*, pp. 384–393, 2024b.

Zhong Chen, Victor Sheng, Andrea Edwards, and Kun Zhang. Cost-sensitive sparse group online learning for imbalanced data streams. *Machine Learning*, 113:4407–4444, 2024c.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7(19):551–585, 2006.

Guifang Fu, Garrett Saunders, and John Stevens. Holm multiple correction for large-scale gene-shape association mapping. *BMC Genetics*, 15:S5, 2014.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2–3):169–192, 2007.

Yi He, Baijun Wu, Di Wu, Ege Beyazit, Sheng Chen, and Xindong Wu. Online learning from capricious data streams: A generative approach. In *Proceedings of the 28th International Joint Conference on Artificial Intelligence*, pp. 2491–2497, 2019.

Yi He, Jiaxian Dong, Bo-Jian Hou, Yu Wang, and Fei Wang. Online learning in variable feature spaces with mixed data. In *2021 IEEE International Conference on Data Mining*, pp. 181–190, 2021.

Yi He, Christian Schreckenberger, Heiner Stuckenschmidt, and Xindong Wu. Towards utilitarian online learning—a review of online algorithms in open feature space. In *Proceedings of the 32nd International Joint Conference on Artificial Intelligence*, pp. 6647–6655, 2023.

Matthew Hoffman, Francis Bach, and David Blei. Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 23, 2010.

Steven C. H. Hoi, Doyen Sahoo, Jing Lu, and Peilin Zhao. Online learning: A comprehensive survey. *Neurocomputing*, 459:249–289, 2021.

Bo-Jian Hou, Lijun Zhang, and Zhi-Hua Zhou. Learning with feature evolvable streams. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Chenping Hou and Zhi-Hua Zhou. One-pass learning with incremental and decremental features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2776–2792, 2018.

Niall Hurley and Scott Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

Heng Lian, John Scovi Atwood, Bo-Jian Hou, Jian Wu, and Yi He. Online deep learning from doubly-streaming data. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 3185–3194, 2022.

Heng Lian, Yu Huang, Xingquan Zhu, and Yi He. Utilitarian online learning from open-world soft sensing. In *2024 IEEE International Conference on Data Mining*, pp. 241–250, 2024.

Heng Lian, Yi He, Di Wu, Zhong Chen, Xingquan Zhu, and Xindong Wu. Online outlier detection in open feature spaces. *IEEE Transactions on Knowledge and Data Engineering*, 2025a.

Heng Lian, Yi He, Di Wu, Zhong Chen, Xingquan Zhu, and Xindong Wu. Online outlier detection in open feature spaces. *IEEE Transactions on Knowledge and Data Engineering*, 37(10):6091–6106, 2025b.

Haipeng Luo, Alekh Agarwal, Nicolò Cesa-Bianchi, and John Langford. Efficient second order online learning by sketching. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Kingsley Okoye and Samira Hosseini. Wilcoxon statistics in r: Signed-rank test and rank-sum test. In *R Programming: Statistical Data Analysis in Research*, pp. 279–303. Springer, 2024.

Lawrence K. Saul. An online passive-aggressive algorithm for difference-of-squares classification. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Christian Schreckenberger, Yi He, Stefan Lüdtke, Christian Bartelt, and Heiner Stuckenschmidt. Online random feature forests for learning in varying feature spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 4587–4595, 2023.

Linqi Song, Cem Tekin, and Mihaela Van Der Schaar. Online learning in large-scale contextual recommender systems. *IEEE Transactions on Services Computing*, 9(3):433–445, 2014.

Abhishek Venketeswaran, Nageswara Lalam, Jeffrey Wuenschell, Paul R Ohodnicki Jr, Mudabbir Badar, Kevin P Chen, Ping Lu, Yuhua Duan, Benjamin Chorpening, and Michael Buric. Recent advances in machine learning for fiber optic sensor applications. *Advanced Intelligent Systems*, 4 (1):2100067, 2022.

Yulin Wang, Yi He, Dianlong You, and Di Wu. Online learning for wild streaming data with delayed feedback. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2026.

Helmut Wasserbacher and Martin Spindler. Machine learning for financial forecasting, planning and analysis: recent developments and pitfalls. *Digital Finance*, 4(1):63–88, 2022.

Sarah Wassermann, Thibaut Cuvelier, Pavol Mulinka, and Pedro Casas. Adaptive and reinforcement learning approaches for online network monitoring and analysis. *IEEE Transactions on Network and Service Management*, 2021.

Brent Wenerstrom and Christophe Giraud-Carrier. Temporal data mining in dynamic feature spaces. In *IEEE International Conference on Data Mining*, pp. 1141–1145. IEEE, 2006.

Cun-Yuan Xing, Meng-Zhang Qian, Wu-Yang Chen, Wei Gao, and Zhi-Hua Zhou. One-pass feature evolvable learning with theoretical guarantees. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 68928–68952, 2025.

Peng Zhang, Hongpeng Yin, and Han Zhou. One-pass online learning from data streams with unpredictable feature evolution. *Pattern Recognition*, 171:112003, 2026.

Qin Zhang, Peng Zhang, Guodong Long, Wei Ding, Chengqi Zhang, and Xindong Wu. Online learning from trapezoidal data streams. *IEEE Transactions on Knowledge and Data Engineering*, 28(10):2709–2723, 2016.

Sheng-Da Zhuo, Jin-Jie Qiu, Chang-Dong Wang, and Shu-Qiang Huang. Online feature selection with varying feature spaces. *IEEE Transactions on Knowledge and Data Engineering*, 36(9): 4806–4820, 2024.

Shengda Zhuo, Di Wu, Yi He, Shuqiang Huang, and Xindong Wu. Online learning from mix-typed, drifted, and incomplete streaming features. *ACM Transactions on Knowledge Discovery from Data*, 19(8), 2025.

Terry Yue Zhuo, Dingmin Wang, Hantian Ding, Varun Kumar, and Zijian Wang. Cyber-zero: Training cybersecurity agents without runtime. In *International Conference on Learning Representations*, volume 2026, pp. 154054–154184, 2026.

Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning*, pp. 928–936, 2003.