Stabilizing Linear Passive-Aggressive Online Learning with Weighted Reservoir Sampling

Skyler Wu, Fred Lu, Edward Raff, James Holt

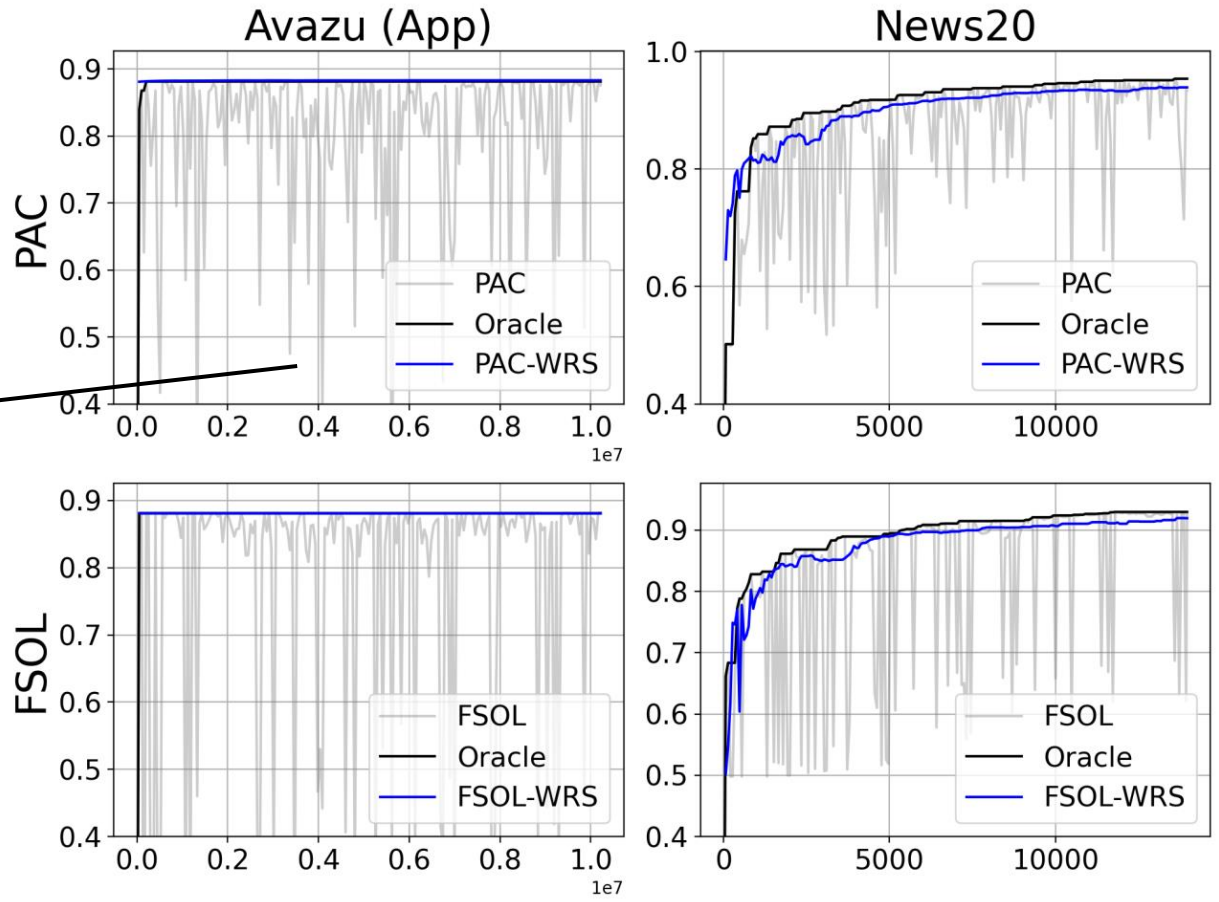
Problem + Deployment Setting (1)

- **Intended Deployment Setting:** Malware Classification.
  - High-volume + high-dimensional data.
  - Require “anytime” deployment.
  - Thus, cannot use complex + offline methods. Need online learning.
- **Proposed Toolkit:** Online Learning.
  - Observe one (**x**, y) datapoint at each timestep.
  - Update solution vector at each timestep (e.g. SGD).
  - Efficient + strong convergence guarantees.

BUT ONLINE LEARNING METHODS CAN BE VERY SENSITIVE TO OUTLIERS + OVERCORRECT …

Problem + Deployment Setting (2)

40%+ decrease in test accuracy within a single timestep!



Test Accuracy Over Time of WRS-Augmented Training (WAT) vs. Base Models and Oracle.

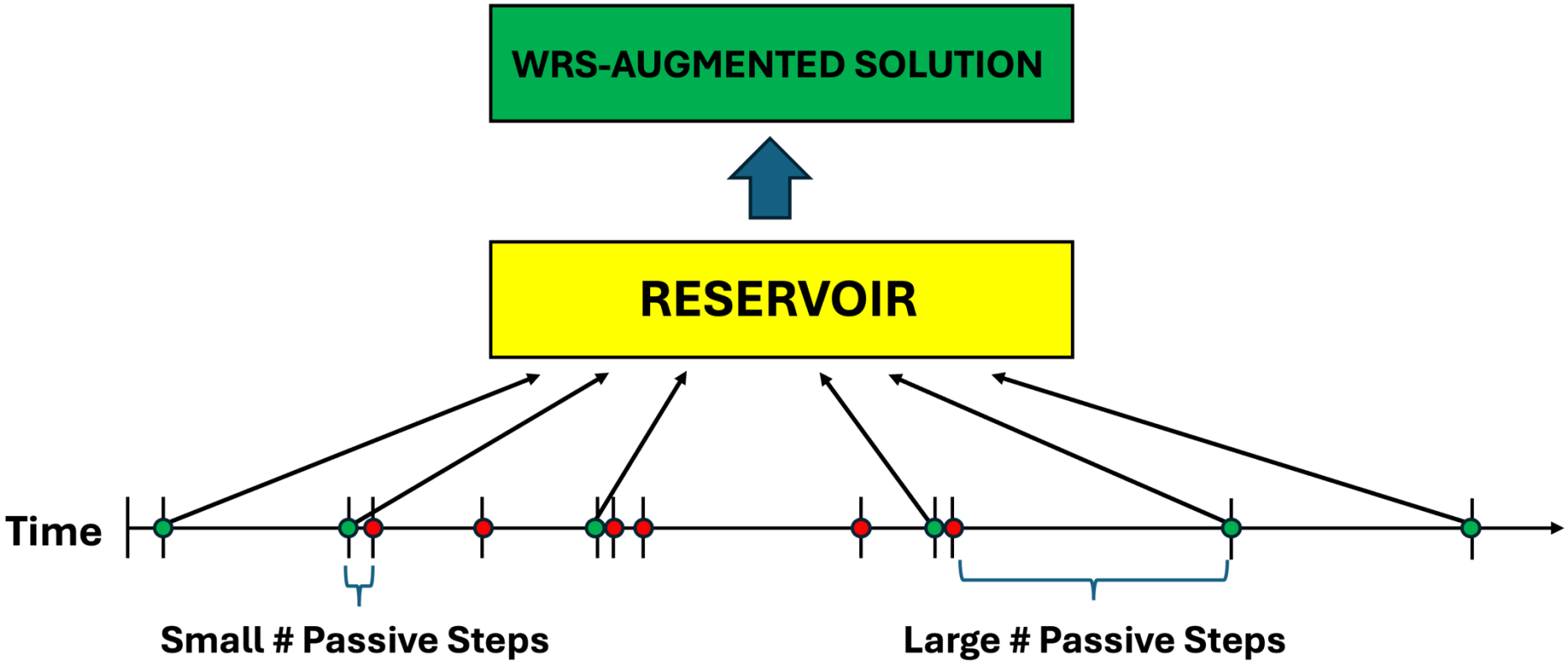
Desiderata

- **Need to stabilize online learning method's test accuracy:**
  - … without multiple passes over the data.
  - … without access to a hold-out / evaluation set.
  - … without excessive memory + storage requirements.
- **"Passive-Aggressive" (PA) online linear binary classification.**
  - Subclass of online learning methods [Crammer 2006, Zhao 2020].
  - Only update solution vector when make mistake, **else no update.**
  - Extra computationally-efficient.
  - Serve as our "base" models.

Method: WRS-Augmented Training (1)

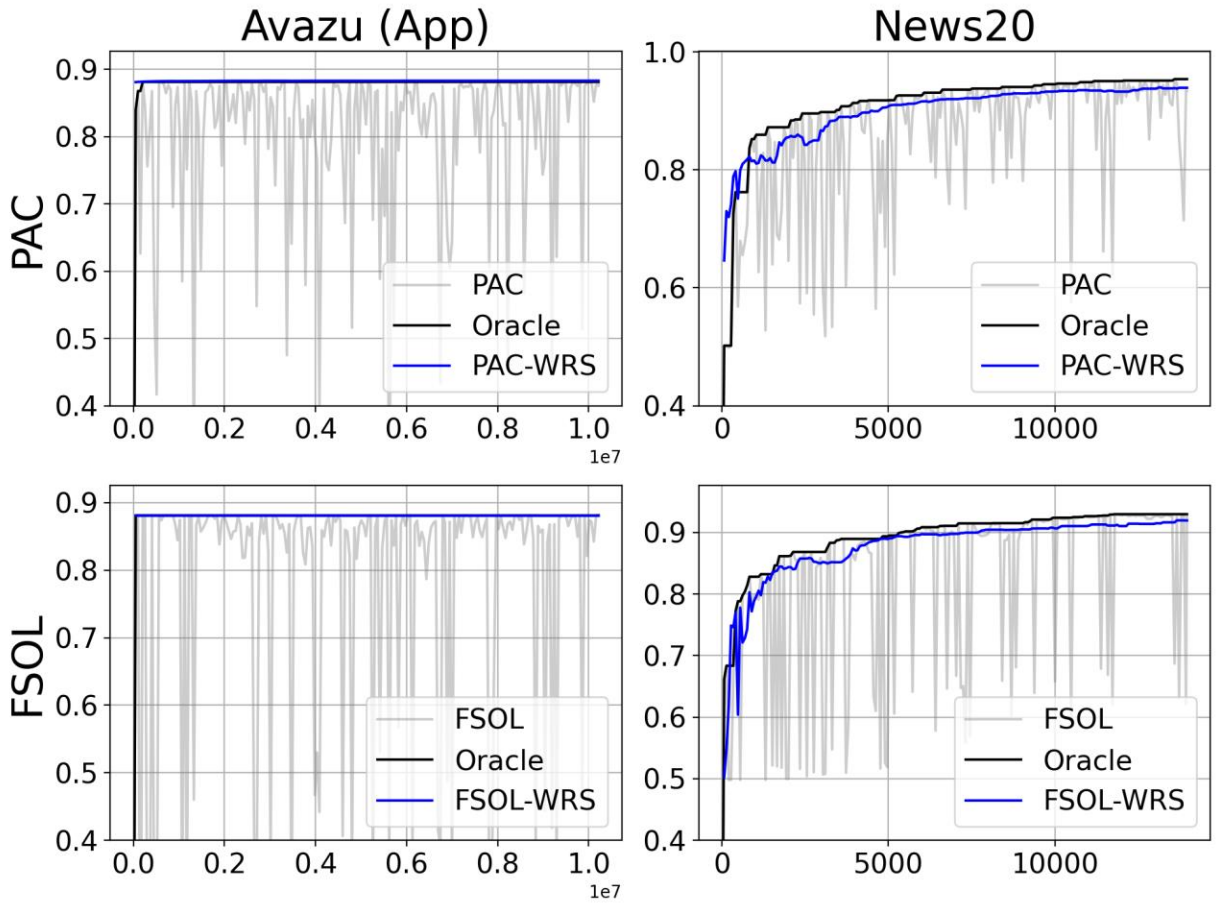
- **Key Insight:** for PA algorithms, the subsequent # of passive steps of a solution vector is a strong proxy for solution quality.
- **Q: How to capture the "best" solution vectors?**
- **A: Weighted Reservoir Sampling (WRS)** [Efraimidis 2006]
  - Given $V = \{v_1, \dots, v_T\}$ with nonnegative weights $w_1, \dots, w_T$, want to collect a size-K weighted random sample from V in **one pass**, even if **population size $|V| = T$ is unknown!**
  - While making our one pass through V, maintain + update a "reservoir" with maximum capacity K. **Members of reservoir are potential candidates for final size-K sample.**

Method: WRS-Augmented Training (2)



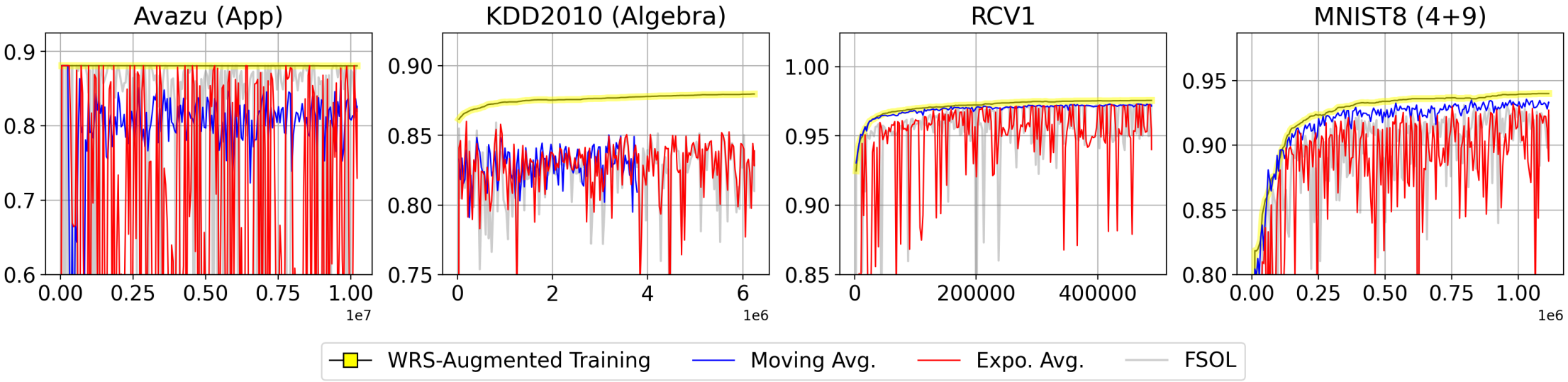
WRS-Augmented Training. Intervals: # of passive steps between aggressive updates. Green/red: accepted/rejected. X-axis represents progression over timestep.

WRS-Augmented Training = Highly Effective Stabilization!



Test Accuracy Over Time of WRS-Augmented Training (WAT) vs. Base Models and Oracle.

WRS-Augmented Training = Stabler than Alternatives.



Test Accuracy Over Time of WAT vs. Moving Average + Exponential Average with FSOL base.

WRS-Augmented Training = Faster than Alternatives.

	PAC		FSOL	
Dataset	**Moving Avg.**	**Expo. Avg.**	**Moving Avg.**	**Expo. Avg.**
Newsgroups (Binary, CS)	3.587	1.042	1.376	0.543
Avazu (App)	8.357	2.507	8.078	2.544
Avazu (Site)	18.704	3.207	11.612	3.683
Criteo	14.267	2.876	7.009	2.381
News20	4.288	1.356	3.376	1.153
URL	10.050	3.945	9.124	4.058
KDD2010 (Algebra)	6.579	2.402	6.778	2.301

Avg. Compute Times Per Iteration (Seconds) of Moving/Expo. Average Relative to WAT.

For example, on URL with PAC, Moving Average was 10.050x slower per iteration than WAT.

Thank You

Paper: https://arxiv.org/abs/2410.23601

Code: https://github.com/FutureComputing4AI/Weighted-Reservoir-Sampling-Augmented-Training

Poster Number: 95981