Strategic Usage in a Multi-Learner Setting

Eliot Seo Shekhtman & Sarah Dean
Cornell University Bowers CIS

Contact Information:
Phone: +1 (518) 810 1582
Email: ess239@cornell.edu

Motivating Example

Retailers (**users**) choose between online marketplace platforms (**services**) for posting their listings. While most retailers are **legitimate**, some are **scammers** who post scam listings to deceive customers. Platforms want legitimate retailers, but not scammers. Platforms **learn algorithms to filter** scam retailers; however, this forces retailers to **adapt strategically** to stay out of the filter.

Related Work

Strategic Classification

- Hardt et al. (2016): **single-service setting**
- Users engaged in strategic **feature manipulation** to adapt to the service (retailers make their listings more believable).
- As the classifiers shift, even certain legitimate retailers would start to be filtered out, forcing them to begin acting strategically as well just to maintain their ability to post listings.

Performative Prediction

- Hardt et al. (2022): feature manipulation could often be prohibitively costly.
- In **multi-service settings**, scam retailers are more likely to **switch services** instead!

Our Contributions

- We **formalize** the game of **strategic usage**, where users only use advantageous services.
- Services only observe the users using them; marketplaces only observe their own listings.
- **Naïve retraining** allows users to game the system by **oscillating** between services.
- When services incorporate **past observations** they reach a **convergent state** with desirable conditions in a **finite time**.
- This status quo would be the starting point from which a feature manipulation game would be incentivized to begin.

At every timestep:
- Users choose between services
- Services observe their user distribution
- Services release models to minimize loss and classify users
- Users get utility from classifications

Formalized Setting

- **Binary classification**, modeling $n \in \mathbb{N}_+$ **users** with d features $x_i \in \mathcal{X}$ and label $y_i \in \{+1, -1\}$.
- Features are used for classifying each user: descriptions, reviews, number of listings, etc.
- They use $m \in \mathbb{N}_+$**services** which each put out **classifiers** $h_j^t : \mathcal{X} \to \{+1, -1\}, h \in \mathcal{H}$ at every timestep t to classify the users.

User Game

- Services give utility to users proportional h, as controlled by **utility function** $u : \mathcal{X} \times \mathcal{H} \to \mathbb{R}$.
- Utility shares sign with h, meaning that users receive utility from positive classifications.
- Users then assign **usage** A to services that give the most utility, optimizing:

$$\max \sum_{j=1}^{m} A_{ij} u(x_i, h_j^t) - \frac{1}{q}\left(\sum_{j=1}^{m} A_{ij}\right)^q \quad (1)$$

- Usage cost can be interpreted as the effort to join a platform, or to create listings.

Service Game

- Usages are observed to update the service's understanding of the user distribution, controlled by **memory** parameter p.

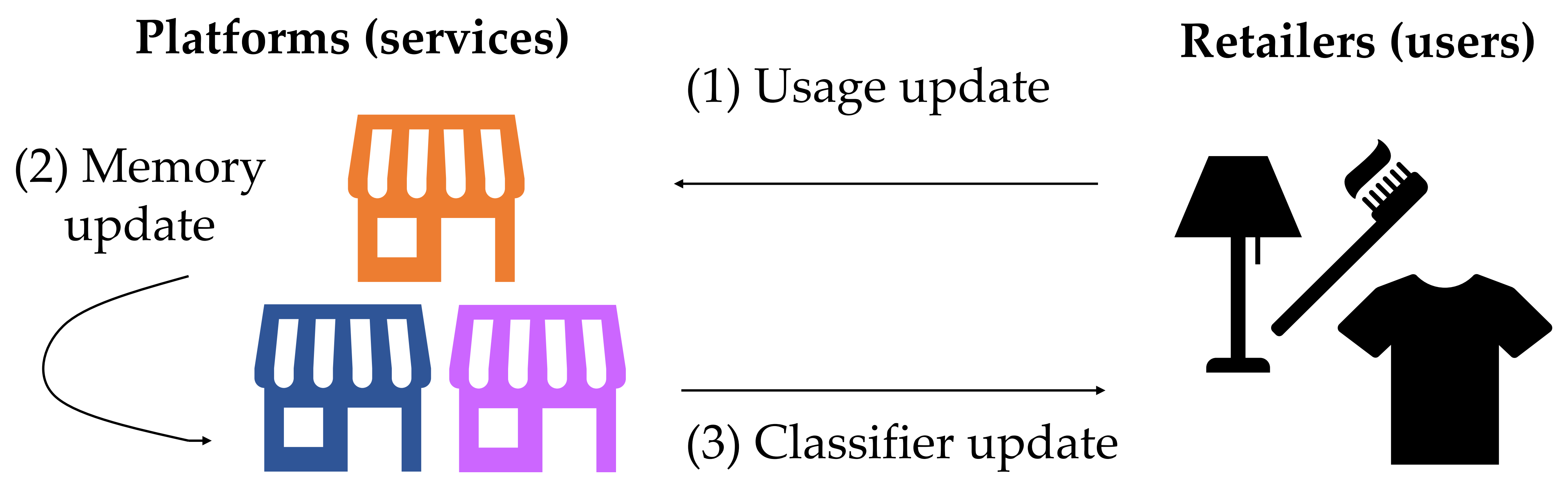
$$M^t = \frac{A^t}{1+p} + \frac{pM^{t-1}}{1+p} \quad (2)$$

- Services then minimize non-negative **loss** $\ell : \mathcal{H} \times \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ over this distribution.
- We assume that increasing utility to positive users will decrease loss, while increasing utility to negative users incurs loss.
- Service objective is a weighted loss:

$$\min \sum_{i=1}^{n} \frac{M_{ij}^t}{\sum_{k=1}^{n} M_{kj}^t} \ell(h_j, x_i, y_i) \quad (3)$$

Full Interaction Dynamics

- We assume **joint updates** due to the independence of usage updates between users and classifier updates between services.
- Users can tiebreak optimal usages however they choose; however, we constrain service tiebreaking to abiding by a **sticky property**.
- If a classifier h_j^t is optimal at timestep t and $t + 1$, the service will re-use the classifier from the previous timestep.



Theoretical Results

Definition 2. A state (H, A) is **zero-loss** if all services j satisfy:

1. $A_{ij}\ell(h_j, x_i, y_i) = 0$ for all $i \in \{1, \ldots, n\}$
2. $u(x_i, h_j) \leq 0$ for all i with $y_i = -1$

Proposition 1. In the memoryless $p = 0$ setting, there exist settings in which the state (H, A) never converges.

Proposition 3. If state (H^t, A^t) is zero-loss, then states (H^τ, A^τ) are zero-loss for all times $\tau \geq t$.

Lemma 5. For any timestep t if there exists no values $M_{ij}^{t-1} = 0$ such that $A_{ij}^t > 0$, then (H^t, A^t) is zero-loss.

Theorem 6. Given nonzero memory $p > 0$, there is a finite time $t \in \mathbb{N}_+$ after which for all $\tau > t$, (H^τ, A^τ) is zero-loss.

Banknote Experiment

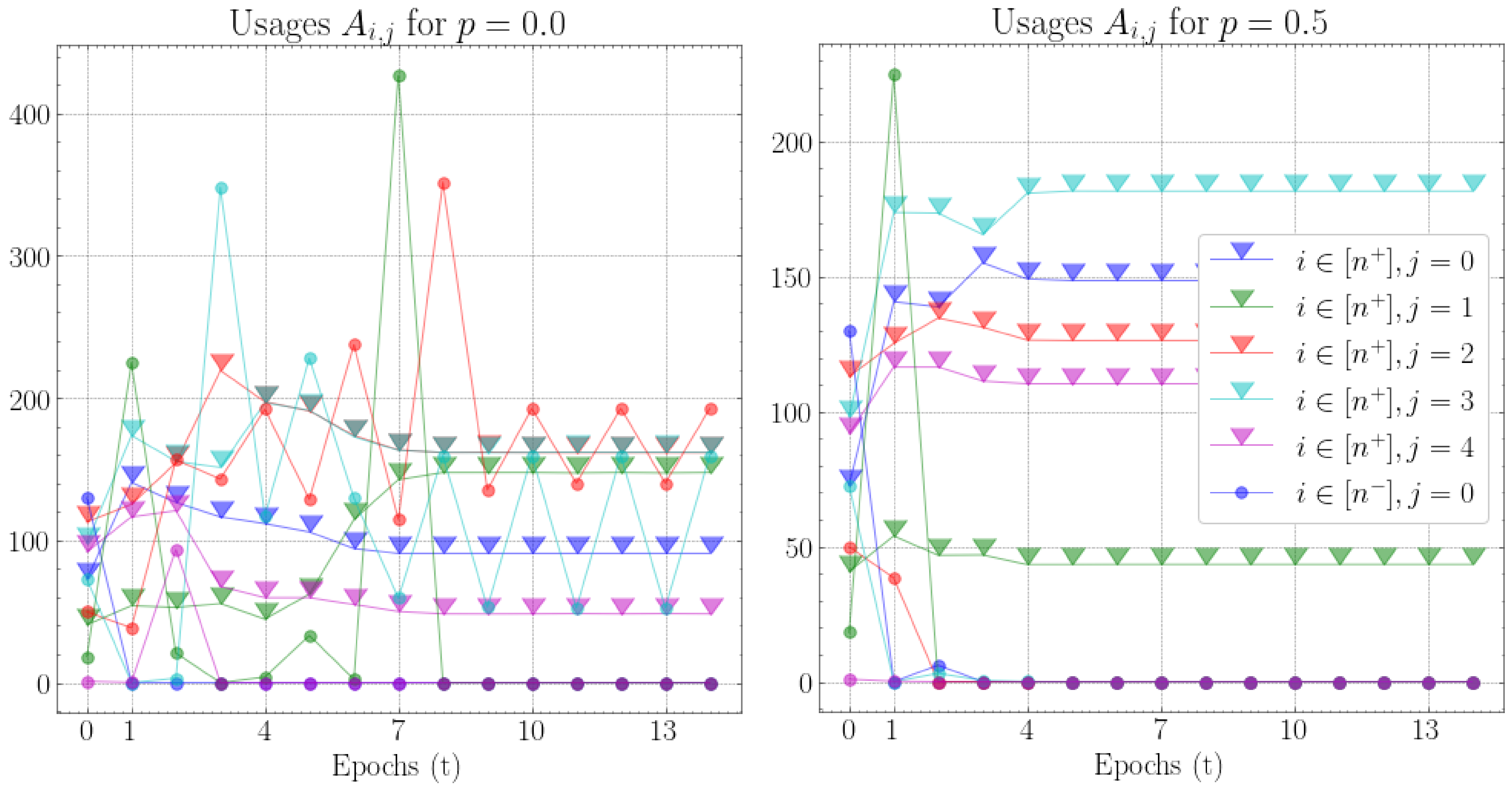
We use the Banknote Authentication dataset (Lohweg, 2013) to experimentally verify our theoretical results. **Legal** banknotes serve as positive users and **forgeries** as negative users, with banks being modeled as services trying to avoid forgeries. In the memoryless setting, oscillation can be observed and negative users never leave the system; however, when $p > 0$ we observe a convergent state after the fourth epoch.

References

M. Hardt, N. Megiddo, C. Papadimitriou, and M. Wootters. Strategic classification. In Proceedings of the 2016 ACM conference on innovations in theoretical computer science, pages 111–122, 2016.

M. Hardt, M. Jagadeesan, and C. Mendler-Dünner. Performative Power. Advances in Neural Information Processing Systems, 36, 2022.

V. Lohweg. Banknote Authentication. UCI Machine Learning Repository, 2013. DOI: https://doi.org/10.24432/C55P57.


Usages $A_{i,j}$ for $p = 0.0$
Epochs (t)
Usages $A_{i,j}$ for $p = 0.5$
Epochs (t)
$i \in [n^+], j = 0$
$i \in [n^+], j = 1$
$i \in [n^+], j = 2$
$i \in [n^+], j = 3$
$i \in [n^+], j = 4$
$i \in [n^-], j = 0$