Heterogeneous Model Reuse via Optimizing Multiparty Multiclass Margin

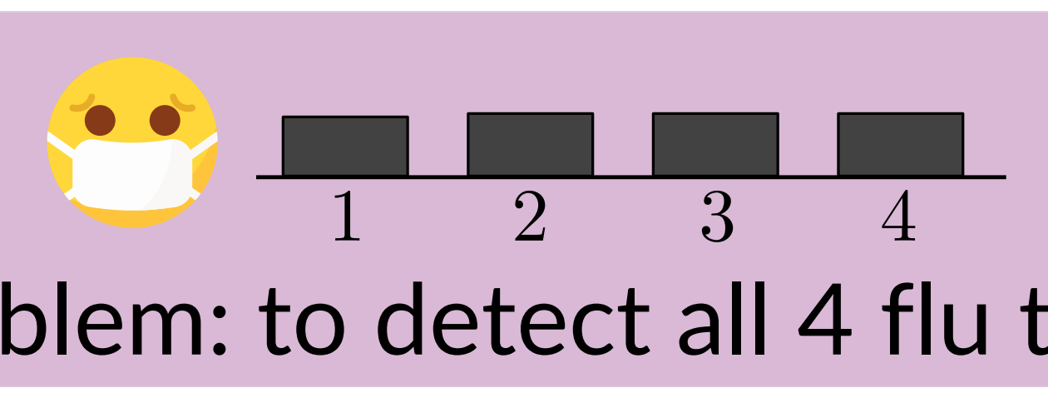
Xi-Zhu Wu[1], Song Liu[2 3], Zhi-Hua Zhou[1]

[1]LAMDA Group, Nanjing University, China
{wuxz,zhouzh}@lamda.nju.edu.cn

[2]University of Bristol, [3]The Alan Turing Institute, United Kingdom
song.liu@bristol.ac.uk

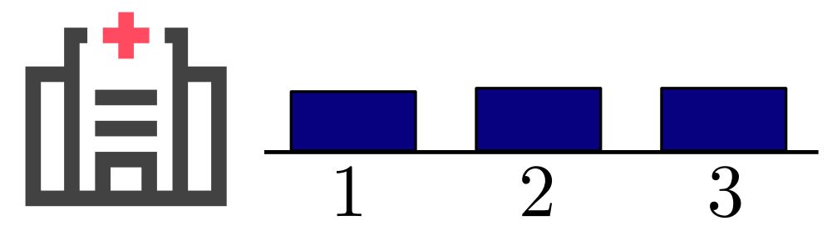
Problem setting

- Problem: Multiparty multiclass classification
- Example: Flu detection

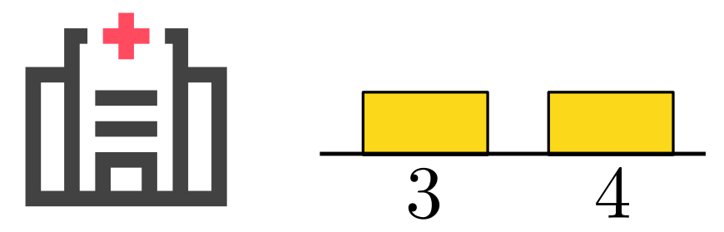


Global problem: to detect all 4 flu types in the US

But, the types of flu diverse geographically, the distribution of patients records collected by a hospital in California is different from Florida. Good local models are built:

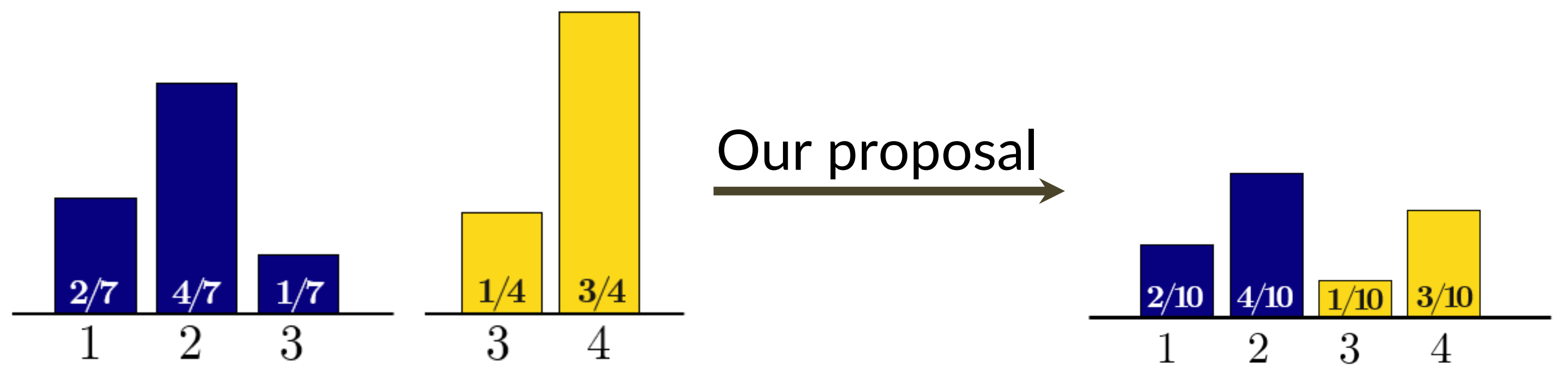


Local model in California detects {1,2,3} types



Local model in Florida detects {3,4} types

The patients' records are confidential. Can we smartly reuse the local models to learn the global problem, instead of building a model on merged local datasets?



Multiple models trained separately with different subsets of the label space

One model works well on the full label space

- Notations

Parties each with local datasets but different label spaces:

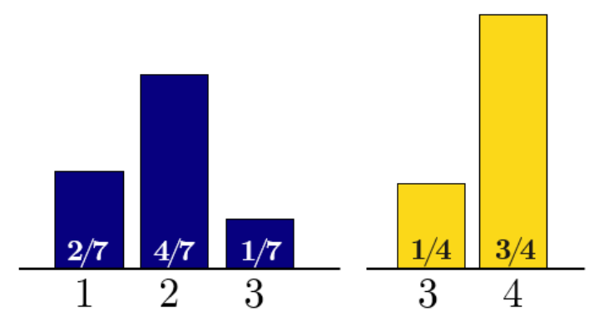
$$S_i = (X_i, Y_i) = \{(x,y) \in \mathcal{X} \times \mathcal{Y}_i\} \subseteq S$$

$$\mathcal{Y}_i \subseteq \mathcal{Y} = \{1, 2, \cdots, k\}$$

Local predictor trained by local algorithm on local dataset:

$$h_i : \mathcal{X} \times \mathcal{Y}_i \to \mathbb{R} \qquad h_i = \mathcal{A}_i(S_i)$$

Example $h_1(x,1)=2/7$ $h_1(x,2)=4/7$ $h_1(x,3)=1/7$



$h_2(x,3)=1/4$ $h_2(x,4)=3/4$

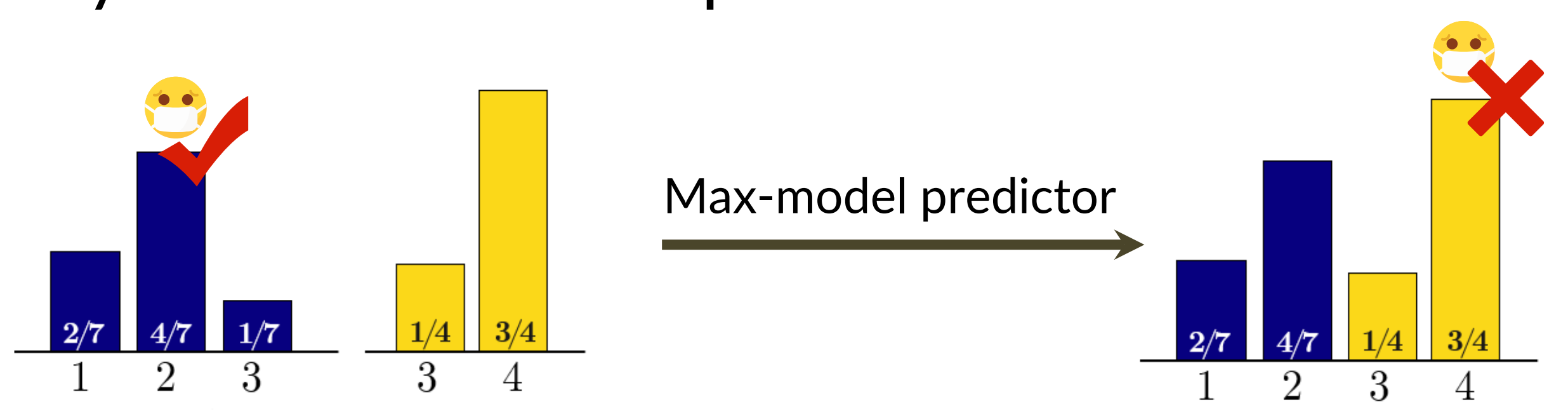
Contact: wuxz@lamda.nju.edu.cn, wuxz.gm@gmail.com
Code: https://github.com/YuriWu/HMR

Behavior of an ensemble of local models

- The intuitive ensemble of local models is to use max-model predictor: Given a set of multi-class predictors $H=\{h_1, \cdots, h_n\}$, the max-model predictor h_H is defined as:

$$h_H(x,y) = \max_{y\in\mathcal{Y}_i, h_i\in H} h_i(x,y)$$

- However, max-model predictor may fail even if each local model is perfect (see Claim 1 in our paper for the formal statement).
- Intuition: another local model which is unaware of the true class may mislead the final prediction.



true label: 2

wrong predicted label: 4

Contribution

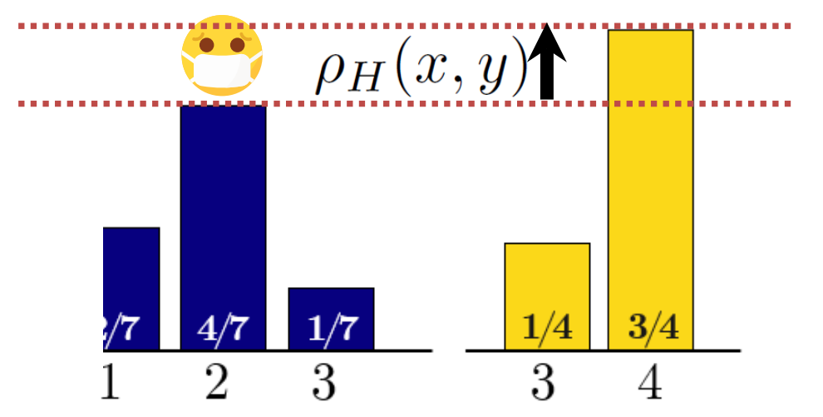
Q: How to measure the global behavior of multiple models?
A: Multiparty multiclass margin. (MPMC-margin)

Q: How to optimize the global behavior?
A: The HMR method, which maximizes MPMC-margin.
by modifying local models,
without merging local datasets.

MPMC-margin

- The multiparty multiclass margin (MPMC-margin) on the local predictors set $H=\{h_1, \cdots, h_n\}$ at a labeled example (x,y) is defined as:

$$\rho_H(x,y) = \max_i h_i(x,y) - \max_{j,y'} h_j(x,y'),$$
$$\text{where } y \in \mathcal{Y}_i, y' \in \mathcal{Y}_j \backslash \{y\}.$$



- Non-positive MPMC-margin causes wrong prediction, so we want to maximize it.

Heterogeneous model reuse method

```
Algorithm 1 HMR
input:
    Parties $1, 2, \cdots, n$, each owns a local dataset $S_i$ and a local model $h_i$. Example communication budget $N$.
output:
    Calibrated local models $h_1, \cdots, h_n$.
procedure:
  Each party broadcasts its local model to others.
  Inner iteration counter $T = 0$
  while $T < N$ do
      Sample a party $i$ according to $|S_i| / \sum_{i=1}^n |S_i|$.
      Party $i$ randomly selects an example $(x,y) \in S_i$.
      Party $i$ computes MPMC-margin $\rho_H(x,y)$ and records the party $i^+, i^-$ and maximum incorrect class $y^-$ as in (8).
      if $\rho_H(x,y) \le 0$ then
          Party $i$ sends $(x, y, y^-)$ to $i^+$ and $i^-$.
          Party $i^+$ calibrates $h_{i^+}$ with $(x, y, y^-)$.
          Party $i^-$ calibrates $h_{i^-}$ with $(x, y, y^-)$.
          Party $i^+$ and $i^-$ broadcast their updated model.
          if $i^+ \ne i$ or $i^- \ne i$ then
              $T = T + 1$.
          end if
      end if
  end while
```

An iterative method exchanges T examples and maximizes MPMC-margin on "unobserved" merged global dataset.

$$\rho_H(x,y) = h_{i^+}(x,y) - h_{i^-}(x,y^-)$$
$$i^+ = \arg\max_i h_i(x,y), \text{ where } y \in \mathcal{Y}_i, \tag{8}$$
$$(i^-, y^-) = \arg\max_{j,y'} h_j(x,y'), \text{ where } y' \in \mathcal{Y}_j \backslash \{y\}.$$

check MPMC-margin

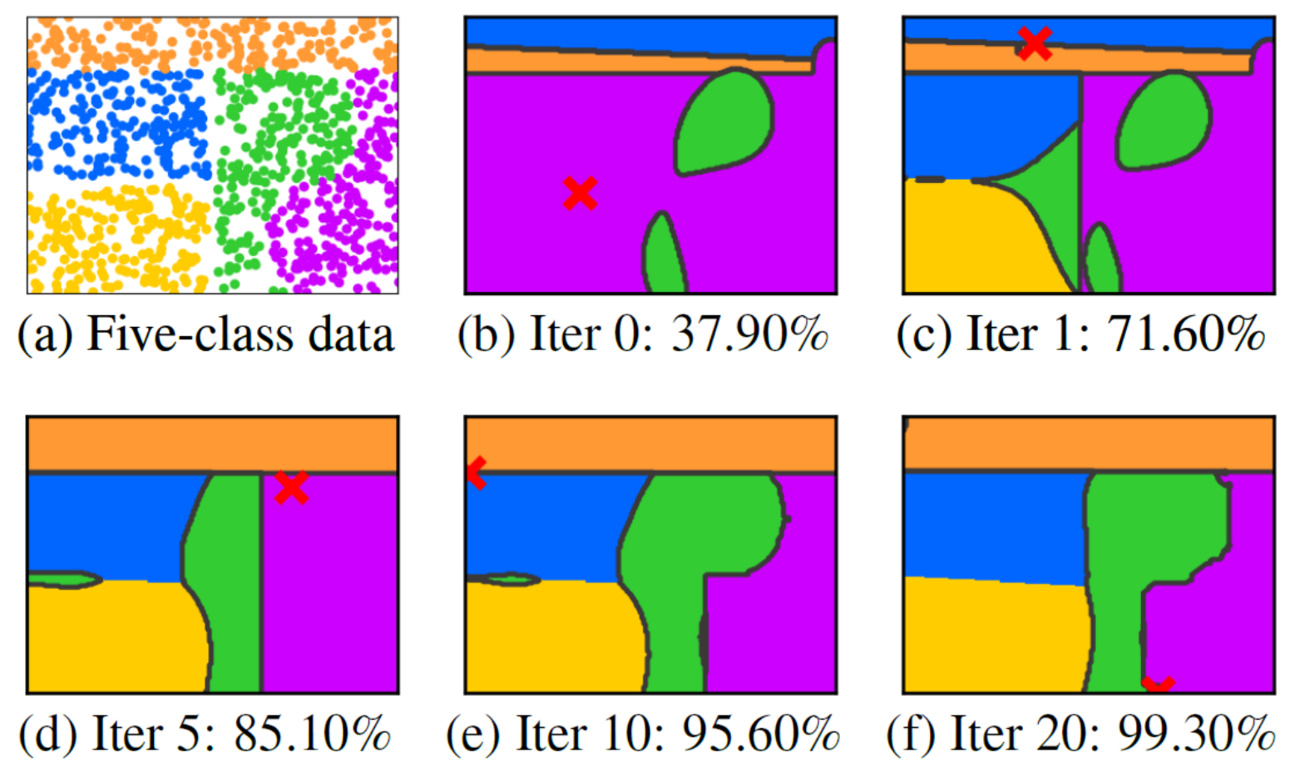
send out one example if non-positive

update local models

Privacy issue: If T is small enough, most of the private local data will be protected. Less than 1% of global data are shared in experiments.

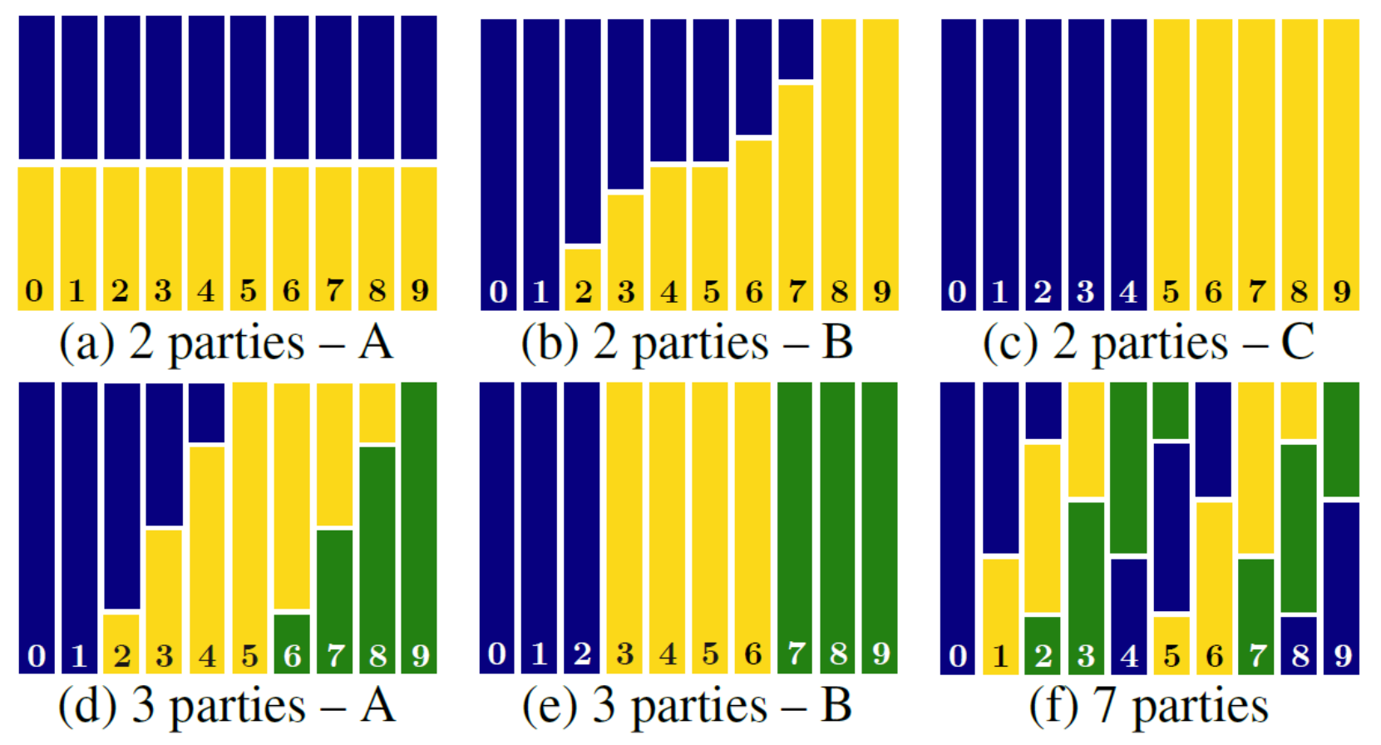
Experiments

- Toy example on LR/SVM/GBDT
  - Heterogeneous learning models
    - LR: green, yellow
    - SVM: green, magenta
    - GBDT: magenta, orange
  - Exchanged 20 examples
  - Nearly perfect performance

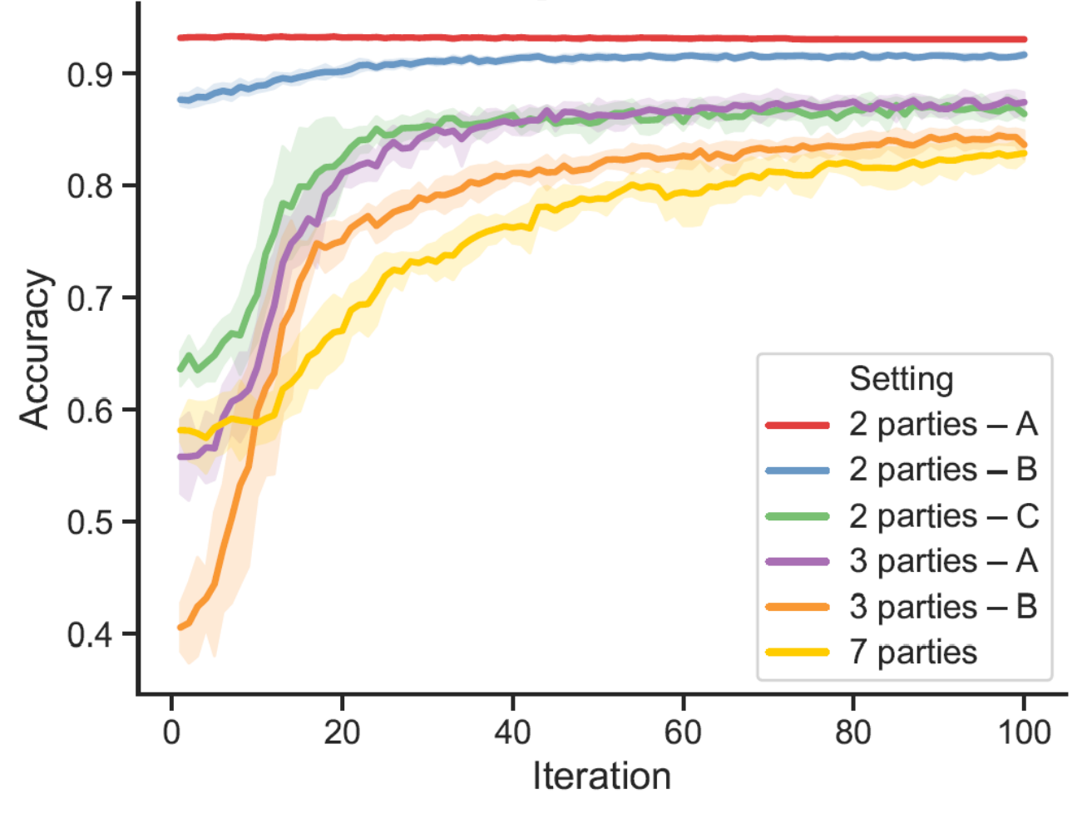


(a) Five-class data (b) Iter 0: 37.90% (c) Iter 1: 71.60%
(d) Iter 5: 85.10% (e) Iter 10: 95.60% (f) Iter 20: 99.30%

- Benchmarking on fashion-MNIST on various data partitions



(a) 2 parties – A (b) 2 parties – B (c) 2 parties – C
(d) 3 parties – A (e) 3 parties – B (f) 7 parties



- Multi-lingual handwriting recognition
  - 6 different structured CNNs trained locally on hiragana, katakana, kanji, devanagari, hangul and English letters
  - 1600+ classes, 94.32% global accuracy
  - Only exchanged 300 out of 420k examples