Communication-Aware Collaborative Learning

Avrim Blum [1] Shelby Heinecke [2] Lev Reyzin [3]

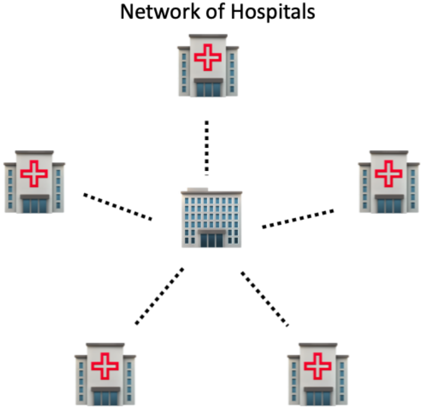
[1]Toyota Technological Institute at Chicago

[2]Salesforce Research

[3]University of Illinois at Chicago

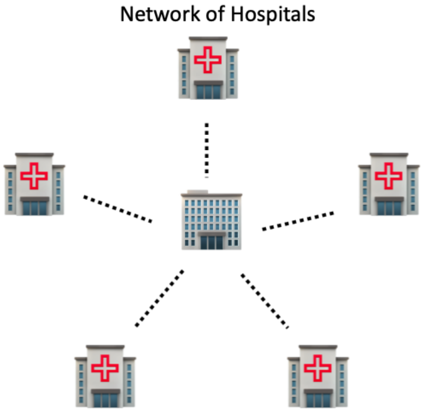
January 7, 2021

Motivating Example



- Each hospital serves a different neighborhood
  - Different distributions on X
- Need highly accurate models for **every** hospital
  - Hospitals can collaborate
  - Harder than traditional distributed learning

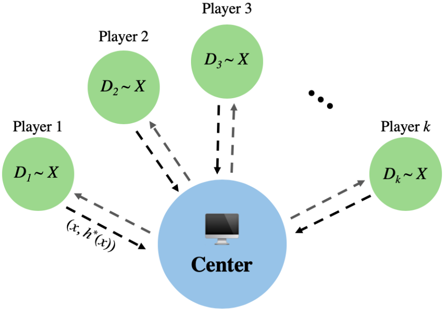
Motivating Example



- **Issue #1**: Communicating data is costly
- **Issue #2**: Data can be noisy
  - clerical errors
  - misdiagnoses
- Can the hospitals collaboratively train accurate classifiers **efficiently**?
  - sample complexity
  - communication complexity

Learning Model: Collaborative PAC Learning

- Formalized as a PAC framework in (Blum et al., 2017).
- **Personalized learning**: learn a classifier for each player that has generalization error $< \epsilon$, with probability $1 - \delta$
- **Centralized learning**: learn one classifier that works for each player



Our Work

Challenge #1: Communication cost

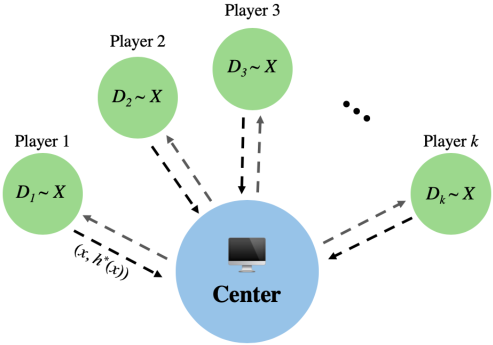
- Communication efficient personalized learning
- Key ingredient: distributed boosting

Challenge #2: Noisy data

- Communication efficient personalized learning with noise
- Key ingredient: distributed agnostic boosting

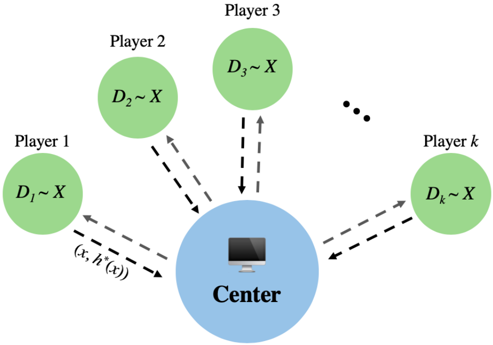
Assumptions

- k players
- d = VC-dim of H
- All players know ϵ, δ, H
- Fix δ be a constant



Assumptions

- Realizable PAC setting: $h^* \in H$
- Broadcast Model: all players can observe data transmitted to center



Baseline

- No previous work on communication efficiency of collaborative PAC learning
  - Compare sample and communication complexities to `Baseline` and `Personalized Learning`
- `Baseline`: Each player learns their own classifier individually
  - Each player draws $\tilde{O}(\frac{d}{\epsilon})$ samples locally
  - Each player learns their own classifier (standard PAC learning)

`Baseline`

Sample Complexity: $\tilde{O}(k\frac{d}{\epsilon})$
Communication Complexity: $\tilde{O}(1)$

Personalized Learning (Blum et al., 2017)

For $O(\log(k))$ rounds:

1. Draw samples, S, from uniform mixture of remaining players
2. Learn consistent hypothesis h on S
3. For each remaining player, compute empirical error of ERM on their distribution:
    - Low error $\implies$ assign ERM to player
4. Repeat with players without assigned classifiers

Personalized Learning

Sample Complexity: $\tilde{O}(\log(k)\frac{d}{\epsilon})$
Communication Complexity: $\tilde{O}(\log(k)\frac{d}{\epsilon})$

Summary

	Sample Complexity	Samples Communicated
`Baseline`	$\tilde{O}(k\frac{d}{\epsilon})$	$\tilde{O}(1)$
`Personalized Learning`	$\tilde{O}(\log(k)\frac{d}{\epsilon})$	$\tilde{O}(\log(k)\frac{d}{\epsilon})$

- `Personalized Learning` logarithmic in k, $k >> 0$
- Can we achieve optimal sample complexity and reduced communication complexity?
  - Highly accurate classifiers, $\epsilon << 0$
  - Can we improve ϵ dependence?

Distributed Boosting (Balcan et al., 2012)

For $\tilde{O}(\log(\frac{1}{\epsilon}))$ rounds:

1. Center gets points from players
2. Center trains weak learner and sends to players
3. Players amplify or reduce weights on their points based on performance of weak learner

Distributed Boosting

Sample Complexity: $\tilde{O}(\frac{d}{\epsilon})$
Communication Complexity: $\tilde{O}\left(d \log\left(\frac{1}{\epsilon}\right)\right)$

PL + Boosting

For $O(\log(k))$ rounds:

1. Draw samples, S, from uniform mixture of remaining players
2. **Use** `Distributed Boosting` **to learn consistent h on S**
3. For each remaining player, compute empirical error of h on their distribution:
   - Low error $\implies$ assign h to player
4. Repeat with players without assigned classifiers

PL + Boosting

Sample Complexity: $\tilde{O}(\log(k)\frac{d}{\epsilon})$

Communication Complexity: $\tilde{O}(\log(k)d\log(\frac{1}{\epsilon}))$

Summary

	Sample Complexity	Samples Communicated
Baseline	$\tilde{O}(k\frac{d}{\epsilon})$	$\tilde{O}(1)$
PL	$\tilde{O}(\log(k)\frac{d}{\epsilon})$	$\tilde{O}(\log(k)\frac{d}{\epsilon})$
PL + Boosting	$\tilde{O}(\log(k)\frac{d}{\epsilon})$	$\tilde{O}(\log(k)d\log(\frac{1}{\epsilon}))$

✓ Achieves optimal sample complexity

✓ Improved communication cost – **logarithmic** in $\frac{1}{\epsilon}$

Our Work

Challenge #1: Communication cost

- Communication efficient personalized learning
- Key ingredient: distributed boosting

Challenge #2: Noisy data

- Communication efficient personalized learning with noise
- Key ingredient: distributed agnostic boosting

Collaborative PAC Learning with Noise

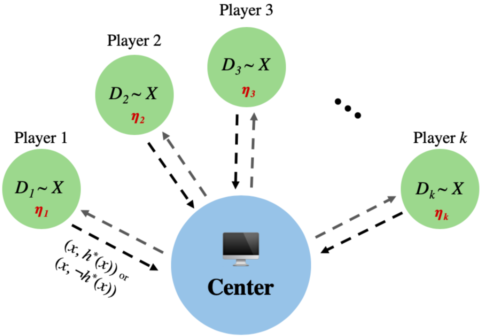
- Previous work: adversarial noise model (Qiao 2018)
- Our work: classification noise, not previously analyzed
  - Personalized and centralized learning are possible
- To achieve communication efficiency and robustness to noise:
  1. Adapt `Personalized Learning` to handle classification noise
  2. Use `Distributed Agnostic Boosting` in noise-robust personalized learning to improve communication cost

Assumptions and Key Classic Result

- Classification noise: each player has error rate $\eta_i < \frac{1}{2}$
- Center knows error rates
- **Theorem (Angluin and Laird, 1988)**: PAC learning in the presence of classification noise is achieved by learning an ERM given at least

$$O\left(\frac{d\log(\frac{1}{\delta})}{\epsilon(1-2\eta_i)^2}\right)$$

samples.



Noisy Baseline

- Analogous to the noiseless baseline - sample cost by Angluin-Laird theorem
- `Noisy Baseline`: Each player learns their own classifier individually
  - Each player draws $O\left(\frac{d}{\epsilon(1-2\eta_i)^2}\right)$ samples locally
  - Each player learns their own classifier (ERM)

Noisy Baseline

Sample Complexity: $\tilde{O}\left(k\frac{d}{\epsilon(1-2\eta_{MAX})^2}\right)$

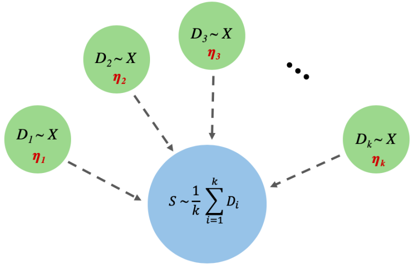
Communication Complexity: $\tilde{O}(1)$

Personalized Learning with Classification Noise

For $O(\log(k))$ rounds:

Step 1: Draw samples from uniform mixture

- Draw $O\left(\frac{d}{\epsilon(1-2\eta_{\mathrm{MAX}})^2} \ln\left(\frac{1}{\delta}\right)\right)$ samples so that ERM has error $< \frac{\epsilon}{4}$ on the mixture



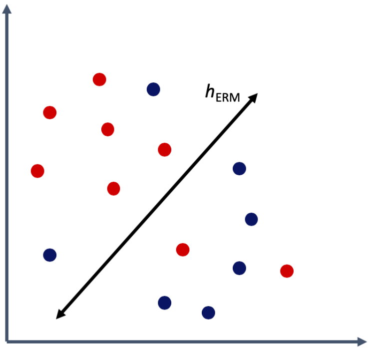
Personalized Learning with Classification Noise

Step 2: Learn ERM

- Learn ERM hypothesis
- By noisy-PAC learning,

$$err_{\frac{1}{k}\sum_{i=1}^{k} D_i}(h_{\mathrm{ERM}}) < \frac{\epsilon}{4}$$

- By Markov's inequality, at least half of players have error $< \frac{\epsilon}{2}$



Personalized Learning with Classification Noise

Step 3: Test ERM on players

This step should identify players for which h_{ERM} performs well.

Problem

Want to generalize on underlying clean player distributions but only have access to noisy data from players.

Step 3: Test ERM on players

Noisy to Clean Distribution (Angluin and Laird 1988)

$$err_{D,\eta}(h) = \eta + err_D(h)(1 - 2\eta)$$

By above and multiplicative Chernoff bounds, center draws

$$T = O\left(\frac{d}{\epsilon(1 - 2\eta_i)} \ln\left(\frac{1}{\delta}\right)\right)$$

samples from each player and computes the empirical error of h_{ERM}.

Noisy PL

Sample Complexity: $\tilde{O}\left(\log(k)\frac{d}{\epsilon(1-2\eta_{MAX})^2}\right)$

Communication Complexity: $\tilde{O}\left(\log(k)\frac{d}{\epsilon(1-2\eta_{MAX})^2}\right)$

Summary

	Sample Complexity	Samples Communicated
`Noisy Baseline`	$\tilde{O}\left(k\frac{d}{\epsilon(1-2\eta_{MAX})^2}\right)$	$\tilde{O}(1)$
`Noisy PL`	$\tilde{O}(\log(k)\frac{d}{\epsilon(1-2\eta_{MAX})^2})$	$\tilde{O}(\log(k)\frac{d}{\epsilon(1-2\eta_{MAX})^2})$

✓ Achieves improved sample complexity

Is it possible to achieve improved sample complexity <u>and</u> reduced communication complexity? Yes.

Distributed Agnostic Boosting (Chen, Balcan, and Chau 2016)

- Distributed implementation of agnostic boosting
- Classification noise is a special case of agnostic learning

Distributed Agnostic Boosting

Restrict to the classification noise setting.

Sample Complexity: $\tilde{O}\left(\frac{d}{\epsilon(1-2\eta_{MAX})^2}\right)$

Communication Complexity: $\tilde{O}\left(d \log\left(\frac{1}{\epsilon(1-2\eta_{MAX})}\right)\right)$

Communication Efficient Noisy PL

For $O(\log(k))$ rounds:

1. Draw samples, S, from uniform mixture of remaining players
2. **Use** `Distributed Agnostic Boosting` **to learn h on S**
3. For each remaining player, compute empirical error of h on their distribution:
   - Low error $\implies$ assign h to player (using classification noise modifications)
4. Repeat with players without assigned classifiers

`Noisy PL with Boosting`

Sample Complexity: $\tilde{O}(\log(k)\frac{d}{\epsilon(1-2\eta_{MAX})^2})$

Communication Complexity: $\tilde{O}(\log(k)d\log(\frac{1}{\epsilon(1-2\eta_{MAX})}))$

Summary

	Sample Complexity	Samples Communicated
Noisy Baseline	$\tilde{O}\left(k\frac{d}{\epsilon(1-2\eta_{MAX})^2}\right)$	$\tilde{O}(1)$
Noisy PL	$\tilde{O}(\log(k)\frac{d}{\epsilon(1-2\eta_{MAX})^2})$	$\tilde{O}(\log(k)\frac{d}{\epsilon(1-2\eta_{MAX})^2})$
Noisy PL + Boosting	$\tilde{O}(\log(k)\frac{d}{\epsilon(1-2\eta_{MAX})^2})$	$\tilde{O}(\log(k)d\log(\frac{1}{\epsilon(1-2\eta_{MAX})}))$

- ✓ Achieves improved sample complexity
- ✓ Improved communication cost – **logarithmic** in $\frac{1}{\epsilon}$

Conclusion

- ✓ Using `Distributed Boosting` improves communication cost of collaborative learning **at no penalty to sample complexity**
- ✓ With classification noise, `Agnostic Distributed Boosting` does the same
- ✓ Results hold analogously for the Centralized Learning setting

Acknowledgements: This work was supported in part by the National Science Foundation under grants CCF-1815011, CCF-1934915, and CCF-1848966. This work was done while Shelby Heinecke was a student at UIC.