# Parallel Online Learning

Daniel Hsu Nikos Karampatziakis John Langford

University of Pennsylvania
Rutgers University
Cornell University
Yahoo! Research

Workshop on Learning on Cores, Clusters and Clouds

# Online Learning

- Learner gets the next example $x_t$, makes a prediction $p_t$, receives actual label $y_t$, suffers loss $\ell(p_t, y_t)$, updates itself
- Simple and fast predictions and updates

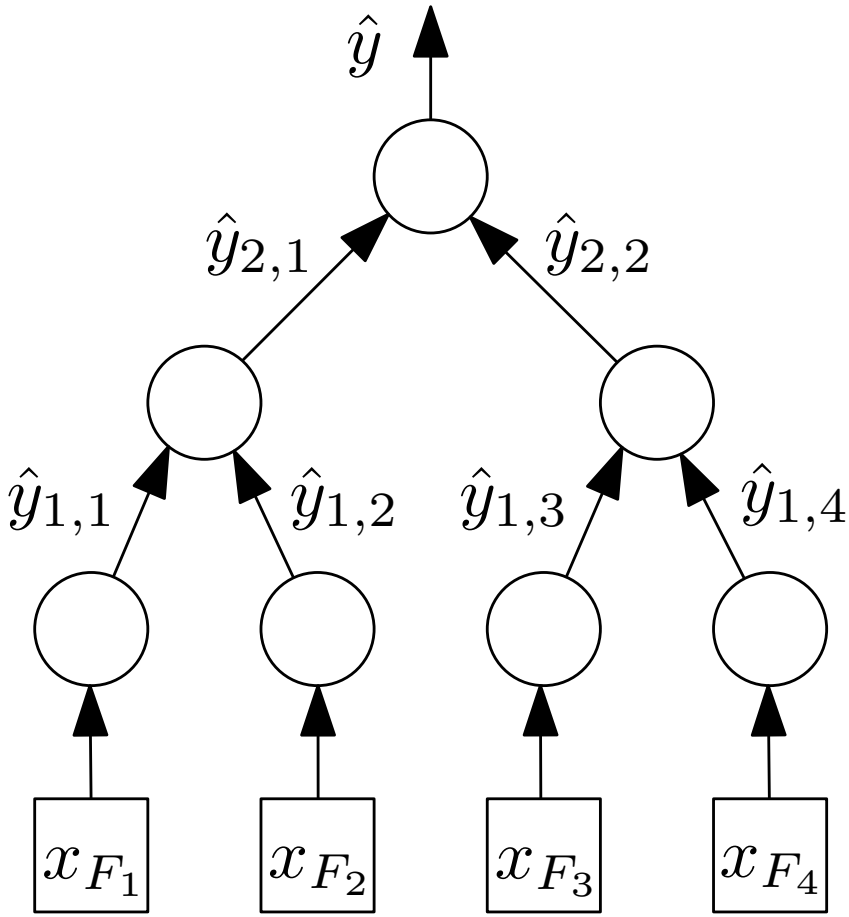
$$\begin{aligned} p_t &= w^\top x_t \\ w_{t+1} &= w_t - \eta_t \nabla \ell(p_t, y_t) \end{aligned}$$

- Online gradient descent asymptotically attains optimal regret
- Online learning scales well . . .
- . . . but it's a sequential algorithm
- What if we want to train on huge datasets?
- We investigate ways of distributing predictions, and updates while minimizing communication.

# Delay

- Parallelizing online learning leads to delay problems.
- Temporally correlated or adversarial examples.
- We investigate no delay and bounded delay schemes.

# Tree Architectures



# Local Updates

Each node in the tree:

- Computes its prediction $p_{i,j}$ based on its weights and inputs
- Sends $\hat{y}_{i,j} = \sigma(p_{i,j})$ to its parent[1]
- Updates its weights based on $\nabla \ell(p_{i,j}, y)$

No delay

Representation power: between Naive Bayes and centralized linear model.

[1] The nonlinearity introduced by $\sigma$ has an interesting effect

# Global Updates

- Local update can help or hurt.
- Improved representation power by more communication.
  - Delayed global training
  - Delayed backprop

For details and experiments come see the poster.