PERSONALIZED AND PRIVATE
PEER-TO-PEER MACHINE LEARNING

Aurélien Bellet (INRIA)

Joint work with R. Guerraoui, M. Taziki (EPFL) and M. Tommasi (INRIA)

NIPS 2017 workshop on “Machine Learning on the Phone and other Consumer Devices”
Long Beach, December 9, 2017
MOTIVATION
• Connected devices are widespread and collect increasingly large and sensitive user data

• Ex: browsing logs, health data, accelerometer, geolocation...

• Great opportunity for providing personalized services but raises serious privacy concerns

• Centralize data from all devices: best for utility, bad for privacy

• Learn on each device separately: best for privacy, bad for utility
OUR FOCUS: FULLY DECENTRALIZED NETWORK

- Personal data stays on user’s device
- Peer-to-peer and asynchronous communication
- No single point of failure/entry as in server-client architecture
- Scalability-by-design to many devices through local updates (see e.g. NIPS 2017 paper [Lian et al., 2017])
• Learn a personalized model for each user (multi-task learning)

• General idea: trade-off between model accuracy on local data and smoothness with respect to similar users
PROBLEM SETTING
PROBLEM SETTING

• A set $V = [n] = \{1, \ldots, n\}$ of $n$ learning agents

• A convex loss function $\ell : \mathbb{R}^p \times X \times Y$

• Personalized and imbalanced data: agent $i$ has dataset $S_i = \{(x^i_j, y^i_j)\}_{j=1}^{m_i}$ of size $m_i \geq 0$ drawn from $\mu_i$

• Purely local model: agent $i$ can learn a model $\theta_i$ on its own by minimizing the loss on its local data

$$L_i(\theta) = \frac{1}{m_i} \sum_{j=1}^{m_i} \ell(\theta; x^i_j, y^i_j) + \lambda_i \|\theta\|^2, \text{ with } \lambda_i \geq 0$$

• How to improve with the help of other users?
PROBLEM SETTING

• Network: weighted connected graph $G = (V, E)$
• $E \subseteq V \times V$ set of undirected edges
• Weight matrix $W \in \mathbb{R}^{n \times n}$: symmetric, nonnegative, with $W_{ij} = 0$ if $(i, j) \notin E$ or $i = j$
• Assumption: network weights are given and represent the underlying similarity between agents
Agents have only a local view of the network: they only know their neighborhood $\mathcal{N}_i = \{j \neq i : W_{ij} > 0\}$ and associated weights.
• Denoting $\Theta = [\Theta_1; \ldots; \Theta_n] \in \mathbb{R}^{np}$, we use a graph regularization formulation [Evgeniou and Pontil, 2004, Vanhaesebrouck et al., 2017]:

$$\min_{\Theta \in \mathbb{R}^{np}} \mathcal{L}(\Theta) = \frac{1}{2} \sum_{i<j}^n W_{ij} \|\Theta_i - \Theta_j\|^2 + \mu \sum_{i=1}^n D_{ii} c_i \mathcal{L}_i(\Theta_i; S_i)$$

• $\mu > 0$ trade-off parameter, $D_{ii} = \sum_j W_{ij}$ normalization factor
• $c_i \in (0, 1] \propto m_i$ is the “confidence” of agent $i$

• Implements a trade-off between having similar models for strongly connected agents and models that are accurate on their respective local datasets
NON-PRIVATE DECENTRALIZED ALGORITHM
Time and communication models:

- **Asynchronous time**: each agent has a random local clock and wakes up when it ticks
- **Broadcast communication**: agents send messages to all their neighbors at once (without expecting a reply)

**Algorithm**: assume agent $i$ wakes up at step $t$

1. Agent $i$ updates its model based on information from neighbors:

   $$\Theta_i(t + 1) = (1 - \alpha)\Theta_i(t) + \alpha \left( \sum_{j \in \mathcal{N}_i} \frac{W_{ij}}{D_{ij}} \Theta_j(t) - \mu C_i \nabla L_i(\Theta_i(t); S_i) \right)$$

2. Agent $i$ sends its updated model $\Theta_i(t + 1)$ to its neighborhood $\mathcal{N}_i$
Proposition ([Bellet et al., 2017])

For any $T > 0$, let $(\Theta(t))_{t=1}^T$ be the sequence of iterates generated by the algorithm running for $T$ iterations from an initial point $\Theta(0)$. Under appropriate assumptions, we have for some $0 < \rho < 1$:

$$
\mathbb{E} [Q_{CL}(\Theta(T)) - Q^*_{CL}] \leq (1 - \rho)^T (Q_{CL}(\Theta(0)) - Q^*_{CL}).
$$
PRIVATE ALGORITHM
• In some applications, data may be sensitive and agents may not want to reveal it to anyone else

• In our algorithms, the agents never communicate their local data but exchange sequences of models computed from data

• Consider an adversary observing all the information sent over the network (but not the internal memory of agents)

• **Goal:** how can we guarantee that no/little information about the local dataset is leaked by the algorithm?
### $(\epsilon, \delta)$-Differential Privacy

Let $\mathcal{M}$ be a randomized mechanism taking a dataset as input, and let $\epsilon > 0, \delta \geq 0$. We say that $\mathcal{M}$ is $(\epsilon, \delta)$-differentially private if for all datasets $S, S'$ differing in a single data point and for all sets of possible outputs $\mathcal{O} \subseteq \text{range}(\mathcal{M})$, we have:

$$\Pr(\mathcal{M}(S) \in \mathcal{O}) \leq e^\epsilon \Pr(\mathcal{M}(S') \in \mathcal{O}) + \delta.$$  

- Guarantees that $\mathcal{M}$ does not leak much information about any individual data point
- Information-theoretic (no computational assumptions)
DIFFERENTIALLY PRIVATE ALGORITHM

• Differentially-private algorithm:
  1. Replace the update of the algorithm by

\[
\tilde{\Theta}_i(t+1) = (1-\alpha)\tilde{\Theta}_i(t) + \alpha \left( \sum_{j \in \mathcal{N}_i} \frac{W_{ij}}{D_{ii}} \tilde{\Theta}_j(t) - \mu c_i(\nabla \mathcal{L}_i(\tilde{\Theta}_i(t); S_i) + \eta_i(t)) \right),
\]

where \( \eta_i(t) \sim \text{Laplace}(0, s_i(t))^p \in \mathbb{R}^p \)

2. Agent \( i \) then broadcasts noisy iterate \( \tilde{\Theta}_i(t + 1) \) to its neighbors
Theorem ([Bellet et al., 2017])

Let \( i \in [n] \) and assume

1. \( \ell(\cdot; x, y) \) \( L_0 \)-Lipschitz w.r.t. the \( L_1 \)-norm for all \( (x, y) \)
2. Agent \( i \) wakes up on iterations \( t^1_i, \ldots, t^T_i \)
3. For some \( \epsilon_i(t^k_i) > 0 \), the noise scale is \( s_i(t^k_i) = \frac{2L_0}{\epsilon_i(t^k_i)m_i} \)

Then for any initial point \( \tilde{\Theta}(0) \) independent of \( S_i \), the mechanism \( M_i(S_i) \) is \((\epsilon_i, 0)\)-DP with \( \bar{\epsilon}_i = \sum_{k=1}^{T_i} \epsilon_i(t^k_i) \).

- **Sweet spot**: the less data, the more noise added by the agent, but the least influence in the network
Theorem ([Bellet et al., 2017])

For any $T > 0$, let $(\tilde{\Theta}(t))_{t=1}^T$ be the sequence of iterates generated by $T$ iterations. Under appropriate assumptions, we have:

$$\mathbb{E} [Q_{CL}(\Theta(T)) - Q^*_{CL}] \leq (1 - \rho)^T (Q_{CL}(\Theta(0)) - Q^*_{CL}) + \text{additive error}.$$ 

- **Second term** gives additive error due to noise
- More results in the paper
EXPERIMENTS
• The private variant outperforms the purely local models for “reasonable” values of $\epsilon$
• Reduces (data) wealth inequality: all agents benefit but those with small dataset get a larger boost
Thank you for your attention!
Come to the poster for more details

