technische universität dortmund

Resource-Constrained On-Device Learning by Dynamic Averaging

Lukas Heppe, Michael Kamp, Linara Adilova,
Danny Heinrich, Nico Piatkowski, Katharina Morik

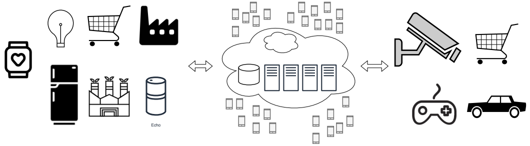
Competence Center Machine Learning Rhine-Ruhr, TU Dortmund University
lukas.heppe@tu-dortmund.de

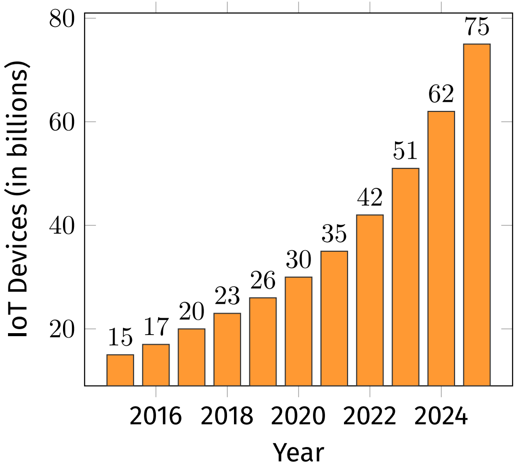
Workshop on Parallel, Distributed and Federated Learning - ECML PKDD 2020

ML2R MONASH University Fraunhofer IAIS

Motivation

- Data is generated by IOT devices
- Data contains valuable information
- Traditional setting does not scale

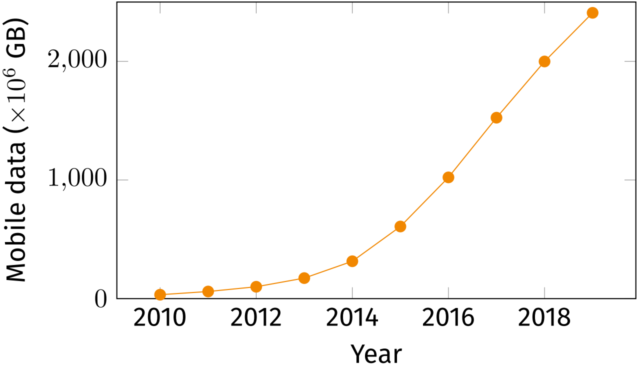




[1] Statista - Number of connected IoT Devices

Motivation - Problems with traditional machine learning

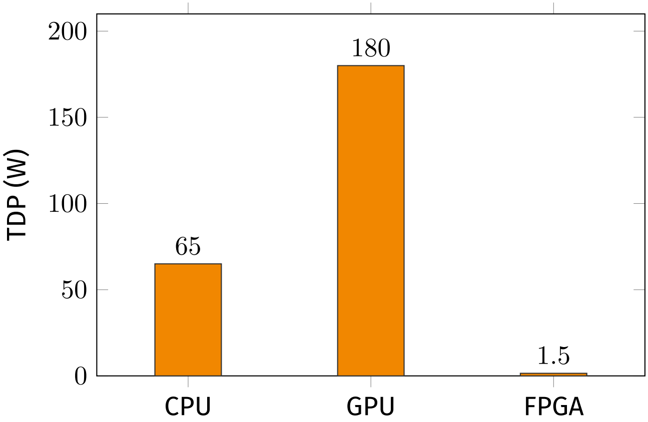
- Storage
- Bandwidth
- Privacy
- High energy consumption



[1] Energy consumption of mobile data transfer – Increasing or decreasing? Evaluating the impact of technology development & user behavior
[2] `https://findikaattori.fi/en/125`

Motivation - Problems with traditional machine learning

- Storage
- Bandwidth
- Privacy
- High energy consumption

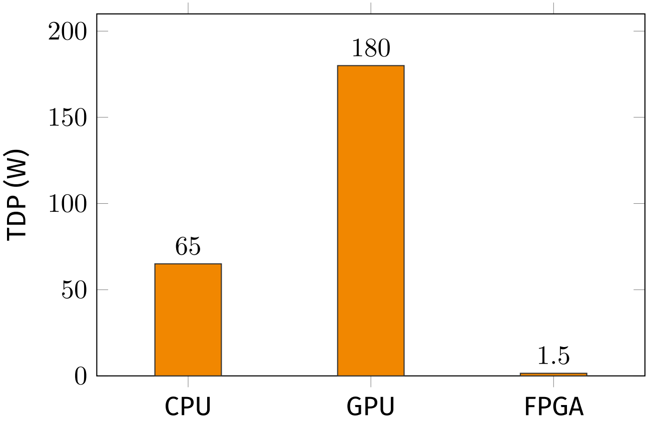


[1] Intel i7-10700
[2] Nvidia GTX 1080
[3] Kintex-7 KC705 - Hardware Accelerated Learning at the Edge, Muecke et. al, 2019

Motivation - Problems with traditional machine learning

- Storage
- Bandwidth
- Privacy
- High energy consumption

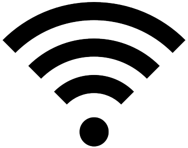
Solution: On-Device Learning



[1] Intel i7-10700
[2] Nvidia GTX 1080
[3] Kintex-7 KC705 - Hardware Accelerated Learning at the Edge, Muecke et. al, 2019

Motivation - Challenges in Resource-Constrained-Systems

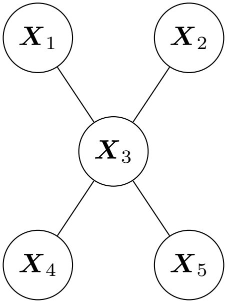
- Limited processing power / instruction sets
- Battery powered
- Network limitations



Goal: Energy- and communication efficient algorithm

Exponential Family Models [1]

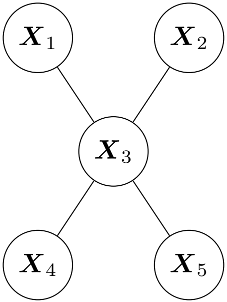
- Model distribution $\mathbb{P}$ of p-dimensional discrete random vector $\mathbf{X}$
- Exploit independencies between $\mathbf{X}_i$'s for compact representation



[1] Martin Wainwright and Michael Jordan, Graphical Models, Exponential Families, and Variational Inference, 2008

Exponential Family Models [1]

- Model distribution $\mathbb{P}$ of p-dimensional discrete random vector $\mathbf{X}$
- Exploit independencies between $\mathbf{X}_i$'s for compact representation



$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) = \frac{\exp(\langle \phi(\boldsymbol{x}), \boldsymbol{\theta} \rangle)}{Z(\boldsymbol{\theta})}$$

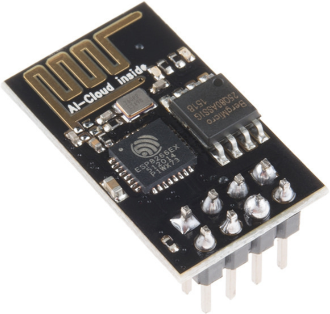
- $\boldsymbol{\theta} \in \mathbb{R}^d$ is our parameter vector
- $\phi(\boldsymbol{x}) : \mathcal{X} \mapsto \{0, 1\}^d$ sufficient statistic
- $Z(\boldsymbol{\theta})$ is the normalizer

[1] Martin Wainwright and Michael Jordan, Graphical Models, Exponential Families, and Variational Inference, 2008

Integer Exponential Family [1]

- Restrict $\boldsymbol{\theta} \subseteq \{p \mid p \in \mathbb{N} \wedge p \leq k\}^d = \mathbb{N}^d_{\leq k}$
- Change base from $\exp$ to 2
- Store probabilities as fraction a/b

$$\mathbb{P}_{\boldsymbol{\theta}}(\boldsymbol{X} = \boldsymbol{x}) = \frac{2^{\langle \phi(\boldsymbol{x}), \boldsymbol{\theta} \rangle}}{Z(\boldsymbol{\theta})}$$



Esp 8266

[1]Nico Piatkowski, Exponential families on resource-constrained systems, 2018

Integer Exponential Family - Learning

- Given a dataset $\mathcal{D} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\}$
- Estimate $\boldsymbol{\theta}$ using maximum-likelihood-estimation
- Denote $\widehat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{\boldsymbol{x}\in\mathcal{D}} \phi(\boldsymbol{x})$ and solve

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} = \log Z(\boldsymbol{\theta}) - \langle \boldsymbol{\theta}, \widehat{\boldsymbol{\mu}} \rangle$$

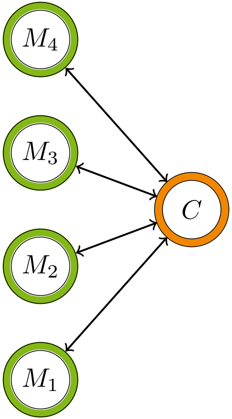
Iterative Learning:

- Update $\widehat{\boldsymbol{\mu}}_t$ as running average
- Solve problem for $\widehat{\boldsymbol{\mu}}_t$

Distributed machine learning

Setting

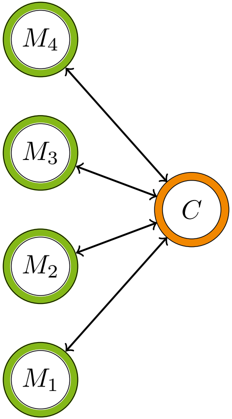
- Set of m learners connected to coordinator
- Data-generating-distribution $\mathcal{Q}(\mathcal{X}, \mathcal{Y})$
- Online round-based learning process



Distributed machine learning

Setting

- Set of m learners connected to coordinator
- Data-generating-distribution $\mathcal{Q}(\mathcal{X}, \mathcal{Y})$
- Online round-based learning process



Questions?

- What do we communicate?
- When do we communicate?
- How to merge information?

What do we communicate?

Protocol	Centralized	Naïve	Privacy
Send	$\boldsymbol{\mu}_i$	$\{\boldsymbol{\mu}_i, \boldsymbol{\theta}_i\}$	$\boldsymbol{\theta}_i$
Receive	$\boldsymbol{\theta}$	$\{\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\theta}}\}$	$\widehat{\boldsymbol{\theta}}$

What do we communicate?

Protocol	Centralized	Naïve	Privacy
Send	$\boldsymbol{\mu}_i$	$\{\boldsymbol{\mu}_i, \boldsymbol{\theta}_i\}$	$\boldsymbol{\theta}_i$
Receive	$\boldsymbol{\theta}$	$\{\widehat{\boldsymbol{\mu}}, \widehat{\boldsymbol{\theta}}\}$	$\widehat{\boldsymbol{\theta}}$

- Assuming 16 learners, problem dimension 1000, 32 bit per parameter, 3 bit for integer

(bytes)	Centralized	Naïve	Privacy
Regular	1.152.000	2.304.000	1.536.000
Integer	456.000	912.000	144.000
Reduction	2.5	2.5	10

When do we communicate?

Periodic synchronization

- Transmit changes on data arrival
- Control frequency via parameter b
- Transmit if:

$$t \mod b = 0$$

When do we communicate?

Periodic synchronization

- Transmit changes on data arrival
- Control frequency via parameter b
- Transmit if:

$$t \mod b = 0$$

Dynamic synchronization

- Define reference-vector $\boldsymbol{r} \in \mathbb{N}^d$
- Define divergence-threshold $\Delta \in \mathbb{N}$
- Transmit if:

$$\|\boldsymbol{\theta}^i - \boldsymbol{r}\|_2^2 \geq \Delta$$

Model aggregation

How to combine the data?

- Focus on simple average

$$\widehat{\boldsymbol{\theta}} = \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\theta}^i$$

- Floored average can be calculated using integer only

$$\overline{\boldsymbol{\theta}} = \left\lfloor \frac{1}{m}\sum_{i=1}^{m}\boldsymbol{\theta}^i \right\rfloor = \left(\sum_{i=1}^{m}\boldsymbol{\theta}^i\right) >> \log_2(m)$$

- Hierarchical reduction to prevent overflows

$$\left\lfloor \frac{a+b}{2} \right\rfloor = (a \wedge b) + ((a \oplus b) >> 1)$$

Theoretical guarantees

Bound approximation errors:

- Integer exponential family models are not arbitrary worse Error = ϵ [1]
- Error of dynamic avg. is bounded over periodic error [2]
- Distance between averages is bounded by dimension: $\|\widehat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}}\|_2^2 \leq \sqrt{d}$

[1] Nico Piatkowski, Exponential families on resource-constrained systems, 2018
[2] Michael Kamp, Black-Box Parallelization for Machine Learning, 2019

Theoretical guarantees

Bound approximation errors:

- Integer exponential family models are not arbitrary worse Error = ϵ [1]
- Error of dynamic avg. is bounded over periodic error [2]
- Distance between averages is bounded by dimension: $\|\widehat{\boldsymbol{\theta}} - \overline{\boldsymbol{\theta}}\|_2^2 \leq \sqrt{d}$

- **Combined error is bounded!**

[1] Nico Piatkowski, Exponential families on resource-constrained systems, 2018
[2] Michael Kamp, Black-Box Parallelization for Machine Learning, 2019

Experimental Evaluation

Questions

- Periodic vs. Dynamic averaging?
- Regular vs. integer exponential family?
- Different communication schemes?

Criteria

- Model quality?
- Bandwidth savings?
- Energy savings?

Experimental Evaluation - General setup

- Distributed learning environment with 16 clients
- Estimation of graph structure on holdout dataset
- Integer learner parameter space $\mathbb{N}_{\leq 8}$
- Horizontal partition of data
- Evaluation of system performance via cumulative 0/1 loss:

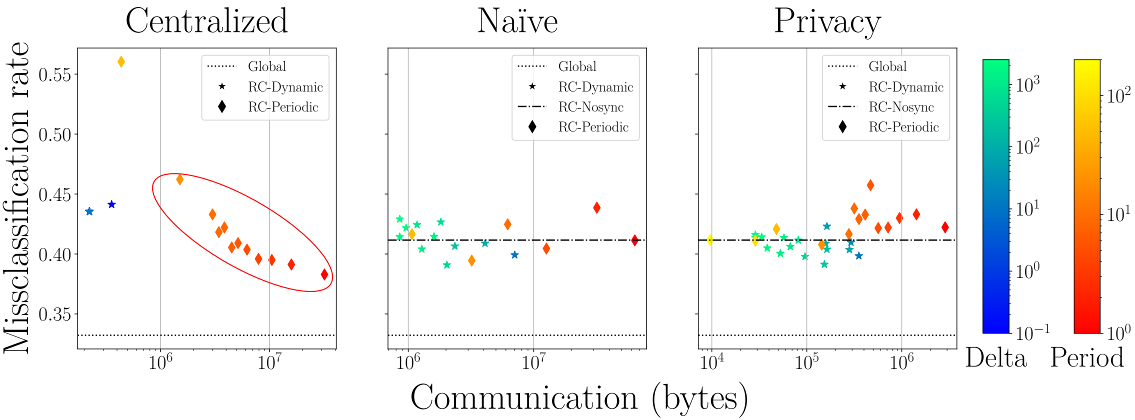
$$\frac{1}{Tm}\sum_{t=1}^{T}\sum_{i=1}^{m}\frac{1}{|S_i^t|}\sum_{(\boldsymbol{x},y)\in S_i^t}\ell(\boldsymbol{x},y|\boldsymbol{\theta}_i^t)$$

Experimental Setup - Procedure and paramters

Online round-based procedure:

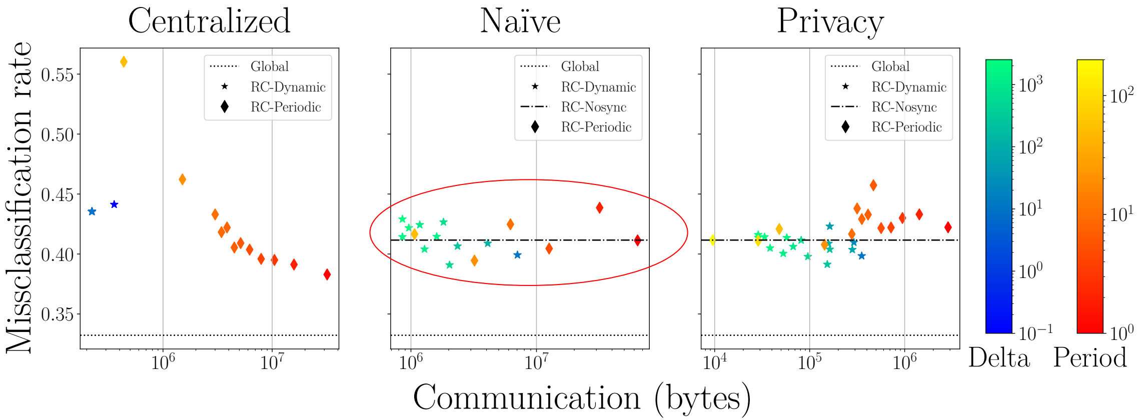
- Receive a batch of data ($bs = 10$)
- Predict labels for batch, compute performance
- Update $\widehat{\mu}$ and run optimization for a given budget ($i = 10$)
- Communicate depending on protocol

Result - Privacy Preserving Averaging vs. Centralized



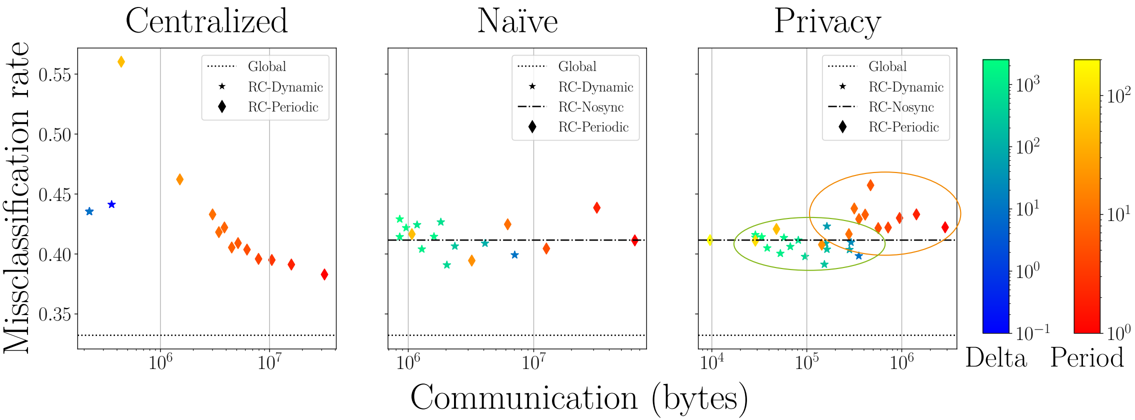
- Privacy-aware averaging retains performance while using less bandwidth

Result - Privacy Preserving Averaging vs. Centralized



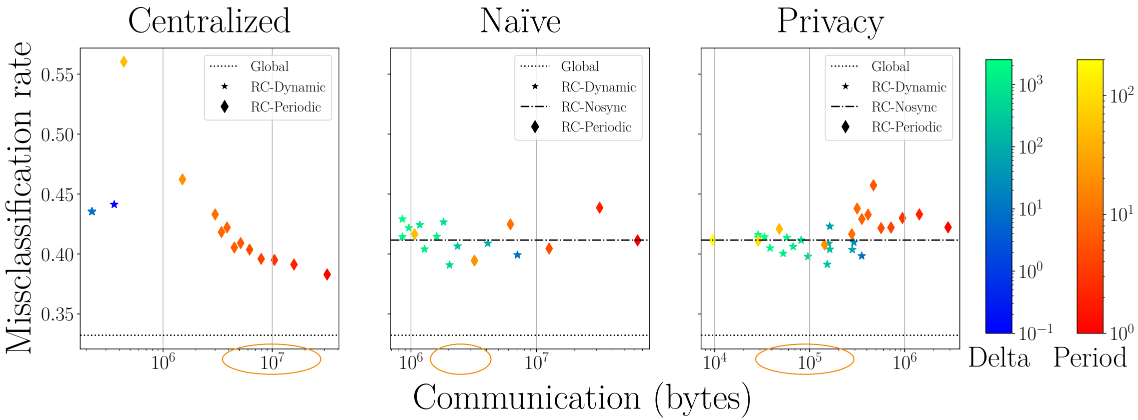
- Privacy-aware averaging retains performance while using less bandwidth

Result - Privacy Preserving Averaging vs. Centralized



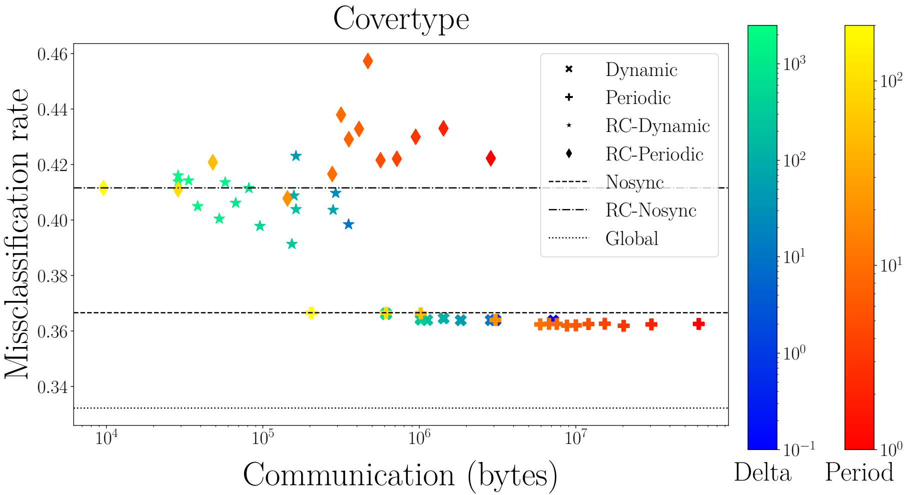
- Privacy-aware averaging retains performance while using less bandwidth

Result - Privacy Preserving Averaging vs. Centralized



- Privacy-aware averaging retains performance while using less bandwidth

Result - Privacy Preserving Averaging - Covertype



Future work

- Scale number of learners
- Communication vs. Rounding impact
- Methods to select optimal hyperparameters
- Modular parameter updates
- Adaptation to non i.i.d / time-variant data

Conclusion

- Distributed integer-only learning is possible
- Energy vs. Performance-Tradeoff
- Bandwidth reduction of $193\times$
- Estimated energy reduction of $67\times$

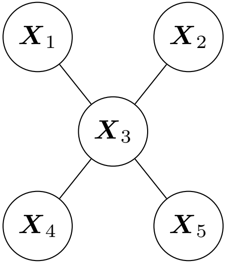
ML2R

MONASH University

Fraunhofer IAIS

Backup

Exponential Family Models - Sufficient statistic



- Assume binary variables ($\forall i\ \mathcal{X}_i = \{0,1\}$)
- Let $\boldsymbol{x} = [0,1,0,0,1]^\top$
- $\phi(\boldsymbol{x}) = [\phi_{x_1x_3}(\boldsymbol{x}), \phi_{x_2x_3}(\boldsymbol{x}), \phi_{x_3x_4}(\boldsymbol{x}), \phi_{x_3x_5}(\boldsymbol{x})]^\top$

$$\phi_{x_1x_3}(\boldsymbol{x}) = \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \quad \phi_{x_2x_3}(\boldsymbol{x}) = \begin{bmatrix}0\\0\\1\\0\end{bmatrix}, \quad \phi_{x_3x_4}(\boldsymbol{x}) = \begin{bmatrix}1\\0\\0\\0\end{bmatrix}, \quad \phi_{x_3x_5}(\boldsymbol{x}) = \begin{bmatrix}0\\1\\0\\0\end{bmatrix}$$

- $d = \sum_{(s,t)\in E} \mathcal{X}_s \cdot \mathcal{X}_t = 4 \cdot 2^2$
- Assuming 10 states per variable, we store $4 \cdot 10^2$ instead of 10^5