Research Group on Artificial Intelligence

UNIVERSITAS SCIENTIARUM SZEGEDIENSIS

Distributed Differentially Private Stochastic Gradient Descent: An Empirical Study

István Hegedűs and Márk Jelasity

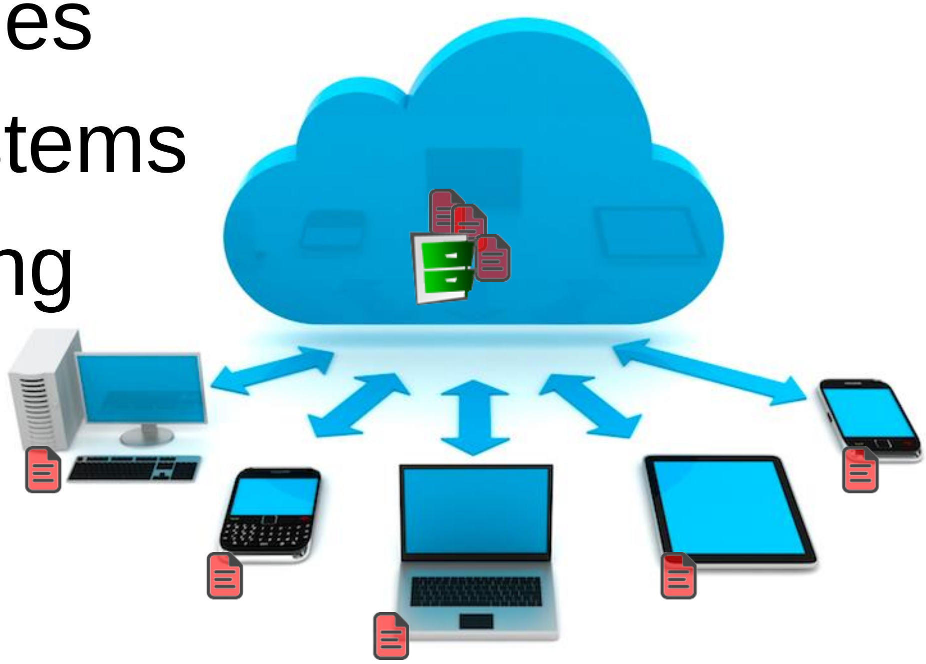
University of Szeged
MTA-SZTE Research Group on AI
Hungary

Motivation

- Data is accumulated in data centers
- Costly storage and processing
  - Maintenence, Infrastructure, Privacy
- Limited access
  - For researchers as well
- But, data was produced by us

Motivation – ML Applications

- Personalized Queries
- Recommender Systems
- Document Clustering
- Spam Filtering

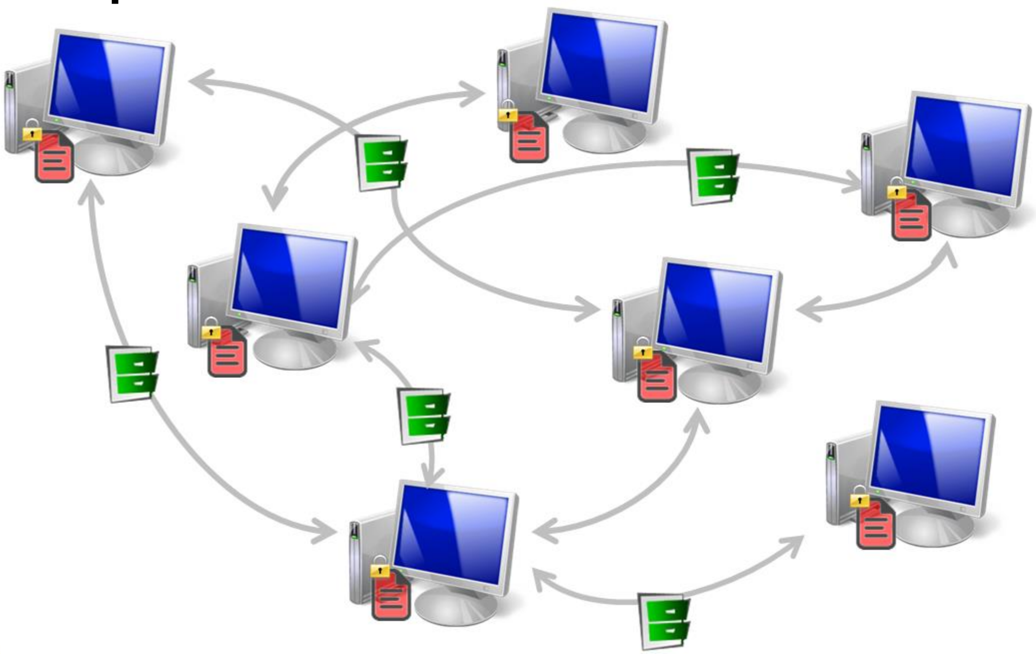


Gossip Learning

- ML is often an optimization problem
- Local data is not enough

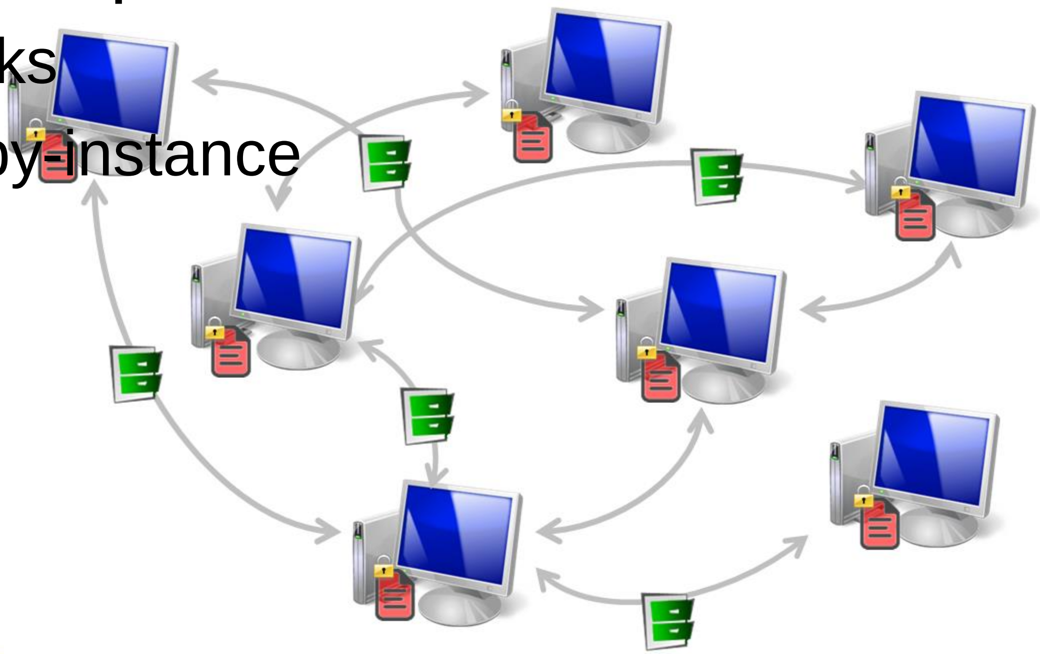
Gossip Learning

- ML is often an optimization problem
- Local data is not enough
- Models are sent and updated on nodes



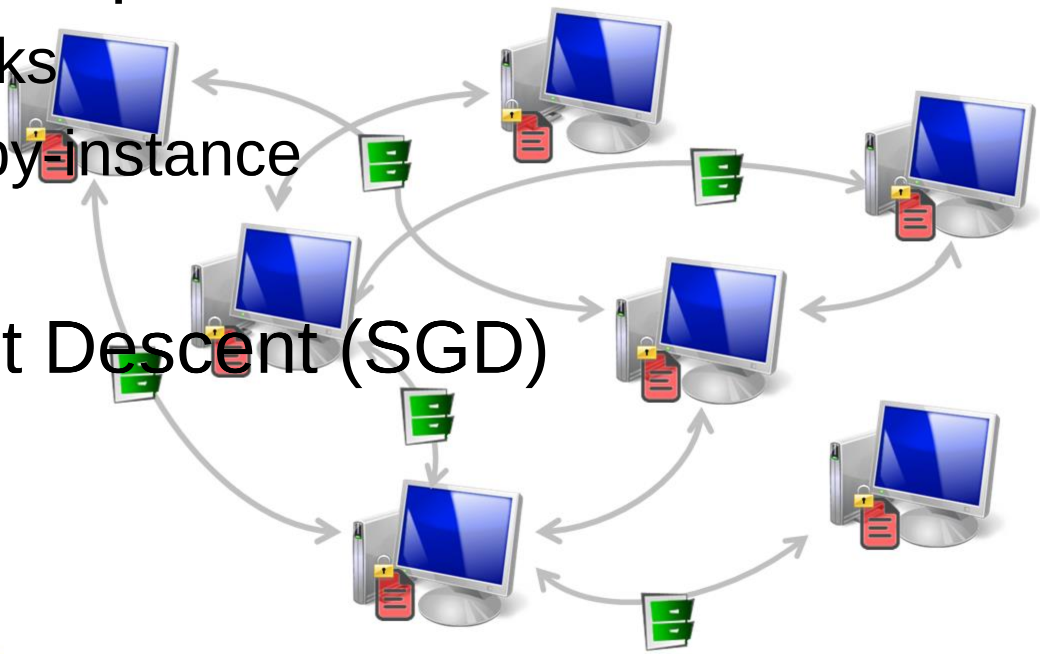
Gossip Learning

- ML is often an optimization problem
- Local data is not enough
- Models are sent and updated on nodes
  - Taking random walks
  - Updated instance-by-instance
  - Data is never sent



Gossip Learning

- ML is often an optimization problem
- Local data is not enough
- Models are sent and updated on nodes
  - Taking random walks
  - Updated instance-by-instance
  - Data is never sent
- Stochastic Gradient Descent (SGD)



SGD

- Objective function

$$w = \arg\min_{w} J(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \frac{\lambda}{2}\|w\|^2$$

SGD

- Objective function

$$w = \arg\min_{w} J(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \frac{\lambda}{2}\|w\|^2$$

- Gradient method

$$\begin{aligned} w_{t+1} &= w_t - \eta_t\left(\frac{\partial J}{\partial w}\right) \\ &= w_t - \eta_t\left(\lambda w + \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(f_w(x_i), y_i)\right) \end{aligned}$$

SGD

- Objective function

$$w = \arg\min_{w} J(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \frac{\lambda}{2}\|w\|^2$$

- Gradient method

$$\begin{aligned} w_{t+1} &= w_t - \eta_t(\frac{\partial J}{\partial w}) \\ &= w_t - \eta_t(\lambda w + \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(f_w(x_i), y_i)) \end{aligned}$$

- SGD, data can be processed online (instance by instance)

$$w_{t+1} = w_t - \eta_t(\lambda w + \nabla \ell(f_w(x_i), y_i))$$

SGD

- Objective function

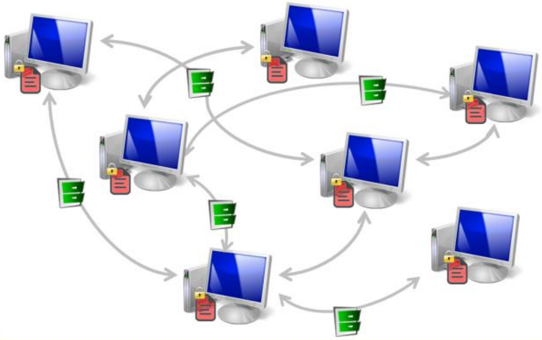
$$w = \arg\min_{w} J(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \frac{\lambda}{2}\|w\|^2$$

- Gradient method

$$\begin{aligned} w_{t+1} &= w_t - \eta_t\left(\frac{\partial J}{\partial w}\right) \\ &= w_t - \eta_t\left(\lambda w + \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(f_w(x_i), y_i)\right) \end{aligned}$$

- SGD, data can be processed online (instance by instance)

$$w_{t+1} = w_t - \eta_t(\lambda w + \nabla \ell(f_w(x_i), y_i))$$



SGD

- Objective function

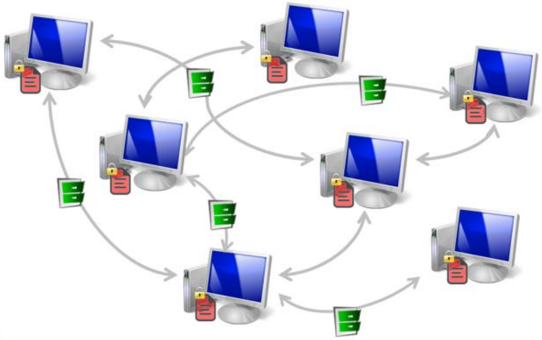
$$w = \arg\min_{w} J(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i) + \frac{\lambda}{2}\|w\|^2$$

- Gradient method

$$\begin{aligned} w_{t+1} &= w_t - \eta_t(\frac{\partial J}{\partial w}) \\ &= w_t - \eta_t(\lambda w + \frac{1}{n}\sum_{i=1}^{n} \nabla \ell(f_w(x_i), y_i)) \end{aligned}$$

- SGD, data can be processed online (instance by instance)
  – Data can be guessed by specifically crafted models

$$w_{t+1} = w_t - \eta_t(\lambda w + \nabla \ell(f_w(x_i), y_i))$$



Differential Privacy

- Privacy: $w_{t+1} = w_t - \eta_t(\lambda w_t + \nabla \ell(f_w(x_i), y_i) + N_t)$

  adding appropriately generated noise

Differential Privacy

- Privacy: $w_{t+1} = w_t - \eta_t(\lambda w_t + \nabla \ell(f_w(x_i), y_i) + N_t)$
  adding appropriately generated noise
- Differential Privacy theoretically guarantees the indistinguishability

$$\forall x : e^{-\epsilon} \leq \frac{P(F(D) = x)}{P(F(D') = x)} \leq e^{\epsilon}$$

  – Based on the global sensitivity

$$Z_F = \max_{D,D' \text{ differ in one record}} \|F(D) - F(D')\|_1$$

Differential Privacy

- Privacy: $w_{t+1} = w_t - \eta_t(\lambda w_t + \nabla \ell(f_w(x_i), y_i) + N_t)$
  adding appropriately generated noise
- Differential Privacy theoretically guarantees the indistinguishability

$$\forall x : e^{-\epsilon} \leq \frac{P(F(D) = x)}{P(F(D') = x)} \leq e^{\epsilon}$$

  - Based on the global sensitivity

$$Z_F = \max_{D,D' \text{ differ in one record}} \|F(D) - F(D')\|_1$$

- Every data instance has a privacy budget

Experimental Setup

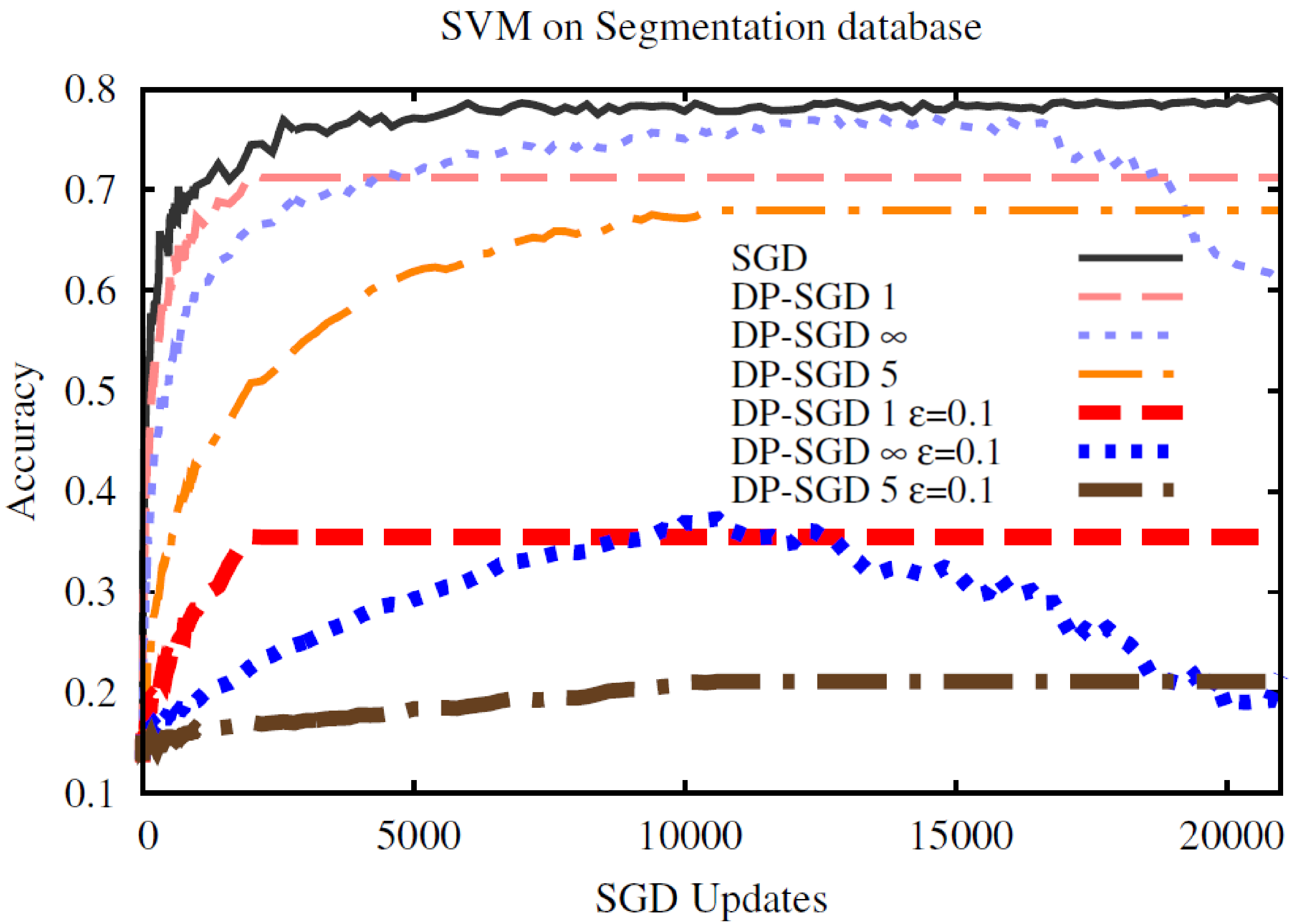
	MNIST	Segmentation	Spambase
Training set size	60 000	2310	4140
Test set size	10 000	210	461
Number of features	784	19	57
Number of classes	10	7	2
Class-label distribution	uniform	uniform	6:4

- Data sets
- Budget management
  - One shot: DP-SGD-1
  - Equipartition: DP-SGD-5
  - Exponential: DP-SGD-∞
- Various normalizations
- Measurement

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^{n}\delta(y_i = f_w(x_i))$$

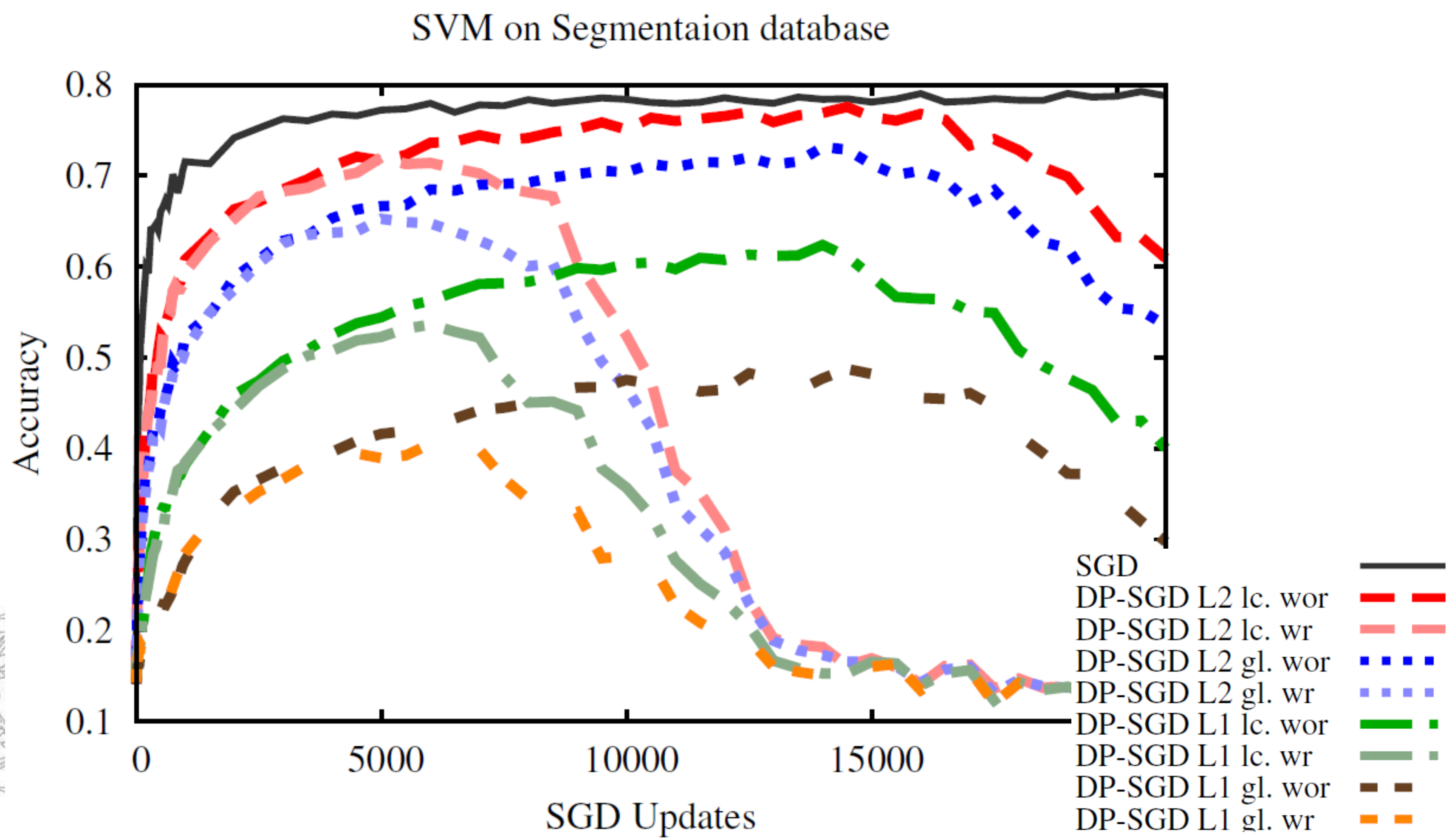
Experimental Results

Budget management and Privacy level



Experimental Results

Norms and Data sampling



Conclusion

- Privacy preserving SGD for fully distributed data mining
- Close to optimal accuracy without additional communication cost
- Influence of the
  - Normalization
  - Budget management
  - Data sampling
- Better performance can be achieved with more local data