sponsored by the German Federal Ministry of Education and Research (BMBF) - 16KIS1004
and by German Research Foundation (DFG) - SCHA 1944/7-1

2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), June 2021 • Toronto, Ontario, Canada

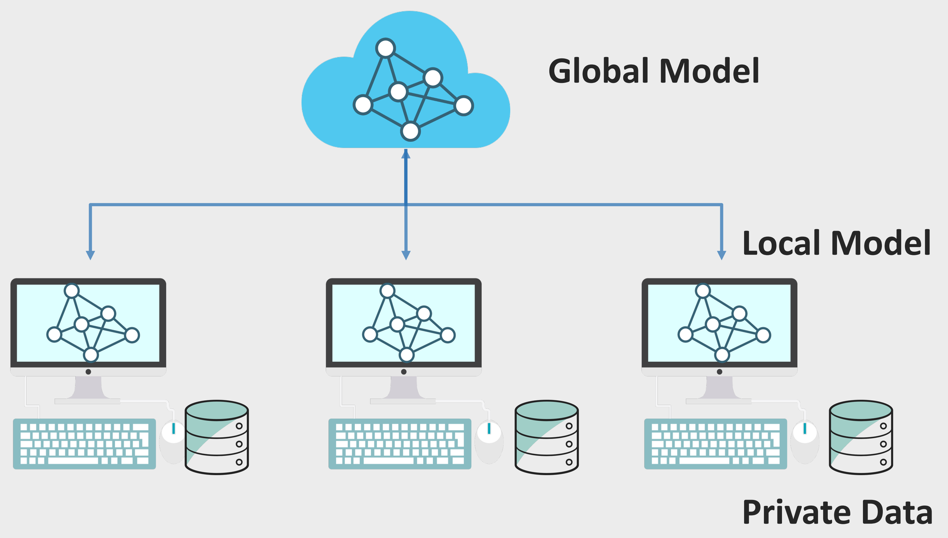
Federated Learning with Local Differential Privacy: Trade-offs between Privacy, Utility, and Communication

Muah Kim[1], Onur Günlü[1], and Rafael F. Schaefer[2]

[1] Information Theory and Applications Chair, TU Berlin, Germany, [2] Chair of Communications Engineering and Security, Universität Siegen, Germany
{muah.kim, guenlue}@tu-berlin.de, {rafael.schaefer}@uni-siegen.de

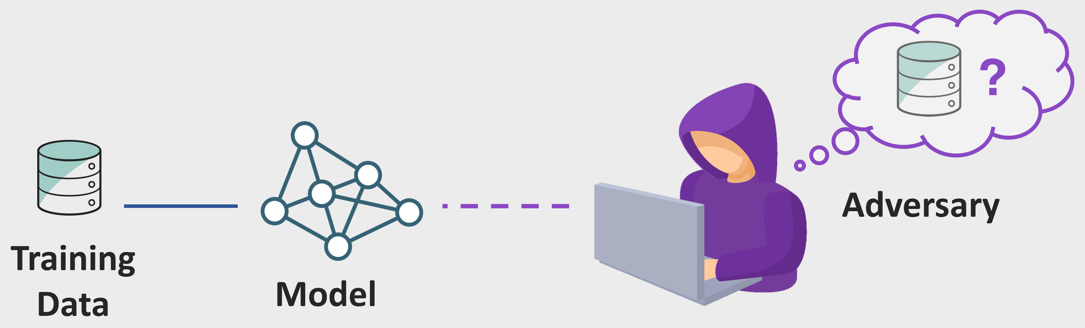
Motivation – Privacy of Machine Learning

1. Federated Learning (FL)



The global model is achieved by merging locally trained models.
⇒ More **private** than a centralized model: users' data remain at users.
⇒ It costs **iterative communication** for every weight updates.

2. Privacy Preserving Scheme



The training data can be partially recovered by using the model.
⇒ **Privacy preserving scheme** is required,
⇒ while minimizing learning **performance degradation** from it.

FL + privacy preserving scheme

Privacy, Utility, Transmission Rate should be **jointly considered**.
Privacy after T iterations should be tightly accounted.

Main Contribution

1. We analyze the bounds of privacy, utility, and transmission rate of an FL model with stochastic gradient descent (SGD) algorithm and Gaussian mechanism.
2. The **trade-offs** between three metrics are observed: Privacy ⇑ – Utility ⇓ – Transmission Rate ⇑
3. The trade-offs are improved by adopting an enhanced privacy accounting method over many iterations. [1]
4. A **generalized** FL model is assumed: heterogeneity of users, variable query sensitivity of the privacy mechanism, parameterized gradient norm clipping threshold, etc.

System Modeling

Algorithm: FL-SGD with Gaussian Mechanism

Input: User datasets $\{\mathcal{D}_k\}_{k=1}^K$, data sampling rates $\{q_k\}_{k=1}^K$, sampled datasets $\left\{\mathcal{J}_k^{(t)}\right\}_{k=1}^K$, total sampled dataset $\mathcal{J}^{(t)} = \cup_{k=1}^K \mathcal{J}_k^{(t)}$, loss function $\mathcal{L}_k\left(\boldsymbol{w}^{(t)}, \mathcal{J}_k^{(t)}\right) = \frac{1}{\left|\mathcal{J}_k^{(t)}\right|}\sum_{x\in\mathcal{J}_k^{(t)}} \ell\left(\boldsymbol{w}^{(t)}, x\right)$.

Parameters: learning rate η_t, noise scale $\{\sigma_k\}_{k=1}^K$, clipping norm threshold C.

Initialize $\boldsymbol{w}^{(0)}$ randomly
for $t \in [0: T-1]$ **do** — Training T iterations in total
  for $k \in [K]$ — Training user k's local model
    Download $\boldsymbol{w}^{(t)}$
    Sample $\mathcal{J}_k^{(t)}$ from $\mathcal{D}_k$ with probability q_k
    Compute gradient
    $\boldsymbol{g}_k^{(t)}\left(\boldsymbol{w}^{(t)}, \mathcal{J}_k^{(t)}\right) \leftarrow \nabla_{\boldsymbol{w}}\mathcal{L}_k\left(\boldsymbol{w}^{(t)}, \mathcal{J}_k^{(t)}\right)$
    Gradient Norm Clipping
    $\overline{\boldsymbol{g}}_k^{(t)} = \frac{\boldsymbol{g}_k^{(t)}}{\max\left(1, \left\|\boldsymbol{g}_k^{(t)}\right\|/C\right)}$
    Add Gaussian Noise
    $\tilde{\boldsymbol{g}}_k^{(t)} = \mathcal{M}_k\left(\overline{\boldsymbol{g}}_k^{(t)}\right) = \overline{\boldsymbol{g}}_k^{(t)} + \mathcal{N}\left(\boldsymbol{0}, C^2\sigma_k^2\boldsymbol{I}_d\right)$
    Upload $\tilde{\boldsymbol{g}}_k^{(t)}$
  Weight Update
  $\boldsymbol{w}^{(t+1)} = \boldsymbol{w}^{(t)} - \eta_t \cdot \sum_{k=1}^K \frac{\left|\mathcal{J}_k^{(t)}\right|}{\left|\mathcal{J}^{(t)}\right|}\tilde{\boldsymbol{g}}_k^{(t)}$

Output: $\boldsymbol{w}^{(T)}$

Performance Metrics

Local Differential Privacy (LDP) of User k

Definition. (ϵ_k, δ_k) – **Local Differential Privacy (LDP)**
Gaussian mechanism $\mathcal{M}_k$ is (ϵ_k, δ_k) –**LDP** w.r.t. dataset $\mathcal{D}_k$, if ∀ two **neighboring** datasets $D \sim D' \subseteq \mathcal{D}_k$ and $\forall \mathcal{S} \subseteq Range(\mathcal{M}_k)$

$$\Pr[\mathcal{M}_k(\overline{\boldsymbol{g}}_k(D)) \in \mathcal{S}] \le e^{\epsilon_k} \cdot \Pr\left[\mathcal{M}_k\left(\overline{\boldsymbol{g}}_k(D')\right) \in \mathcal{S}\right] + \delta_k.$$

Gaussian Mechanism for (ϵ_k, δ_k) –LDP
A Gaussian mechanism $\mathcal{M}_k\left(\overline{\boldsymbol{g}}_k^{(t)}\right) = \overline{\boldsymbol{g}}_k^{(t)} + \mathcal{N}\left(\boldsymbol{0}, C^2\sigma_k^2\boldsymbol{I}_d\right)$ satisfies (ϵ_k, δ_k) –LDP

$$\text{if } \delta_k \ge \tfrac{4}{5}\exp(-(C\sigma_k\epsilon_k)^2/2).$$

Global Utility is defined by the multiplicative inverse of the convergence rate, i.e.,

$$\mathcal{U}(T) = \frac{1}{\mathbb{E}[\mathcal{L}(\boldsymbol{w}^{(T)}, \mathcal{J}^{(T)}) - \mathcal{L}(\boldsymbol{w}^*, \cup_{k=1}^K \mathcal{D}_k)]}.$$

$$\boldsymbol{w}^* = arg\min_{\boldsymbol{w}\in\mathbb{R}^d} \mathcal{L}\left(\boldsymbol{w}^{(t)}, \cup_{k=1}^K \mathcal{D}_k\right)$$: optimal weight vector

Transmission rate is defined by the differential entropy of a noisy gradient: $R_{tr,k} = h\left(\tilde{\boldsymbol{g}}_k^{(t)}\right)$.

Theoretical Analysis

Theoretical Bounds on the Metrics

Theorem* 1.** *For* $\epsilon_k > 2\log(\delta_k^{-1})\max\left\{\delta_k, \frac{1}{\sigma_k^2 \ln\frac{1}{q_k\sigma_k}}\right\}$, $q_k < \frac{1}{16\sigma_k}$, *and* $\sigma_k \ge 1$, *user* k*'s Gaussian mechanism* $\mathcal{M}_k$ *is* (ϵ_k, δ_k) – ***LDP *after* T *iterations if*

$$\sigma_k^2 \ge \frac{4q_k^2T}{1-q_k}\left[\frac{2}{\epsilon_k^2}\log\frac{1}{\delta_k} + \frac{1}{\epsilon_k} - \frac{2}{\epsilon_k^2}\left(\log(2\log\delta_k^{-1}) + 1 - \log\epsilon_k\right)\right] + \mathcal{O}\left(\frac{\log^2(\log\delta_k^{-1})}{\log\delta_k^{-1}}\right).$$

For a μ – *smooth and* λ – *strongly convex loss* $\mathcal{L}(\boldsymbol{w}; \mathcal{S})$ *w.r.t. a* d – *dimensional weight vector* $\boldsymbol{w} \in \mathbb{R}^d$ *given an arbitrary subset* $\mathcal{S}$ *of* $\mathcal{D} = \cup_{k=1}^K \mathcal{D}_k$, *i.e.,* $\mathcal{S} \subseteq \mathcal{D}$, *and for a learning rate* $\eta_t = \frac{G}{C\lambda t}$,

$$\mathcal{U}(T) \ge \frac{\lambda^2T}{\mu G^2}\min\left\{\frac{1}{2}, \frac{1}{1+d\sigma^2}\right\}$$

where $\sigma^2 = \frac{\sum_{k=1}^K(|\mathcal{D}_k|q_k\sigma_k)^2}{(\sum_{k=1}^K|\mathcal{D}_k|q_k)^2}$, *and* G *is the maximum norm of the gradient. The transmission rate* $R_{tr,k}$ *of user* k *wth Gaussian noise* $\mathcal{N}\left(0, C^2\sigma_k^2\boldsymbol{I}_d\right)$ *satisfies*

$$R_{\mathrm{tr},k} \le d\log_2\left(\frac{2\pi e C^2\sigma_k}{\sqrt{d}}\right) \quad (\textit{bits per gradient}).$$

Theoretical Analysis (cont'd)

- We set the number of iterations T and obtain the range of σ_k^2 that can achieve a target LDP level (ϵ_k, δ_k).
- The noise variance σ_k^2 connects the target privacy (ϵ_k, δ_k) with the bounds of utility $\mathcal{U}(T)$ and transmission rate $R_{\mathrm{tr},k}$:
- Privacy ⇑ & ϵ_k ⇓ – σ_k ⇑ – Utility ⇓ – Transmission Rate ⇑.

⇒ **Trade-off relationship**

Comparison with other Privacy Accounting Methods

Moment Accountant (MA)

$$\sigma_k^2 \ge \frac{4q_k^2T}{1-q_k}\left[\frac{2}{\epsilon_k^2}\log\frac{1}{\delta_k} + \frac{1}{\epsilon_k} + \mathcal{O}\left(\log\delta_k^{-1}\right)\right]$$

Advanced Composition 1 (AC1)

$$\sigma_k^2 \ge \frac{4q_k^2}{1-q_k}\cdot\frac{2}{\epsilon_0}\log\left(\frac{4}{5\delta_0}\right)$$

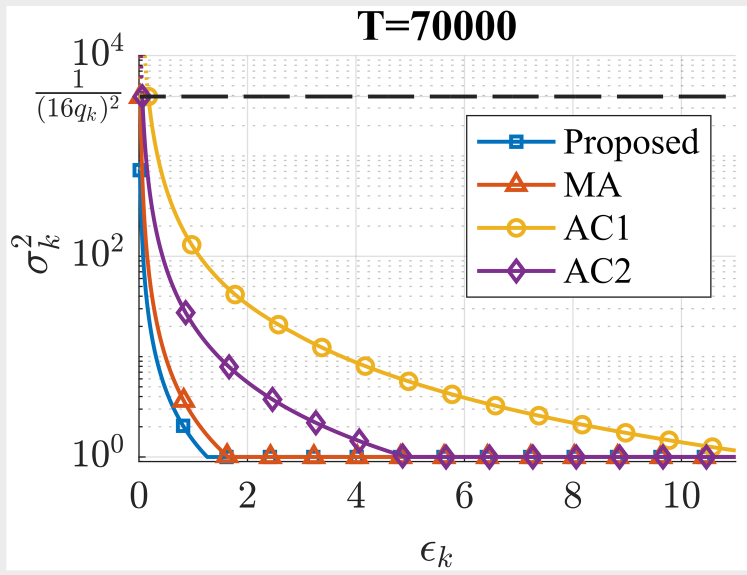
Advanced Composition 2 (AC2)

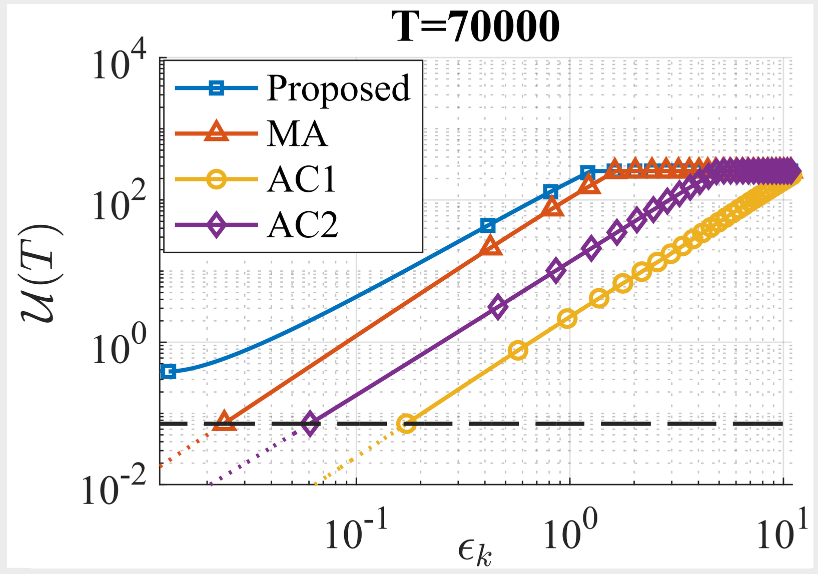
$$\sigma_k^2 \ge \frac{4q_k^2}{1-q_k}\cdot\frac{8T\log\left(e+\frac{\epsilon_k}{\delta_k}\right)}{\epsilon_k^2}$$

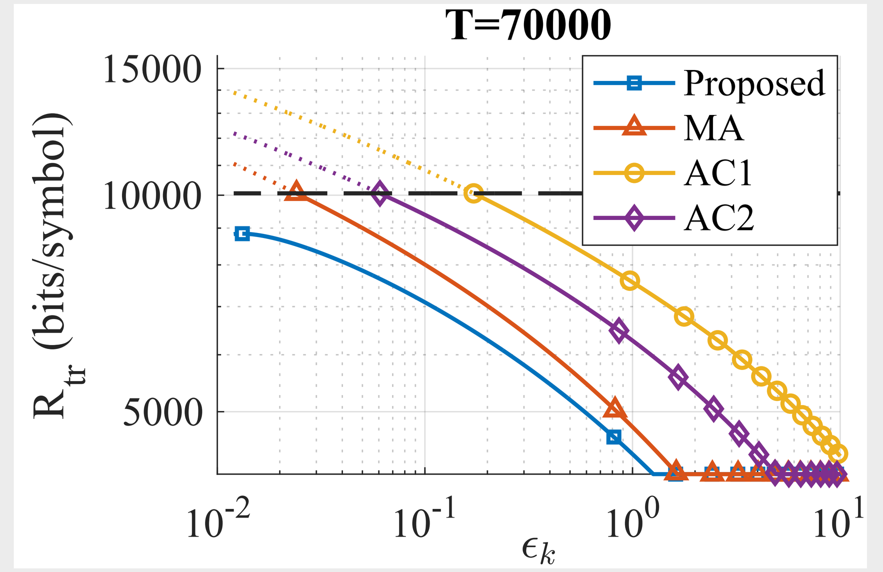
- These three methods are commonly used for privacy analysis.
- The bound of σ_k^2 in ***Theorem* 1**. is smaller than that of **(MA)**.
- **(MA)** is shown to outperform **(AC1)** and **(AC2).** [2]

⇒ The **proposed** bound of σ_k^2 is expected to be the smallest.

Interpretation by Simulations







- Parameters: $K = 100, \delta_k = 10^{-4}, q_k = 10^{-3}$ *for all* $k = 1,2,\dots,100, d = 10^4, \mu = 1, \lambda = 1, C = 1,$ *and* $G = 5$.
- The bounds of σ_k^2, **utility** $\mathcal{U}(T)$ and **transmission rate** $R_{\mathrm{tr},k}$ are plotted for varying ϵ_k by using σ_k^2 as a parameter.
- **Trade-offs** between three metrics are observed: the utility increases and transmission rate decreases as ϵ_k grows, i.e., the target privacy becomes weaker.
- The proposed scheme accounts the differential privacy **the tightest** compared to using the other privacy accounting methods, which results in a smaller noise level σ_k^2 to achieve the same (ϵ_k, δ_k) –LDP after $T = 70{,}000$ iterations and, accordingly, a greater guaranteed value of utility and a smaller worst case transmission rate. ⟹ **Better Trade-offs**
- Due to the condition $q_k < \frac{1}{16\sigma_k}$, the domain of ϵ_k is restricted, and it restricts the ranges by the black dashed lines in the graphs. By the condition $\sigma_k \ge 1$, the curves show saturating behavior when σ_k reaches to 1 as ϵ_k increases.

[1] S. Asoodeh, J. Liao, F. P. Calmon, O. Kosut and L. Sankar, "A Better Bound Gives a Hundred Rounds: Enhanced Privacy Guarantees via f-Divergences," *2020 IEEE International Symposium on Information Theory (ISIT)*, Los Angeles, CA, USA, 2020, pp. 920-925, doi: 10.1109/ISIT44484.2020.9174015.
[2] M. Abadi, A. Chu, I. Goodfellow, H B. McMahan, I. Mironov, K. Talwar, and L. Zhang, "Deep learning with differential privacy," in ACM Conf. Comput. Commun. Security, Vienna, Austria, Oct. 2016, pp. 308–318.