Personalization Improves Privacy–Accuracy Tradeoffs in Federated Learning

Alberto Bietti (NYU) Chen-Yu Wei (USC) Miro Dudík (MSR)
John Langford (MSR) Zhiwei Steven Wu (CMU)

ICML 2022, Baltimore.

Personalized Federated Learning

$$\min_{w,\theta_{1:N}} \left\{ f(w, \theta_{1:N}) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\xi \sim P_i}[f_i(w, \theta_i, \xi)] \right\},$$

- N: number of users/clients
- P_i: data distribution of user i
- w: **global** (shared) parameters
- θ_i: **local** parameters for user i

Personalized Federated Learning

$$\min_{w,\theta_{1:N}} \left\{ f(w, \theta_{1:N}) := \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\xi \sim P_i}[f_i(w, \theta_i, \xi)] \right\},$$

- N: number of users/clients
- P_i: data distribution of user i
- w: **global** (shared) parameters
- θ_i: **local** parameters for user i

Federated Learning

- Decentralized data $\rightarrow$ federated optimization algorithms (Wang et al., 2021)
- Privacy leaks on w $\rightarrow$ **user-level** (joint) differential privacy (Kearns et al., 2014)

Local vs Global Learning

Local learning

$$\min_{\theta_i} \mathbb{E}_{(x,y)\sim P_i}[\ell(y, \theta_i{}^\top x)]$$

Global learning

$$\min_{w} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{(x,y)\sim P_i}[\ell(y, w^\top x)]$$

Local vs Global Learning

Local learning

$$\min_{\theta_i} \mathbb{E}_{(x,y)\sim P_i}[\ell(y, \theta_i{}^\top x)]$$

- Personalized models
- Perfectly private!
- Statistically inefficient

Global learning

$$\min_{w} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{(x,y)\sim P_i}[\ell(y, w^\top x)]$$

Local vs Global Learning

Local learning

$$\min_{\theta_i} \mathbb{E}_{(x,y)\sim P_i}[\ell(y, \theta_i{}^\top x)]$$

- Personalized models
- Perfectly private!
- Statistically inefficient

Global learning

$$\min_{w} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{(x,y)\sim P_i}[\ell(y, w^\top x)]$$

- No personalization
- Cost of privacy
- Statistical gains (N times more samples)

Local vs Global Learning

Local learning

$$\min_{\theta_i} \mathbb{E}_{(x,y)\sim P_i}[\ell(y, \theta_i{}^\top x)]$$

- Personalized models
- Perfectly private!
- Statistically inefficient

Global learning

$$\min_{w} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{(x,y)\sim P_i}[\ell(y, w^\top x)]$$

- No personalization
- Cost of privacy
- Statistical gains (N times more samples)

Q: How does personalization affect this tradeoff?

Main Algorithm

- Federated SGD on personalized models
- Vary personalization level through **step-size ratio** α
- DP updates on w through clipping + noise injection

Main Algorithm

- Federated SGD on personalized models
- Vary personalization level through **step-size ratio** α
- DP updates on w through clipping + noise injection

Algorithm: (n: rounds = samples per user)

- For $t = 1, \dots, n$
  - For all clients i in parallel
    - ⋆ Sample data $\xi_{i,t} \sim P_i$
    - ⋆ Compute $g^t_{\theta,i} = \nabla_\theta f_i(w_{t-1}, \theta_{i,t-1}, \xi_{i,t})$
      $g^t_{w,i} = \nabla_w f_i(w_{t-1}, \theta_{i,t-1}, \xi_{i,t})$
    - ⋆ Update $\theta_{i,t} = \theta_{i,t-1} - \frac{\eta}{N} g^t_{\theta,i}$ (**local update**)
  - Sample $\zeta_t \sim \mathcal{N}(0, \sigma^2_\zeta I)$
  - Update $w_t = w_{t-1} - \alpha\eta(\frac{1}{N}\sum_{i=1}^N \text{clip}(g^t_{w,i}) + \zeta_t)$ (**global update**)

Generalization guarantee

- Define α-norm of $z = (w, \theta_{1:N})$ which controls inductive bias:

$$\|z\|_\alpha^2 := \frac{1}{\alpha}\|w\|^2 + \|\theta_{1:N}\|^2$$

- f_i jointly convex, L-smooth in (w, θ_i), G-bounded gradients, gradient variances $\sigma_w^2, \sigma_\theta^2$
- Set privacy noise ζ_t such that the algorithm satisfies (ϵ, δ) (joint) DP

Generalization guarantee

- Define α-norm of $z = (w, \theta_{1:N})$ which controls inductive bias:

$$\|z\|_\alpha^2 := \frac{1}{\alpha}\|w\|^2 + \|\theta_{1:N}\|^2$$

- f_i jointly convex, L-smooth in (w, θ_i), G-bounded gradients, gradient variances $\sigma_w^2, \sigma_\theta^2$
- Set privacy noise ζ_t such that the algorithm satisfies (ϵ, δ) (joint) DP

Theorem (Generalization)

Let z^* be a minimizer of f, and $\bar{z}_n = \frac{1}{n}\sum_{t=1}^n z_t$. After n rounds/samples, we have

$$\mathbb{E}[f(\bar{z}_n) - f(z^*)] \lesssim \underbrace{\frac{L\max(\alpha, \frac{1}{N})\|z^*\|_\alpha^2}{n}}_{\text{bias}} + \underbrace{\|z^*\|_\alpha\sqrt{\frac{\alpha\sigma_w^2 + \sigma_\theta^2}{Nn}}}_{\text{variance}} + \underbrace{\|z^*\|_\alpha\sqrt{\frac{\alpha d_w G^2 \log(\frac{1}{\delta})}{N^2\epsilon^2}}}_{\text{privacy cost}}$$

Example: additive model

$$\min_{w,\theta_{1:N}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{(x,y)\sim P_i}[\ell(y, (w + \theta_i)^\top x)]$$

- Assume global-only minimizer v^* exists: $v^* \in \arg\min_v \mathbb{E}_{(x,y)\sim P_i}[\ell(y, v^\top x)]$

Example: additive model

$$\min_{w,\theta_{1:N}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{(x,y)\sim P_i}[\ell(y, (w + \theta_i)^\top x)]$$

- Assume global-only minimizer v^* exists: $v^* \in \arg\min_v \mathbb{E}_{(x,y)\sim P_i}[\ell(y, v^\top x)]$
- **Local learning** guarantee ($\alpha \to 0$), $z^* = (0, (v^*, \ldots, v^*))$: (ignoring the bias term)

$$\mathbb{E}[f(\bar{z}_n) - f(z^*)] \lesssim \|v^*\| \sqrt{\frac{\sigma_\theta^2}{n}}$$

Example: additive model

$$\min_{w,\theta_{1:N}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{(x,y)\sim P_i}[\ell(y, (w+\theta_i)^\top x)]$$

- Assume global-only minimizer v^* exists: $v^* \in \arg\min_v \mathbb{E}_{(x,y)\sim P_i}[\ell(y, v^\top x)]$
- **Local learning** guarantee ($\alpha \to 0$), $z^* = (0, (v^*, \ldots, v^*))$: (ignoring the bias term)

$$\mathbb{E}[f(\bar{z}_n) - f(z^*)] \lesssim \|v^*\| \sqrt{\frac{\sigma_\theta^2}{n}}$$

- **Global learning** guarantee ($\alpha \to \infty$), $z^* = (v^*, 0)$:

$$\mathbb{E}[f(\bar{z}_n) - f(z^*)] \lesssim \|v^*\| \sqrt{\frac{\sigma_w^2}{Nn}} + \|v^*\| \sqrt{\frac{d_w G^2 \log(\frac{1}{\delta})}{N^2 \epsilon^2}}$$

Example: additive model

$$\min_{w,\theta_{1:N}} \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{(x,y)\sim P_i}[\ell(y, (w + \theta_i)^\top x)]$$

- Assume global-only minimizer v^* exists: $v^* \in \arg\min_v \mathbb{E}_{(x,y)\sim P_i}[\ell(y, v^\top x)]$
- **Local learning** guarantee ($\alpha \to 0$), $z^* = (0, (v^*, \ldots, v^*))$: (ignoring the bias term)

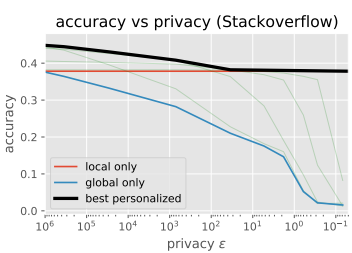
$$\mathbb{E}[f(\bar{z}_n) - f(z^*)] \lesssim \|v^*\| \sqrt{\frac{\sigma_\theta^2}{n}}$$

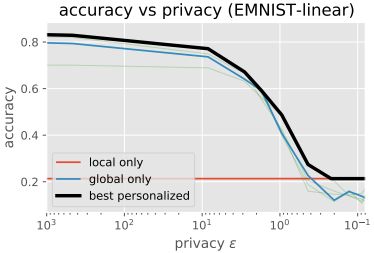
- **Global learning** guarantee ($\alpha \to \infty$), $z^* = (v^*, 0)$:

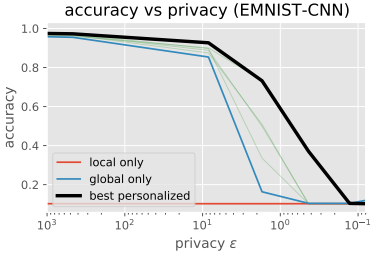
$$\mathbb{E}[f(\bar{z}_n) - f(z^*)] \lesssim \|v^*\| \sqrt{\frac{\sigma_w^2}{Nn}} + \|v^*\| \sqrt{\frac{d_w G^2 \log(\frac{1}{\delta})}{N^2 \epsilon^2}}$$

- Decreasing α helps when n gets larger, or when more personalization is useful

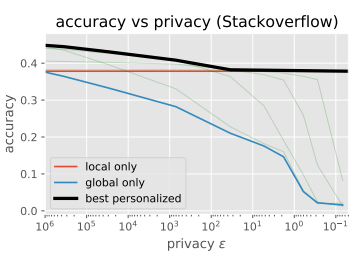
Experiments

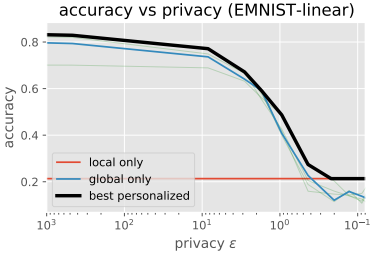


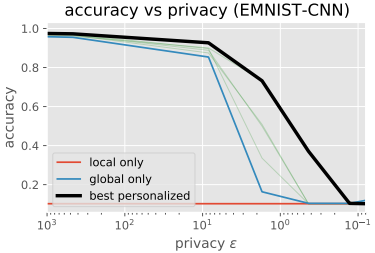




Experiments







Thank you!

References I

M. Kearns, M. M. Pai, A. Roth, and J. Ullman. Mechanism design in large games: Incentives and privacy. *American Economic Review*, 104(5):431–35, May 2014.

J. Wang, Z. Charles, Z. Xu, G. Joshi, H. B. McMahan, M. Al-Shedivat, G. Andrew, S. Avestimehr, K. Daly, D. Data, et al. A field guide to federated optimization. *arXiv preprint arXiv:2107.06917*, 2021.