Adaptive Latent-Space Constraints in Personalized Federated Learning

Sana Ayromlou[1,2,*] D. B. Emerson[1]

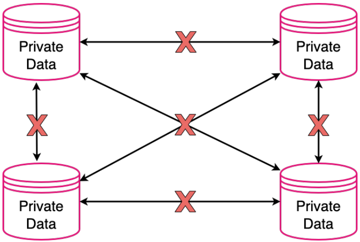
[1]Vector Institute, Toronto, Ontario, CA

[2]Google, Toronto, Ontario, CA

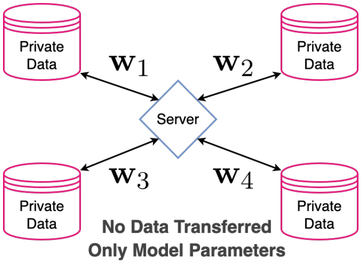
[*]Work done while at the Vector Institute.

Federated Learning: Training on Distributed Data

ML models are most commonly trained on a centralized pool of data. Federated Learning (FL) is used to train models on decentralized data.



Data transfer is discouraged or impossible.



Data remains in place, while model weights are communicated with a trusted entity.

VECTOR INSTITUTE | INSTITUT VECTEUR

Horizontal, Cross-Silo, Homogenous Model PFL

There are many FL settings, each of which may require unique approaches to facilitate distributed deep learning.

Horizontal FL: **Feature and label spaces are shared** between clients. Primary benefit is access to more training data.

Cross-Silo: Small to medium pool of clients, large compute resources, reliable training participation.

Homogenous Models: Each client is training the **same model architecture**.

Personalized FL: Each client trains a **unique set of model parameters**, aiming to overcome heterogeneity in distributed datasets.[1]

[1] Li et al., "Federated Learning: Challenges, Methods, and Future Directions"

Globally Constrained Local Model Training: Ditto[3] and Beyond

Algorithm 1: Ditto with FedAvg aggregation and batch SGD.

Input: N, T, s, λ, η, $\bar{w}$.
Set $w_L^{(i)} = \bar{w}$ for each client i.
for $t = 0, \ldots, T-1$ **do**
  for *each client i in parallel* **do**
    Set $w_G^{(i)} = \bar{w}$.
    for *s iterations, draw batch b* **do**
      $w_G^{(i)} = w_G^{(i)} - \eta \nabla \ell_i \left(b; w_G^{(i)} \right)$.
      $w_L^{(i)} = w_L^{(i)} - \eta \nabla \left(\ell_i \left(b; w_L^{(i)} \right) + \frac{\lambda}{2} \| w_L^{(i)} - \bar{w} \|_2^2 \right)$.
    end
    Send $w_G^{(i)}$ to server for aggregation.
  end
  $\bar{w} = \frac{1}{n} \sum_{i=1}^{N} n_i \cdot w_G^{(i)}$.
end

Target model at the end of training is parameterized by w_L.

Local model training is constrained to not "drift" too far from an averaged global model.

Ditto is a state-of-the-art algorithm for many FL settings with data heterogeneity.[2]

[2] Matsuda et al., "Benchmark for Personalized Federated Learning"
[3] Li et al., "Ditto: Fair and Robust Federated Learning Through Personalization"

Globally Constrained Local Model Training: Ditto[3] and Beyond

Algorithm 1: Ditto with FedAvg aggregation and batch SGD.

Input: N, T, s, λ, η, $\bar{w}$.
Set $w_L^{(i)} = \bar{w}$ for each client i.
for $t = 0, \ldots, T-1$ **do**
  for *each client i in parallel* **do**
    Set $w_G^{(i)} = \bar{w}$.
    for *s iterations, draw batch b* **do**
      $w_G^{(i)} = w_G^{(i)} - \eta \nabla \ell_i \left(b; w_G^{(i)} \right)$.
      $w_L^{(i)} = w_L^{(i)} - \eta \nabla \left(\ell_i \left(b; w_L^{(i)} \right) + \frac{\lambda}{2} \| w_L^{(i)} - \bar{w} \|_2^2 \right)$.
    end
    Send $w_G^{(i)}$ to server for aggregation.
  end
  $\bar{w} = \frac{1}{n} \sum_{i=1}^{N} n_i \cdot w_G^{(i)}$.
end

This measure is static and does not consider specific properties of training data.

Replace or augment this penalty with an adaptive measure targeting a different kind of drift.

Define a strong distance measure on model latent spaces: $d(f(x; \theta_L); f(x; \bar{\theta}))$.

VECTOR INSTITUTE | INSTITUT VECTEUR

[3] Li et al., "Ditto: Fair and Robust Federated Learning Through Personalization"

Adaptable Latent-Space Measures

Let $X \subset \mathbb{R}^m$ represent a model latent space and P and Q be probability measures on X induced by distinct feature maps.

Consider two maximum mean discrepancy (MMD) measures that can be optimized to tell P and Q apart.

Adaptable Latent-Space Measures[4]

Let $X \subset \mathbb{R}^m$ represent a model latent space and P and Q be probability measures on X induced by distinct feature maps.

Consider two maximum mean discrepancy (MMD) measures that can be optimized to tell P and Q apart.

$$\text{MK-MMD}^2(P, Q; \mathcal{H}_k) = \sum_{j=1}^{d} \beta_j \text{MMD}^2(P, Q; \mathcal{H}_{k_j}),$$

where $k_j(x, y) = e^{-\gamma_j \|x-y\|_2^2}$ for a set of $\{\gamma_j\}_{j=1}^d$.

$$\beta_* = \underset{\substack{\sum_{j=1}^d \beta_j = 1 \\ \beta \geq \mathbf{0}}}{\arg\max} \frac{\text{MK-MMD}^2(P, Q; \mathcal{H}_k)}{\sigma(P, Q, \mathcal{H}_k)}.$$

[4]Gretton et al., "Optimal kernel choice for large-scale two-sample tests"

Adaptable Latent-Space Measures[5]

Let $X \subset \mathbb{R}^m$ represent a model latent space and P and Q be probability measures on X induced by distinct feature maps.

Consider two maximum mean discrepancy (MMD) measures that can be optimized to tell P and Q apart.

Define a featurization network, $\varphi(\cdot;\omega)$, parameterized by ω and deep kernel

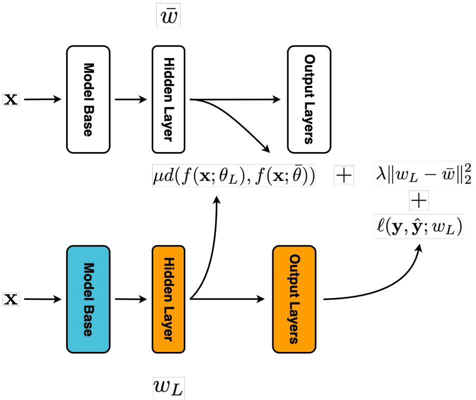
$$k_\omega(x,y) = (1-\epsilon)k(\varphi(x;\omega),\varphi(y;\omega)) + \epsilon q(x,y),$$

where $k(x,y) = e^{-\gamma_k \|x-y\|_2^2}$, $q(x,y) = e^{-\gamma_q \|x-y\|_2^2}$.

$$\max_{\omega,\epsilon,\gamma_k,\gamma_q} \frac{\text{MMD-D}^2(P,Q;\mathcal{H}_{k_\omega})}{\sigma(P,Q;\mathcal{H}_{k_\omega})}.$$

[5] Liu et al., "Learning deep kernels for non-parametric two-sample tests"

Numerical Setup



$d(\cdot,\cdot)$ is either MK-MMD or MMD-D, acting on the latent spaces of the local and frozen global models.

MK-MMD or MMD-D measures are re-optimized periodically using training data, adapting to changing feature maps.

Weights $\mu \geq 0$ and $\lambda \geq 0$ balance the MMD and standard Ditto constraints.

Experimental Results: Ditto

Dataset	FedAvg	Ditto	Without Ditto ($\lambda = 0$)		With Ditto ($\lambda > 0$)	
			MMD-D	MK-MMD	MMD-D	MK-MMD
$\text{Synthetic}_{0.0}$[6]	84.733	89.129	**90.237***	90.066*	89.458*	89.258*
$\text{Synthetic}_{0.5}$[6]	85.458	85.533	**91.270***	90.262*	89.695*	88.104*
RxRx1[7]	35.207	65.629	67.478*	67.078*	**67.755***	66.892*
$\text{CIFAR-10}_{0.1}$	71.220	84.930	83.789	84.439	**85.214***	84.900
$\text{CIFAR-10}_{0.5}$	75.575	80.702	75.094	76.564	80.696	**80.976***
$\text{CIFAR-10}_{5.0}$	77.284	77.658	67.729	68.832	**77.739***	**77.739***
Fed-ISIC2019[8]	64.057	71.350	64.302	62.677	**72.226***	71.267

[6] Li et al., "Federated Optimization in Heterogeneous Networks"
[7] Sypetkowski et al., "RxRx1: A Dataset for Evaluating Experimental Batch Correction Methods"
[8] Terrail et al., "FLamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings"

Experimental Results: MR-MTL

Dataset	MR-MTL	+MK-MMD	+MMD-D
$\text{Synthetic}_{0.0}$	90.879	90.708	**91.142***
$\text{Synthetic}_{0.5}$	86.750	**90.958***	90.337*
RxRx1	64.065	65.791*	**66.673***
$\text{CIFAR-10}_{0.1}$	79.516	**81.269***	80.307*
$\text{CIFAR-10}_{0.5}$	73.361	74.333*	**74.446***
$\text{CIFAR-10}_{5.0}$	68.224	69.487*	**70.353***
Fed-ISIC2019	**70.628**	68.180	70.366

Thank you for listening!

- Gretton, Arthur et al. "Optimal kernel choice for large-scale two-sample tests". In: *Advances in neural information processing systems* 25 (2012).
- Li, T. et al. "Ditto: Fair and Robust Federated Learning Through Personalization". In: *Proceedings of the 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Proceedings of Machine Learning Research. PMLR, 2021, pp. 6357–6368. URL: https://proceedings.mlr.press/v139/li21h.html.
- Li, Tian et al. "Federated Learning: Challenges, Methods, and Future Directions". In: *IEEE Signal Processing Magazine* 37.3 (2020), pp. 50–60.
- Li, Tian et al. "Federated Optimization in Heterogeneous Networks". In: *Proceedings of Machine Learning and Systems*. Ed. by I. Dhillon, D. Papailiopoulos, and V. Sze. Vol. 2. 2020, pp. 429–450. URL: https://proceedings.mlsys.org/paper/2020/file/38af86134b65d0f10fe33d30dd76442e-Paper.pdf.
- Liu, Feng et al. "Learning deep kernels for non-parametric two-sample tests". In: *International conference on machine learning*. PMLR. 2020, pp. 6316–6326.

VECTOR INSTITUTE | INSTITUT VECTEUR

- Matsuda, K. et al. "Benchmark for Personalized Federated Learning". In: *IEEE Open Journal of the Computer Society* 5.01 (2024), pp. 2–13. ISSN: 2644-1268. DOI: 10.1109/OJCS.2023.3332351.
- Sutherland, Danica J. et al. "Generative Models and Model Criticism via Optimized Maximum Mean Discrepancy". In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL: https://openreview.net/forum?id=HJWHIKqgl.
- Sypetkowski, Maciej et al. "RxRx1: A Dataset for Evaluating Experimental Batch Correction Methods". In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2023, pp. 4285–4294.
- Terrail, Jean Ogier Du et al. "FLamby: Datasets and Benchmarks for Cross-Silo Federated Learning in Realistic Healthcare Settings". In: *NeurIPS 2022 - Thirty-sixth Conference on Neural Information Processing Systems*. Proceedings of NeurIPS. New Orleans, United States, Nov. 2022. URL: https://hal.science/hal-03900026.

VECTOR INSTITUTE | INSTITUT VECTEUR