Appendix D. Experiments

In this section we describe in more detail the experimental setup and present additional results referenced in the main manuscript. All dataset processing, models and hyperparameters are available at github.com/annoymous-submissions/ot_cost.

D.1. Setup: Datasets, Training, and models

Table D.1 describes how we partitioned datasets to produce datasets with specific dataset dissimilarity scores. We employ a mixture of feature and label skew **??**.

- **Synthetic:** We draw samples from 2 distinct distributions, $\mathcal{D}_a$ and $\mathcal{D}_b$. For partition 1 we always draw samples from $\mathcal{D}_a$ and for partition 2 we vary the proportion of samples drawn from $\mathcal{D}_a$ and $\mathcal{D}_b$.
- **Credit:** We introduce increasing amounts of label skew and feature noise to the datasets. Label skew is achieved by introducing label distribution bias. For example, a 10% bias means 10% more positive labels and negative labels in partition 1 and 2, respectively. Feature noise is achieved by adding Gaussian noise.
- **Weather:** We partition datasets based on climate as described by **?**. This introduces real-world feature and label skew.
- **EMNIST:** We partition datasets based on labels.
- **CIFAR:** We partition the dataset based on labels.
- **IXITiny:** We partition based on site. This is a natural partition. Sites Guys and Hammersmith use a Philips system with similar settings. Site IOP uses a GE system with unknown settings (see **??**).
- **ISIC2019:** We partition based on site. This is a natural paritition.

Dataset	OT score	Partition 1	Partition 2
Synthetic	0.00*	100% $\mathcal{D}_a$	Identical to Partition 1
	0.01*	100% $\mathcal{D}_a$	50% subset of Partition 1
	0.03	100% $\mathcal{D}_a$	100% $\mathcal{D}_a$
	0.1	100% $\mathcal{D}_a$	85% $\mathcal{D}_a$, 15% $\mathcal{D}_b$
	0.2	100% $\mathcal{D}_a$	65% $\mathcal{D}_a$, 35% $\mathcal{D}_b$
	0.3	100% $\mathcal{D}_a$	43% $\mathcal{D}_a$, 57% $\mathcal{D}_b$
	0.4	100% $\mathcal{D}_a$	25% $\mathcal{D}_a$, 75% $\mathcal{D}_b$
	0.5	100% $\mathcal{D}_a$	11% $\mathcal{D}_a$, 89% $\mathcal{D}_b$
Credit	0.12	No feature noise or label bias	
	0.23	No feature noise and 30% label bias	
	0.3	feature noise $\sim N(0.5, 0.5)$ and 45% label bias	
	0.4	feature noise $\sim N(1, 0.5)$ and 45% label bias	
Weather	0.11	Tropical, Mild temperate	
	0.19	Tropical, Mild temperate	Dry, Mild temperate
	0.30	Tropical, Mild temperate	Dry
	0.40	Tropical, Mild temperate	Dry, Snow
	0.48	Tropical, Mild temperate	Snow
EMNIST	0.11	1-10	
	0.19	1-10, 12,18,24,28	1-10, 38,44,50,54
	0.25	1-10, 11,12,13,14,16,18,24,28	1-10, 37,38,39,40,42,44,50,54
	0.34	1-10, 11-25	1-10, 36-41
	0.39	1-10, 11-35	1-10, 36-61
CIFAR	0.08	1-10	
	0.21	11,98,29,73, 78, 49, 97, 51, 55, 92	11,98,29,73, 78, 49, 42, 83, 72, 82
	0.30	11,50,78,1,92, 78, 49, 97, 55, 16, 14	11, 36, 29, 73, 82, 78, 49, 42, 12, 23, 51
	0.38	11,50,78,8,92,2,49,98,89,3	17, 36, 30, 73, 83,28, 34, 42, 10, 20
IXITiny	0.08	Guys	Hammersmith
	0.28	Guys	IOP
	0.30	Hammersmith	IOP
ISIC2019	0.06	ViDIR Group, Vienna (FOTO)	ViDIR Group, Vienna (FOTO)
	0.15	ViDIR Group, Vienna (FOTO)	Hospital Clínic de Barcelona
	0.19	ViDIR Group, Vienna (FOTO)	ViDIR Group, Vienna (Dermaphot)
	0.25	ViDIR Group, Vienna (FOTO)	ViDIR Group, Vienna (MoleMax)
	0.3	ViDIR Group, Vienna (MoleMax)	Cliff Rosendahl, Australia

TABLE D.1
Creation of partitioned datasets

* Note these examples are not assessed during model training and are for illustrative purposes only.

A. *Setup: Training and Models*

For each combination of dataset, cost, and training, we performed learning rate and optimizer tuning via grid search. We evaluate five learning rates ($5e^{-1}, 1e^{-1}, 5e^{-2}, 1e^{-2}, 5e^{-3}, 5e^{-4}$) and two optimizers (ADAM and SGD). The SGD optimizer is exclusively used for federated algorithms. For pFedME we use the pFedME optimizer released by the authors. For CIFAR, IXITiny and ISIC2019 we utilize pre-trained models that we fine-tune. Below we describe the model, learning rates, optimizers, and regularization parameters (for pFedME and Ditto only) for each dataset.

a) Synthetic: We use a MLP with 2 hidden layers of size 18 and 6.

TABLE D.2

OT score	Single	Joint	FedAvg	pFedME	Ditto
0.03	$5e^{-3}$, ADM	$5e^{-2}$, ADM	$5e^{-2}$, ADM	$5e^{-2}$, pFedME-opt, $5e^{-1}$	$1e^{-1}$, ADM, $5e^{-1}$
0.10	$5e^{-3}$, ADM	$5e^{-2}$, ADM	$1e^{-1}$, ADM	$1e^{-1}$, pFedME-opt, $5e^{-1}$	$1e^{-1}$, ADM, $1e^{0}$
0.20	$1e^{-2}$, ADM	$1e^{-1}$, ADM	$1e^{-1}$, ADM	$1e^{-1}$, pFedME-opt, $1e^{-1}$	$1e^{-1}$, ADM, $1e^{-1}$
0.30	$5e^{-2}$, ADM	$5e^{-3}$, ADM	$5e^{-2}$, SGD	$5e^{-2}$, pFedME-opt, $1e^{-2}$	$1e^{-1}$, SGD, $1e^{-2}$
0.40	$5e^{-2}$, ADM	$1e^{-2}$, ADM	$5e^{-2}$, SGD	$5e^{-2}$,pFedME-opt, $1e^{-2}$	$1e^{-1}$, SGD, $1e^{-2}$
0.50	$5e^{-2}$, ADM	$3e^{-2}$, ADM	$5e^{-2}$, SGD	$1e^{-1}$, pFedME-opt, $1e^{-3}$	$1e^{-1}$, SGD, $1e^{-3}$

b) Credit: We use a MLP with 3 hidden layers of size 56, 56 and 28.

TABLE D.3

OT score	Single	Joint	FedAvg	pFedME	Ditto
0.12	$5e^{-3}$, ADM	$5e^{-2}$, ADM	$5e^{-2}$, ADM	$1e^{-1}$, pFedME-opt, $5e^{-1}$	$1e^{-1}$, ADM, $1e^{0}$
0.23	$5e^{-3}$, ADM	$1e^{-2}$, ADM	$1e^{-2}$, ADM	$1e^{-1}$, pFedME-opt, $1e^{-1}$	$1e^{-1}$, ADM, $1e^{-1}$
0.30	$1e^{-2}$, ADM	$1e^{-2}$, ADM	$1e^{-2}$, ADM	$1e^{-1}$, pFedME-opt, $1e^{-2}$	$1e^{-1}$, ADM, $1e^{-3}$
0.40	$5e^{-2}$, ADM	$1e^{-3}$, ADM	$5e^{-3}$, ADM	$1e^{-1}$, pFedME-opt, $1e^{-2}$	$1e^{-1}$, ADM, $1e^{-3}$

c) Weather: We use a MLP with 3 hidden layers of size 123, 123 and 50.

TABLE D.4

OT score	Single	Joint	FedAvg	pFedME	Ditto
0.11	$3e^{-3}$, ADM	$1e^{-2}$, ADM	$5e^{-2}$, SGD	$1e^{-1}$, pFedME-opt, $5e^{-1}$	$1e^{-1}$, ADM, $1e^{0}$
0.19	$5e^{-3}$, ADM	$5e^{-2}$, ADM	$1e^{-1}$, SGD	$1e^{-1}$, pFedME-opt, $5e^{-1}$	$1e^{-1}$, ADM, $1e^{0}$
0.30	$1e^{-2}$, ADM	$1e^{-2}$, ADM	$5e^{-2}$, SGD	$1e^{-1}$, pFedME-opt, $1e^{-1}$	$1e^{-1}$, ADM, $5e^{-1}$
0.40	$1e^{-2}$, ADM	$1e^{-3}$, ADM	$2e^{-2}$, SGD	$1e^{-1}$, pFedME-opt, $1e^{-2}$	$1e^{-1}$, ADM, $1e^{-2}$
0.48	$1e^{-2}$, ADM	$1e^{-3}$, ADM	$5e^{-2}$, SGD	$1e^{-1}$, pFedME-opt, $1e^{-3}$	$1e^{-1}$, ADM, $1e^{-2}$

d) EMNIST: We use a LeNet-5 CNN.

TABLE D.5

OT score	Single	Joint	FedAvg	pFedME	Ditto
0.11	$5e^{-2}$, ADM	$1e^{-2}$, ADM	$5e^{-2}$, ADM	$5e^{-2}$, pFedME-opt, $1e^{-1}$	$5e^{-2}$, ADM, $5e^{-1}$
0.19	$5e^{-3}$, ADM	$5e^{-3}$, ADM	$5e^{-2}$, ADM	$5e^{-2}$, pFedME-opt, $1e^{-1}$	$5e^{-2}$, ADM, $1e^{-1}$
0.25	$1e^{-2}$, ADM	$5e^{-3}$, ADM	$5e^{-2}$, ADM	$5e^{-2}$, pFedME-opt, $1e^{-2}$	$1e^{-1}$, ADM, $1e^{-2}$
0.34	$1e^{-2}$, ADM	$1e^{-2}$, ADM	$1e^{-2}$, ADM	$5e^{-2}$, pFedME-opt, $1e^{-2}$	$1e^{-1}$, ADM, $1e^{-3}$
0.39	$1e^{-2}$, ADM	$5e^{-3}$, ADM	$3e^{-2}$, ADM	$1e^{-1}$, pFedME-opt, $1e^{-3}$	$5e^{-2}$, ADM, $1e^{-3}$

e) CIFAR: We use a pre-trained ResNet-18, freezing layers 1-3.

TABLE D.6

OT score	Single	Joint	FedAvg	pFedME	Ditto
0.08	$1e^{-3}$, ADM	$5e^{-4}$, ADM	$1e^{-2}$, ADM	$1e^{-2}$, pFedME-opt, $5e^{-1}$	$5e^{-3}$, ADM, $1e^{0}$
0.21	$1e^{-3}$, ADM	$5e^{-3}$, ADM	$1e^{-2}$, SGD	$5e^{-2}$, pFedME-opt, $1e^{-1}$	$1e^{-2}$, ADM, $5e^{-1}$
0.30	$5e^{-4}$, ADM	$5e^{-2}$, ADM	$1e^{-2}$, SGD	$5e^{-2}$, pFedME-opt, $1e^{-2}$	$1e^{-2}$, ADM, $1e^{-2}$
0.38	$5e^{-4}$, ADM	$5e^{-3}$, ADM	$5e^{-2}$, SGD	$5e^{-2}$, pFedME-opt, $1e^{-3}$	$5e^{-3}$, ADM, $1e^{-2}$

f) IXITiny: We use a pre-trained 3D Unet and allow training of all layers.

TABLE D.7

OT score	Single	Joint	FedAvg	pFedME	Ditto
0.08	$1e^{-1}$, ADM	$1e^{-1}$, ADM	$5e^{-2}$, ADM	$1e^{-1}$, pFedME-opt, $5e^{-1}$	$1e^{-1}$, ADM, $1e^{-1}$
0.28	$1e^{-1}$, ADM	$1e^{-1}$, ADM	$1e^{-1}$, SGD	$5e^{-2}$, pFedME-opt, $1e^{-2}$	$1e^{-1}$, SGD, $1e^{-2}$
0.30	$1e^{-1}$, ADM	$1e^{-2}$, ADM	$1e^{-1}$, SGD	$5e^{-2}$, pFedME-opt, $1e^{-3}$	$1e^{-3}$, SGD, $1e^{-2}$

g) ISIC2019: We use a pre-trained efficientNet and allow training of all layers.

TABLE D.8

OT score	**Single**	**Joint**	**FedAvg**	**pFedME**	**Ditto**
0.06	$5e^{-4}$, ADM	$5e^{-3}$, ADM	$1e^{-1}$, ADM	$5e^{-3}$, pFedME-opt, $5e^{-1}$	$1e^{-2}$, ADM, $5e^{-1}$
0.15	$5e^{-3}$, ADM	$5e^{-3}$, ADM	$1e^{-2}$, ADM	$5e^{-3}$, pFedME-opt, $1e^{-1}$	$1e^{-1}$, ADM, $1e^{-1}$
0.19	$5e^{-3}$, ADM	$5e^{-3}$, ADM	$1e^{-2}$, ADM	$5e^{-3}$, pFedME-opt, $1e^{-1}$	$1e^{-1}$, ADM, $1e^{-1}$
0.25	$5e^{-3}$, ADM	$1e^{-2}$, ADM	$1e^{-2}$, ADM	$1e^{-2}$, pFedME-opt, $1e^{-2}$	$1e^{-2}$, ADM, $1e^{-2}$
0.3	$5e^{-3}$, ADM	$5e^{-3}$, ADM	$5e^{-2}$, ADM	$1e^{-2}$, pFedME-opt, $1e^{-3}$	$1e^{-2}$, ADM, $1e^{-2}$

*ADM = Adam optimizer, SGD = SGD optimizer, pFedME-opt = pFedME optimizer