CAN FAIR FEDERATED LEARNING REDUCE THE NEED FOR PERSONALIZATION?

Anonymous authors
Paper under double-blind review

ABSTRACT

Federated Learning (FL) allows edge devices to collaboratively train machine learning models without sharing local data. Since the data distribution varies across client partitions, the performance of the federated model on local data also varies. To solve this, *fair* FL approaches attempt to reduce the accuracy disparity between local partitions by focusing on clients with larger losses during training; while *local adaptation* personalizes the federated model by re-training it on local data providing a device participation incentive in cases where a federated model underperforms *relatively* to one trained locally—their accuracy difference is less than zero. This paper evaluates Q-Fair Federated Learning (Q-FFL) in this relative domain and determines whether it provides a better starting point for personalization or supplants it. Contrary to expectation, Q-FFL does not significantly reduce the number of underperforming clients in a language task while doubling them in an image recognition task. Furthermore, fairness levels which maintain average accuracy provide no benefit to relative accuracy in federated or adapted models. We postulate that Q-FFL is unsuitable for our goal since clients with highly accurate local models require the federated model to have a disproportionate local partition accuracy to receive a benefit. Instead, we propose using knowledge distillation during FL training to create models with a higher local accuracy floor without forfeiting the ceiling. Our preliminary evaluation shows a 50% reduction in underperforming clients in the language task with no increase in underperforming clients for the image task. Thus, we argue that this simple change represents a more promising avenue for reducing the need for personalization than fairness.

1 INTRODUCTION

Edge devices represent a pool of both computational power and data for ML tasks. In order to access this pool in a manner which minimizes communication costs while preserving privacy, McMahan et al. (2017) introduced Federated Learning (FL). Their FL algorithm, FedAvg, achieved these aims by training ML models directly on client devices and then performing aggregation over model weights, thus bypassing the need to share training data. As the data distribution may differ considerably across clients, emphasizing global model convergence leads to a trade-off between federated performance and performance on client data partitions. In some cases, a federated model may have lower accuracy on a client than a local model trained from scratch. This eliminates one of the main incentives for the client to participate in the FL process.

The existing body of work on balancing the trade-off between federated and local performance has focused on two primary means of improving accuracy on heterogeneous data distributions. Li et al. (2019a) propose Q-Fair Federated Learning (Q-FFL) as a branch of FL which tunes the aggregate objective function to focus on clients with large losses when creating a *single* shared global model. This is done in an attempt to smooth-out the accuracy distribution across client data partitions. Alternatively, *local adaptation* (personalization) methods such as Freezebase (FB), Multi-task Learning (MTL) and Knowledge Distillation (KD) have been recommended by Yu et al. (2020) and Mansour et al. (2020) as a means of quickly constructing effective local models by using the federated model as a starting point. Since local adaptation requires further on-device training, the natural baseline of comparison is a fully local model trained on the specific client for a comparable amount of time. On the other hand, Li et al. (2019a) evaluates Q-FFL in terms of the *absolute* accuracy distribution of the federated model over local data partitions, without regard to the potential performance of a fully

local model for a client. For the rest of this work, *relative accuracy* shall refer to the difference in achieved accuracy between a federated and local model on a client test set.

While the sets of potential use cases for these techniques are not identical—for instance, personalization would be inappropriate when clients do not have sufficient data for re-training—, it is possible that using Q-FFL may result in a fairer relative accuracy distribution without significantly hurting average performance. Locally adapting on a client requires sufficiently powerful hardware to retrain it, high-quality data, and a large amount of dedicated training time. If Q-FFL can reduce the need for personalization by addressing the same concerns, it may be considered preferable for a federated network whose clients do not fulfil the aforementioned requirements. For this to be the case, Q-FFL would need to lower the number of underperforming clients or improve their average relative accuracy enough so that they require *less* adaptation. This is not what we observe in practice, as our experiments show Q-FFL to have either a negative impact on relative accuracy or none at all.

Our contribution is threefold. First, we construct an initial empirical evaluation of the relative accuracy distribution of models trained with Q-FFL on the Reddit, CIFAR-10, and FEMNIST datasets for either next word prediction or image recognition. We then compare the effects of Q-FFL on the number of underperforming clients to a FedAvg baseline. During our evaluation, we show that Q-FFL does not significantly improve the relative accuracy distribution for next word prediction and brings little discernible benefit over a combination of FedAvg and local adaptation. Concerningly, it doubled the number of underperforming clients for image recognition on FEMNIST.

Second, we investigate any potential synergies between Q-FFL and personalization. We adapt fair federated models and show that the adapted models do not outperform those initially trained with FedAvg on the language task in terms of either average accuracy or the number of underperforming clients. Effects for the image recognition task are unclear for both datasets as the potential effects are too small to evaluate clearly.

Third, we propose using Knowledge Distillation (Hinton et al., 2015) during local training after the halfway point of the FL process as an alternative to Q-FFL. Preliminary experimental results on the language task show a significant reduction in the number of underperforming clients when compared to Q-FFL without any downsides to subsequent local adaptation. This technique also avoids the increase in the number of underperforming clients observed for image recognition on FEMNIST. We speculate that Knowledge Distillation outperforms the loss-based weighted averaging mechanism of Q-FFL because it is able to better compromise between the data partitions of different clients, although further research is required to test this hypothesis.

2 Background and related work

2.1 Foundational challenges

Statistical (data) heterogeneity Data generation and accrual naturally vary across devices in terms of both quantity and characteristics. Factors such as sensor characteristics, geographic location, time, and user behaviour may influence the precise deviations in data distribution seen by a client—-e.g feature, label, or quantity skew as reported by Kairouz et al. (2019, sec.3.1)—, which in turn prevents treating client data as Independent and Identically Distributed (IID). Non-IID data has been shown to negatively impact both global accuracy (Zhao et al., 2018; Hsieh et al., 2020) and theoretical convergence bounds (Li et al., 2019b). A variety of techniques have been used to restrict the impact of data heterogeneity by either improving the global model (Zhao et al., 2018; Li et al., 2018; 2019b; Kairouz et al., 2019; Li et al., 2019a; Hsieh et al., 2020) or creating a personalized one (Kulkarni et al., 2020; Yu et al., 2020; Mansour et al., 2020; Jiang et al., 2019) for the clients.

System (hardware) heterogeneity Devices within the federated network may differ from one-another in terms of characteristics such as computational ability, network speed, reliability and data-gathering hardware. System heterogeneity together with device unreliability create barriers to achieving a fault and straggler-tolerant algorithm capable of accommodating different client training capabilities. Despite this fact, Li et al. (2019b) indicate that system heterogeneity is reflected similarly to data heterogeneity in the training process and largely benefits from similar solutions.

2.2 FEDERATED LEARNING OBJECTIVE

The standard loss function of FL is formulated by Li et al. (2020) as seen in Eq. (1)

$$\min_{w} f(w) = \sum_{k=1}^{m} p_k F_k(w) \ , \tag{1}$$

where f is the federated global loss function, m is the total number of devices, w is the model at the beginning of a round, and F_k is the local loss function of client k weighted by the associated p_k. For a total number of training examples n, p_k is defined as either the proportion of total training examples on the client $\frac{n_k}{n}$ or as the inverse of the number of clients $\frac{1}{m}$. The Federated Averaging (FedAvg) algorithm constructed by McMahan et al. (2017) optimizes this objective by performing local training on each client and then summing the parameters of each model weighted by p_k into an update to the previous model using an aggregation learning rate η, as seen in Eq. (2)

$$G^{t+1} = G^t + \eta \left(\sum_{k=1}^{m} p_k G_k^t \right) \ . \tag{2}$$

2.3 FAIR FEDERATED LEARNING

Li et al. (2019a) propose a form of FL, Fair Federated Learning (FFL), which attempts to construct a global model with a lower performance variance across client data partitions. The authors construct the objective function for q-FFL, a specific version of FFL, as seen in Eq. (3)

$$\min_{w} f(w) = \sum_{k=1}^{m} \frac{p_k}{q+1} F_k^{q+1}(w) \ , \tag{3}$$

where the new parameter q controls the degree of desired fairness. A value of $q = 0$ corresponds to standard FedAvg while larger values prioritize clients with higher losses. As q approaches infinity, the objective function approaches optimizing solely for the client with the largest loss.

2.4 LOCAL ADAPTATION

The analysis of Yu et al. (2020) established that not only does the federated model perform worse on highly heterogeneous clients, as previously noted by Li et al. (2019a); Kairouz et al. (2019), but it may offer inferior performance to purely local models. Thus, certain clients may receive no benefit from participating in the FL process. Yu et al. (2020) recommend addressing this accuracy gap via personalization methods using the federated model as a starting point for local training.

Multi-task Learning (MTL) The task of the global model is to perform well on the distributions of all clients while a local one must perform on the distribution of a single client. Equation (4) frames this as a Multi-task Learning (MTL) problem

$$l(C, x) = L(C, x) + \sum_{i} \tfrac{\lambda}{2} M[i](C[i] - G[i])^2 \ , \tag{4}$$

where λ determines the weighting between the two tasks, M is the Fisher information matrix and i indexes model parameters. The loss L optimizes for the client model while the second term represents the task of maintaining performance over all participants. Multiplying the squared difference of parameters by $M[i]$ mitigates Catastrophic forgetting (French, 1999; Goodfellow et al., 2013) via the Elastic Weight Consolidation technique introduced by Kirkpatrick et al. (2017).

Fine-tuning (FT) and Freezebase (FB) When a client receives a global model after the FL process, it can apply Fine-tuning (see Wang et al. (2019); Paulik et al. (2022) and Mansour et al. (2020, Section D.2)) to retrain the model on its data. To avoid potential Catastrophic forgetting, Yu et al. (2020) also opt to apply Freezebase (FB) as a variant of FT which retrains only the top layer.

Knowledge Distillation (KD) As an alternative to MTL and FT, Knowledge Distillation (Hinton et al., 2015) uses the global model as a teacher for a student client model. For the pure logit outputs of the global model $G(x)$ and client model $C(x)$, the client minimizes the loss seen in Eq. (5)

$$l(C, x) = \alpha K^2 L(C, x) + (1 - \alpha) K_L(\sigma(G(x) \,/\, K), \sigma(C(x) \,/\, K)) \,, \tag{5}$$

where L is the client loss function, K_L is the Kullback-Leibler divergence (Kullback & Leibler, 1951), σ is the softmax of the final logit output, α is the weighting of the clients' loss and K is the temperature. Equation (5) states the optimization objective as a mixture of minimizing the local loss and the KL-distance between the temperature-adjusted outputs of G and C for input x.

3 METHODS

3.1 PERFORMANCE EVALUATION

Federated and absolute local performance The first set of experiments intends to compare the accuracy on the federated test and the *absolute*—i.e. without subtracting the accuracy of a purely local model—local accuracy distribution of multiple models trained with increasing fairness levels. Too large of a drop in federated performance may cause a certain fairness level to be considered overall unusable. For a given fair model to perform well locally, it must not bring noticeable harm to average local accuracy while reducing variance when compared to FedAvg. Locally adapted models are omitted from such comparisons as they receive a similar amount of explicit on-device training to the purely local models. If local training and adaptation are unfeasible on an underperforming client due to lacking data or computational power, Q-FFL could still allow for improvements.

Relative local performance The second set involves assessing the difference in accuracy between federated or adapted models and purely local ones when evaluated directly on client data. The most important factors for the utility of such models over purely local ones are the number of clients which receive an improvement and the average improvement of the whole population. For local adaptation or a given fairness level to be worthwhile, it must reduce the number of clients which do not benefit from a federated model while maintaining or increasing the average accuracy difference. If a synergistic effect were to exist between Q-FFL and local adaptation, fairer models should either receive a larger improvement in average accuracy or result in a lower number of underperforming clients following adaptation. Conversely, fairer models may benefit less from local adaptation—especially if they already bring comparable benefits.

For both absolute accuracy and relative accuracy, we track the average and variance of the entire distribution as well as the average and variance of the best and worst 10% of performers.

3.2 Q-FEDAVG CONSIDERATIONS

The implementation of Q-FFL which is relevant to this work is q-FedAvg. In their empirical evaluation of q-FedAvg, Li et al. (2019a) show that choosing a q value is highly domain specific with optimal q values ranging from $q = 0.001$ to $q = 5$ across evaluated tasks. Two implementation details of q-FedAvg are worth mentioning. First, the choice of q impacts the optimal aggregation learning rate η. Rather than determining η for each q, the authors optimize it for $q = 0$ and then scale it for $q > 0$. Since task training parameters are already optimized for FedAvg (i.e. Q-FedAvg with $q = 0$) in all tasks, we do not change η. Second, the original publication (Li et al., 2019a) uses weighted sampling of devices based on their share of the total training examples, followed by uniform averaging. This methodology is untenable in many real-world scenarios, as it would require a server being aware of the amount of data available in each client a priori. Thus, uniform sampling and weighted averaging are preferred for the purposes of this work.

3.3 TASKS

Following the lead of Yu et al. (2020) and McMahan et al. (2017), fair FL algorithms were trained and adapted for next-word prediction on the Reddit (Caldas et al., 2018) dataset with $80\,000$ participants, and image recognition on CIFAR-10 partitioned into 100 participants. The naturally-heterogeneus Federated Extended MNIST (FEMNIST) (Caldas et al., 2018) dataset was also used

for image recognition. In order to allow direct comparison, the training pipelines and model architectures of Yu et al. (2020) and Caldas et al. (2018) were adapted. As such, we only note necessary changes from their approach below—the CIFAR-10 setup remains the same.

Next word prediction During the FL training process $\approx 5\%$ of the federated test set is used to track convergence, with the full test set being used for the final evaluation after training. Federated models are trained for $1\,000$ rounds using 20 clients per round, rather than the $5\,000$ rounds of 100 clients used by Yu et al. (2020). This will be shown to be sufficient for the baseline FedAvg model to reach an accuracy of 17.8%, which is close to the original optimum of 19.29%. Only $\sim 65\,500$ clients are locally evaluated and $\sim 18\,500$ clients adapted due to resource constraints (Appendix A.1).

FEMNIST image classification Rather than subsampling 5% of the data, we keep the first 350 clients with more than 10 data points out of the total $3\,597$ clients in the FEMINIST dataset. Since we require both local and federated testing sets, we use 70% of a client's data for training, 10% for local testing, and add the remaining 20% to the federated test set. For the FL process, we use an aggregation learning rate of $\eta = 1.0$ with 10 clients per round for 500 rounds, instead of the 3 clients per round used by Caldas et al. (2018). For each client, we use SGD with a learning rate of 0.1 and a batch size of 32. During adaptation, we lower the learning rate to 0.01.

4 Results

4.1 Q-FFL baseline absolute performance

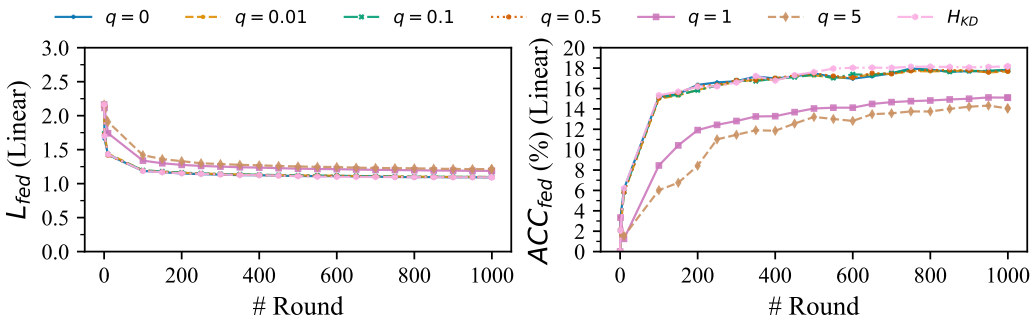


Figure 1: Federated loss (left) and accuracy (right) of fair models on the Reddit dataset. The performance of q-FedAvg rapidly decays as the exponent increases from $q = 0.5$ to $q = 1.0$. Note that the model trained using our proposed technique (H_{KD}) outperforms all others throughout.

The first set of experiments aimed to establish the performance of q-FedAvg variants on the global federated task prior to evaluating its interactions with client data partitions in either an absolute or relative sense. Figure 1 provides an overview of the convergence process for next-word prediction while Table 1 showcases summary data for both federated performance and absolute local performance. From Fig. 1 it can be seen that the impact of q on both federated training loss and accuracy is neutral to negative. Small values ($q \in [0.01, 0.5]$) lead to performance which is indistinguishable from FedAvg ($q = 0$), while larger ones ($q \in [1, 5]$) show a noticeable dip in federated and average accuracy. Fairness only seems to reliably reduce the variance of the whole distribution for $q \geq 1$, however, variance across the worst 10% (i.e., the underperformers) decreases linearly for all q. Overall, q-FedAvg shows a viable level of federated and local performance for $q \in [0.01, 0.5]$.

Table 2 indicates the CIFAR-10 image classification task to be more resilient in terms of performance than next word prediction, with only a small decrease being noticeable in models trained using $q \geq 10$. The lower sensitivity of this task to training parameters is consistent with previous findings (Yu et al., 2020) on differential private FL algorithms and robust FL algorithms. Unlike next word prediction on Reddit, the variance-reducing effects of Q-FFL peak at $q = 10$ alongside federated

Objective	$ACC_{fed}(\%)$	$Avg_{loc}(\%)$	$B_{loc}10\%(\%)$	$W_{loc}10\%(\%)$	(Var_{Avg})	(Var_B)	(Var_W)
$q = 0$	17.826	18.636	24.552	14.803	9.169	22.066	1.127
$q = 0.01$	17.633	18.431	24.035	14.676	8.487	**21.068**	1.23
$q = 0.1$	17.789	18.652	24.826	14.726	9.797	22.961	1.078
$q = 0.5$	17.718	18.55	24.639	14.676	9.583	23.096	1.055
$q = 1$	15.104	15.836	21.391	12.566	8.35	27.308	0.825
$q = 5$	14.056	14.813	20.2	11.653	**8.003**	26.969	**0.714**
H_{KD}	**18.177**	**19.17**	**26.406**	**14.879**	12.442	26.101	1.083
Local	NaN	4.458	10.25	1.208	8.884	32.008	0.889

Table 1: Results showing the federated and absolute local performance on Reddit. While fairness does decrease variance at $q \geq 1.0$, the harm to accuracy is too great. For lower levels of fairness, no clear benefits except for decreased variance amongst the worst performers may be observed. The proposed KD-based model outperforms on accuracy across the distribution but increases variance for everyone except the worst performers.

performance. This lack of sensitivity to fairness caused the omission of the CIFAR-10 convergence graph from this section, although it is available in the appendix (Fig. 4).

Objective	$ACC_{fed}(\%)$	$Avg_{loc}(\%)$	$B_{loc}10\%(\%)$	$W_{loc}10\%(\%)$	(Var_{Avg})	(Var_B)	(Var_W)
$q = 0$	81.28	**81.37**	**82.255**	**79.864**	0.568	0.022	1.067
$q = 0.1$	81.03	80.972	81.874	79.35	0.61	0.014	1.188
$q = 1$	81.66	81.16	82.086	79.508	0.603	0.008	0.821
$q = 5$	**81.86**	81.221	82.011	79.794	**0.446**	**0.004**	**0.643**
$q = 10$	76.64	80.004	81.213	77.826	0.991	0.011	1.484
$q = 15$	78.16	79.935	81.178	77.885	0.945	0.02	1.267
Local	NaN	31.718	38.297	24.649	16.3	1.543	0.906

Table 2: Results for CIFAR-10. Unlike the language task (Table 1), $q = 5$ represents a clear optimum in terms of variance while maintaining performance. However, differences between all models are very small and cannot be guaranteed to be significant.

On the opposite end of the spectrum from CIFAR-10, Fig. 2 and Table 3 show FEMNIST image recognition models to be highly sensitive to fairness as their performance either oscillates repeatedly or diverges. Since all models except for those trained with FedAvg ($q = 0$) diverge before the end of FL, we test and adapt the last version of a model prior to divergence. The natural heterogeneity of FEMNIST results in a much wider gap between average, the worst performer, and best performer accuracy for both federated and fully local models. Unlike the previous datasets, nothing resembling a trend emerges for any metric. Nevertheless, a good balance between performance and average variance is struck by $q = 10$—except for the fairly high worst-performer variance.

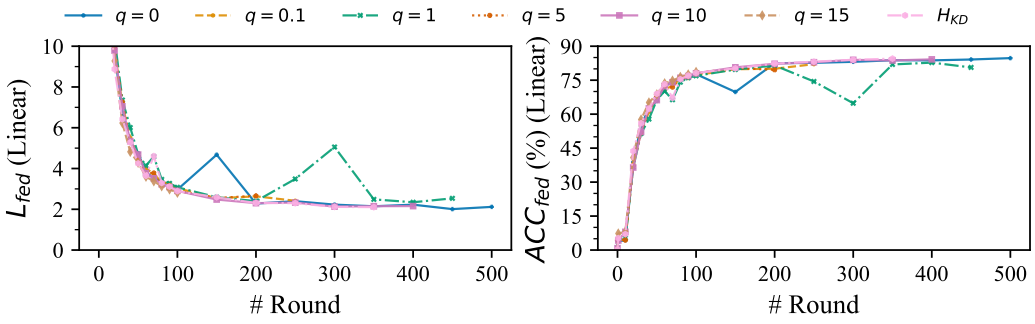


Figure 2: Federated loss (left) and accuracy (right) of fair models on FEMNIST. Fairness causes increased instability in the training process or outright divergence past a certain round. Our proposed technique also diverges, however it reaches a similar accuracy to FedAvg ($q = 0$) prior to doing so.

Objective	$ACC_{fed}(\%)$	$Avg_{loc}(\%)$	$B_{loc}10\%(\%)$	$W_{loc}10\%(\%)$	(Var_{Avg})	(Var_B)	(Var_W)
$q = 0$	**84.739**	75.341	99.435	35.717	432.507	1.151	73.747
$q = 0.1$	82.196	**73.749**	98.503	38.901	348.465	1.979	46.692
$q = 1$	80.628	73.333	98.425	37.728	339.837	2.33	59.295
$q = 5$	79.556	71.37	97.561	35.985	355.291	4.312	**31.672**
$q = 10$	84.19	76.591	99.013	**42.055**	**320.385**	1.714	64.049
$q = 15$	78.634	69.749	96.681	34.988	374.637	5.177	47.761
H_{KD}	84.51	75.243	**99.491**	34.34	443.015	**1.07**	62.91
Local	NaN	46.322	92.848	0.0	1006.77	18.144	0.0

Table 3: Results for FEMNIST. Given the much larger variability in performance compared to CIFAR-10, the only acceptable fair model is the one trained with $q = 10$. Note that the average accuracy of the best-performing local clients is close to that of the best-performing partitions for the federated model. Unlike the language task, using KD during FL seems to disproportionately help the best performers—which is likely tied to the high potential of local models.

4.2 Q-FFL FAILS TO IMPROVE RELATIVE PERFORMANCE

Objective	Adapt	$Avg(\%)$	$Acc < 0$	$B10\%(\%)$	$W10\%(\%)$	(Var_{Avg})	(Var_B)	(Var_W)
$q = 0$	$q = 0$	14.185	53	20.715	9.392	**13.323**	34.379	20.201
	A_FB	15.87	0	25.849	**11.311**	29.216	149.736	3.246
	A_MTL	**16.046**	0	<u>27.558</u>	11.304	36.067	178.387	3.337
	A_KD	15.538	0	24.376	11.209	23.112	115.016	<u>3.183</u>
$q = 0.01$	$q = 0.01$	13.989	106	20.458	8.793	<u>14.421</u>	**33.874**	28.638
	A_FB	15.799	0	25.7	<u>11.238</u>	29.093	149.929	3.228
	A_MTL	<u>15.833</u>	1	<u>27.419</u>	11.057	36.547	180.574	3.326
	A_KD	15.495	0	24.308	11.175	23.343	117.768	<u>3.173</u>
$q = 0.1$	$q = 0.1$	14.208	50	20.907	9.359	<u>13.742</u>	<u>35.005</u>	20.733
	A_FB	15.827	0	25.964	<u>11.261</u>	29.505	149.011	3.212
	A_MTL	<u>15.839</u>	0	<u>27.692</u>	11.024	37.108	179.066	3.336
	A_KD	15.546	0	24.614	11.19	23.95	118.471	<u>3.166</u>
$q = 0.5$	$q = 0.5$	14.097	53	20.705	9.287	<u>13.514</u>	<u>34.546</u>	20.655
	A_FB	15.837	0	25.838	<u>11.285</u>	29.2	148.925	3.232
	A_MTL	<u>15.892</u>	0	<u>27.555</u>	11.133	36.531	179.11	3.26
	A_KD	15.52	0	24.42	11.186	23.402	116.556	<u>3.199</u>
H_{KD}	H_{KD}	14.729	27	22.038	9.971	<u>14.969</u>	<u>41.225</u>	18.409
	A_FB	15.772	0	26.533	11.154	31.068	150.552	**3.156**
	A_MTL	<u>15.824</u>	2	**28.217**	10.966	38.439	178.543	3.469
	A_KD	15.698	0	25.358	<u>11.214</u>	25.367	119.418	3.241

Table 4: Relative performance on the Reddit dataset of the acceptable fair models and our proposed KD-based alternative. The best value in a column is in bold while the best in a group is underlined. The worst fair model trained with $q = 0.01$ doubles the number of underperforming ($Acc < 0$) clients while the others leave it virtually unchanged. Alternatively, H_{KD} lowers it to half. Local adaptation seems to provide nearly-equivalent results regardless of starting point.

Having established baselines of accuracy for fair models, we can now evaluate the local relative performance of Q-FFL and its interactions with local adaptation. There are several considerations involved in doing this. First, the CIFAR-10 data (Table 7) is unsuitable for multiple reasons including its artificial partitioning (unlike FEMNIST) and the lack of sensitivity to either fairness or local adaptation—the federated model outperforms a fully local one for all clients across all models. Second, in this section, we only showcase data for Reddit models in the viable $[0, 0.5]$ range and for FEMNIST at the optimum $q = 10$. However, result sets for Reddit (Table 6), CIFAR-10 (Table 7) and FEMNIST (Table 8) are available in their entirety in the appendix.

The results for the language task showcased in Table 4 are less than satisfactory as fair models fail to provide benefits in terms of the number of underperforming clients ($q = 0.01$ doubles it), relative accuracy, or variance on average. Furthermore, fair models do not offer improved accuracy once adapted. This lack of benefits is also directly visible in the Fig. 3a scatter plot of relative accuracy against fully local accuracy.

Objective	Adapt	$Avg(\%)$	$Acc < 0$	$B10\%(\%)$	$W10\%(\%)$	(Var_{Avg})	(Var_B)	(Var_W)
$q = 0$	$q = 0$	29.02	16	65.768	2.729	338.387	137.366	12.33
	A_FB	28.954	17	65.463	2.72	334.754	134.824	11.853
	A_MTL	28.994	16	65.672	**2.802**	336.25	137.507	**11.325**
	A_KD	28.986	16	65.684	2.788	337.14	137.796	11.594
$q = 10$	$q = 10$	**30.269**	39	**79.687**	-14.844	673.351	111.144	206.995
	A_FB	28.613	**12**	64.818	2.699	**320.729**	123.679	15.552
	A_MTL	28.612	14	64.818	2.516	321.3	123.679	15.869
	A_KD	28.563	14	64.645	2.52	320.957	127.618	15.934
H_{KD}	H_{KD}	28.921	16	66.178	1.65	353.698	154.217	16.231
	A_FB	28.916	17	66.605	1.644	350.631	152.111	16.922
	A_MTL	28.871	16	66.605	1.719	350.742	152.111	16.865
	A_KD	28.967	15	66.329	1.869	351.991	150.583	17.124

Table 5: FEMNIST performance of the best fair model and our proposed alternative to Q-FFL. Despite providing the highest average accuracy and accuracy amongst the best 10%, the model trained using $q = 10$ has more than double the number of underperformers ($Acc < 0$) of FedAvg ($q = 0$). The model trained using H_{KD} on the other hand stays close to the FedAvg baseline.

The results for image recognition on FEMNIST are more unusual yet similarly discouraging. Direct numerical data in Table 5 makes it clear that the fair model actually achieves a higher relative accuracy on average and amongst the top 10% of performers at the cost of obtaining a *negative* average accuracy on the worst performers. Moreover, it has more than twice as many clients with negative relative accuracies. Since the federated model has a lower absolute accuracy variance, it cannot obtain a good enough level of performance on clients which are able to train very high-quality local models. This is corroborated by the final distribution shown in Fig. 3b, as almost *all* the underperforming clients have a high purely local model accuracy. Another factor to consider is the atypical behaviour of local adaptation on FEMNIST and CIFAR-10, as the FedAvg-trained federated and adapted models tend to converge to nearly the same relative accuracy regardless of the specific adaptation technique. Thus, FedAvg is likely near-optimal for the dataset and architecture.

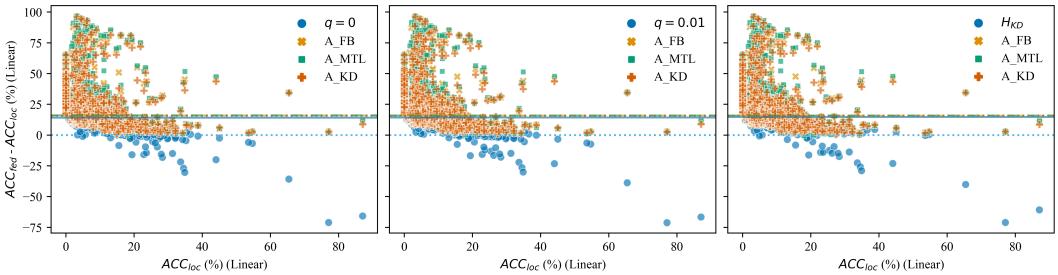
4.3 Knowledge Distillation as an Alternative

Having shown the inability of Q-FFL to replace or enhance local adaptation, we argue that it is not the right approach for this domain fundamentally and propose an alternative. In principle, for an FL algorithm to provide benefits in terms of relative accuracy, it must achieve two goals. First, it must make sure that the clients on whose data the *federated model performs worst* still receive a sufficient level of accuracy to match or exceed local models. Second, for the clients with the *best* local models, it must provide disproportionately high accuracy. While Q-FFL may fulfil the first requirement, its inability to only raise the floor of the worst performers without hurting the ceiling of those that might have a good local model makes it incapable of fulfilling the second requirement.

We propose replacing Q-FFL with a system where clients use the Knowledge Distillation loss function (Eq. (5)) during federated training rather than only their local loss. This alternative loss function allows local models to diverge while maintaining performance on the global task; and thus, could be better adept at dealing with data heterogeneity on the client. The intent is to raise the accuracy of every client, especially highly heterogeneous ones, consequently improving the relative accuracy distribution. As a preliminary form of this paradigm, we train a model with standard FedAvg for half the rounds and then switch to KD using the same parameters as for local adaptation. We notate a model trained this way H_{KD}, results for this technique have already been included in previous tables to save space and allow for easy comparison.

Preliminary results for the language task are promising, Fig. 1 and Table 1 indicate that H_{KD} performs better than FedAvg and q-FedAvg-trained models in every metric except average and best-performer variance. Importantly, variance is not increased for the worst performers. In terms of relative accuracy, Table 4 shows that H_{KD} halves the number of underperformers and provides the best average relative accuracy at baseline. However, this higher baseline does not seem to translate to increased relative accuracy for later-adapted models. Overall, lowering the number of clients

(a) Language task results, fairness shows no benefit while H_{KD} reduces the number of underperformers.



(b) FEMNIST results, clients with highly accurate local models are underserved by federated models trained using $q = 10$. Alternatively, those trained using FedAvg or H_{KD} achieve nearly identical performance.

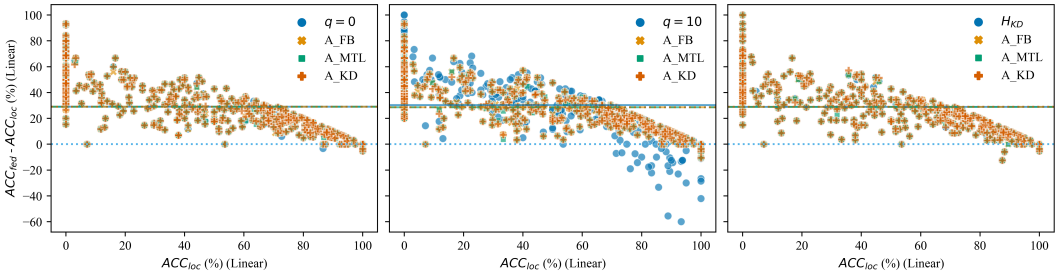


Figure 3: Federated model relative accuracy on a client plotted against the accuracy of a fully local model trained for that client.

which *require* adaptation in order to receive an incentive to participate H_{KD} successfully reduces the need for personalization on the language task.

For image recognition on FEMNIST, H_{KD} is satisfactory in terms of federated and average accuracy according to Fig. 2 and Table 3. On the other hand, the results related to relative accuracy shown in Fig. 3b and Table 5 are mixed. While H_{KD} avoids the doubling in underperforming clients that fair models suffer, locally adapted models starting from H_{KD} as a baseline seem to perform slightly worse than those starting from the fair model—though more experiments are needed to verify this.

5 Conclusion

This paper set out to find a means of reducing the amount of personalization needed to incentivize FL participation for clients for whom the federated model underperforms relative to a purely local one. Such a reduction would be relevant for federated networks containing clients with limited capabilities for retraining. Our investigation began with Q-Fair Federated Learning due to the possibility that reducing disparity in the local accuracy distribution would translate to reducing disparity in the relative accuracy distribution. Our experimental results indicate that Q-FFL is unlikely to provide the desired properties as it has been shown to maintain the number of underperformers on our language task while increasing it by 143% for image recognition on FEMNIST. We speculate that the reason for this is that although it can help clients for which the global model is highly inaccurate, it cannot help those for whom relative underperformance is caused by an extremely accurate local model. As an alternative, we propose using knowledge distillation during FL training given its ability to allow local training to diverge while maintaining performance on the federated task—thus leading to a better compromise. Initial performance across tasks showed promising results in reducing the number of underperformers by up to 50% on the language task without increasing the number of underpeformers on the FEMNIST image recognition task. Consequently, we recommend using KD to incentivize FL participation without explicit local adaptation and advise further research into applying either KD or similar techniques such as Multi-task Learning during FL training.

REFERENCES

Sebastian Caldas, Peter Wu, Tian Li, Jakub Konečný, H. Brendan McMahan, Virginia Smith, and Ameet Talwalkar. LEAF: A benchmark for federated settings. *CoRR*, abs/1812.01097, 2018. URL http://arxiv.org/abs/1812.01097.

Robert M. French. Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135, 1999. ISSN 1364-6613. doi: https://doi.org/10.1016/S1364-6613(99)01294-2. URL https://www.sciencedirect.com/science/article/pii/S1364661399012942.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks, 2013. URL https://arxiv.org/abs/1312.6211.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015. URL https://arxiv.org/abs/1503.02531.

Kevin Hsieh, Amar Phanishayee, Onur Mutlu, and Phillip Gibbons. The non-IID data quagmire of decentralized machine learning. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4387–4398. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hsieh20a.html.

Yihan Jiang, Jakub Konečný, Keith Rush, and Sreeram Kannan. Improving federated learning personalization via model agnostic meta learning. *CoRR*, abs/1909.12488, 2019. URL http://arxiv.org/abs/1909.12488.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *arXiv preprint arXiv:1912.04977*, 2019.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526, 2017.

Viraj Kulkarni, Milind Kulkarni, and Aniruddha Pant. Survey of personalization techniques for federated learning. In *2020 Fourth World Conference on Smart Trends in Systems, Security and Sustainability (WorldS4)*, pp. 794–797, 2020. doi: 10.1109/WorldS450073.2020.9210355.

S. Kullback and R. A. Leibler. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79 – 86, 1951. doi: 10.1214/aoms/1177729694. URL https://doi.org/10.1214/aoms/1177729694.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. *arXiv preprint arXiv:1812.06127*, 2018.

Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. *arXiv preprint arXiv:1905.10497*, 2019a.

Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37(3):50–60, 2020.

Xiang Li, Kaixuan Huang, Wenhao Yang, Shusen Wang, and Zhihua Zhang. On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*, 2019b.

Y. Mansour, Mehryar Mohri, Jae Ro, and Ananda Theertha Suresh. Three approaches for personalization with applications to federated learning. *ArXiv*, abs/2002.10619, 2020.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pp. 1273–1282. PMLR, 2017.

Matthias Paulik, Matt Seigel, Henry Mason, Dominic Telaar, Joris Kluivers, Rogier van Dalen, Chi Wai Lau, Luke Carlson, Filip Granqvist, Chris Vandevelde, Sudeep Agarwal, Julien Freudiger, Andrew Byde, Abhishek Bhowmick, Gaurav Kapoor, Si Beaumont, Áine Cahill, Dominic Hughes, Omid Javidbakht, Fei Dong, Rehan Rishi, and Stanley Hung. Federated evaluation and tuning for on-device personalization: System design & applications, 2022. URL https://arxiv.org/pdf/2102.08503.pdf.

Kangkang Wang, Rajiv Mathews, Chloé Kiddon, Hubert Eichner, Françoise Beaufays, and Daniel Ramage. Federated evaluation of on-device personalization. *CoRR*, abs/1910.10252, 2019. URL http://arxiv.org/abs/1910.10252.

Tao Yu, Eugene Bagdasaryan, and Vitaly Shmatikov. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data, 2018.

A APPENDIX

A.1 HARDWARE

Each node of the cluster that the experiments were run on holds four Nvidia A100 GPUs with 80GiB VRAM and 2 x AMD EPYC 7763 64-Core processors alongside a 1000GiB shared memory. Given the quotas and service levels of the cluster, the number of clients on which the federated model could be tested locally for the language task was limited to $\sim 65\,500$ while the number that could be adapted was limited to (~ 18500)—the entire client set was available during FL training. To compensate for this fact, all charts and tables comparing local model performance or adaptation performance only use data from the client-set common to all results. This limitation did not impact any other sets of experiments.

A.2 FULL RESULTS

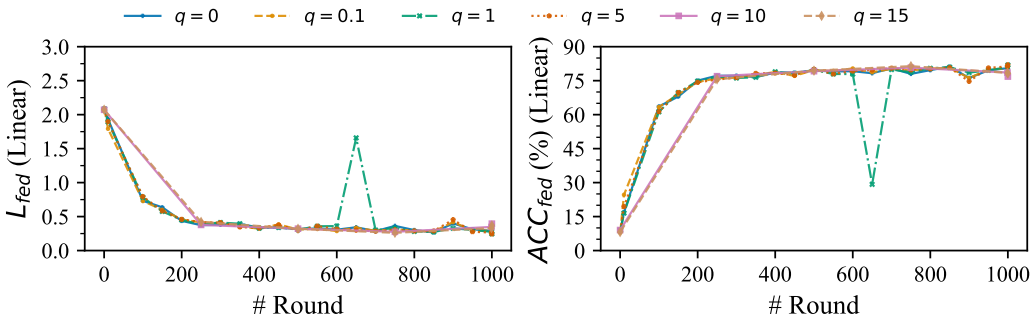


Figure 4: Federated performance of fair models on CIFAR-10. q-FedAvg performs marginally worse for $q \geq 10.0$, however, it must be concluded that the accuracy of the given task is not sensitive enough to fairness to draw strong conclusions.

Objective	Adapt	$Avg(\%)$	$Acc < 0$	$B10\%(\%)$	$W10\%(\%)$	(Var_{Avg})	(Var_B)	(Var_W)
$q = 0$	$q = 0$	14.185	53	20.715	9.392	13.323	34.379	20.201
	A_FB	15.87	0	25.849	11.311	29.216	149.736	3.246
	A_MTL	16.046	0	27.558	11.304	36.067	178.387	3.337
	A_KD	15.538	0	24.376	11.209	23.112	115.016	3.183
$q = 0.01$	$q = 0.01$	13.989	106	20.458	8.793	14.421	33.874	28.638
	A_FB	15.799	0	25.7	11.238	29.093	149.929	3.228
	A_MTL	15.833	1	27.419	11.057	36.547	180.574	3.326
	A_KD	15.495	0	24.308	11.175	23.343	117.768	3.173
$q = 0.1$	$q = 0.1$	14.208	50	20.907	9.359	13.742	35.005	20.733
	A_FB	15.827	0	25.964	11.261	29.505	149.011	3.212
	A_MTL	15.839	0	27.692	11.024	37.108	179.066	3.336
	A_KD	15.546	0	24.614	11.19	23.95	118.471	3.166
$q = 0.5$	$q = 0.5$	14.097	53	20.705	9.287	13.514	34.546	20.655
	A_FB	15.837	0	25.838	11.285	29.2	148.925	3.232
	A_MTL	15.892	0	27.555	11.133	36.531	179.11	3.26
	A_KD	15.52	0	24.42	11.186	23.402	116.556	3.199
$q = 1$	$q = 1$	11.397	70	17.618	7.047	12.546	37.951	19.031
	A_FB	13.463	0	22.632	9.256	25.775	136.853	1.987
	A_MTL	13.364	4	24.519	8.993	35.506	187.701	2.376
	A_KD	13.238	0	21.567	9.245	20.821	105.756	2.016
$q = 5$	$q = 5$	10.384	121	16.583	5.904	12.952	39.358	21.502
	A_FB	12.771	0	21.703	8.678	24.874	133.58	1.694
	A_MTL	12.838	6	23.37	8.548	33.262	181.204	2.146
	A_KD	12.492	0	20.608	8.614	19.704	99.431	1.695
H_{KD}	H_{KD}	14.729	27	22.038	9.971	14.969	41.225	18.409
	A_FB	15.772	0	26.533	11.154	31.068	150.552	3.156
	A_MTL	15.824	2	28.217	10.966	38.439	178.543	3.469
	A_KD	15.698	0	25.358	11.214	25.367	119.418	3.241

Table 6: Reddit full table of results.

Objective	Adapt	$Avg(\%)$	$Acc < 0$	$B10\%(\%)$	$W10\%(\%)$	(Var_{Avg})	(Var_B)	(Var_W)
$q = 0$	$q = 0$	49.652	0	56.54	43.045	15.557	0.79	2.563
	A_FB	49.669	0	56.522	43.145	15.364	0.839	2.165
	A_MTL	49.647	0	56.489	43.064	15.468	0.829	2.267
	A_KD	49.647	0	56.527	43.066	15.545	0.789	2.076
$q = 0.1$	$q = 0.1$	49.254	0	56.078	42.839	15.408	0.858	1.94
	A_FB	49.23	0	55.997	42.758	15.492	0.987	1.923
	A_MTL	49.257	0	55.925	42.933	15.569	1.129	2.077
	A_KD	49.248	0	56.047	42.82	15.486	0.996	1.973
$q = 1$	$q = 1$	49.442	0	56.297	42.865	15.614	1.053	1.991
	A_FB	49.448	0	56.308	42.812	15.729	0.98	2.423
	A_MTL	49.425	0	56.352	42.786	15.728	0.966	2.287
	A_KD	49.439	0	56.314	42.78	15.742	0.981	2.307
$q = 5$	$q = 5$	49.503	0	56.25	42.864	15.583	0.755	2.07
	A_FB	49.517	0	56.315	42.974	15.617	0.77	2.091
	A_MTL	49.522	0	56.404	42.909	15.651	0.603	1.874
	A_KD	49.548	0	56.416	42.991	15.531	0.678	1.902
$q = 10$	$q = 10$	48.286	0	54.949	42.033	14.417	0.608	1.956
	A_FB	48.3	0	54.948	42.096	14.275	0.424	1.847
	A_MTL	48.306	0	55.025	42.087	14.468	0.565	1.71
	A_KD	48.297	0	54.987	42.021	14.504	0.507	1.873
$q = 15$	$q = 15$	48.217	0	54.745	41.842	14.449	0.585	2.434
	A_FB	48.22	0	54.718	41.883	14.482	0.619	1.859
	A_MTL	48.202	0	54.761	41.824	14.444	0.614	2.191
	A_KD	48.193	0	54.779	41.713	14.763	0.564	2.473

Table 7: CIFAR-10 full table of results.

Objective	Adapt	$Avg(\%)$	$Acc < 0$	$B10\%(\%)$	$W10\%(\%)$	(Var_{Avg})	(Var_B)	(Var_W)
$q = 0$	$q = 0$	29.02	16	65.768	2.729	338.387	137.366	12.33
	A_FB	28.954	17	65.463	2.72	334.754	134.824	11.853
	A_MTL	28.994	16	65.672	2.802	336.25	137.507	11.325
	A_KD	28.986	16	65.684	2.788	337.14	137.796	11.594
$q = 0.1$	$q = 0.1$	27.427	45	80.262	-23.681	761.163	117.165	165.003
	A_FB	26.174	14	59.483	2.09	282.14	128.526	14.212
	A_MTL	26.205	14	59.567	2.345	281.689	128.659	15.377
	A_KD	26.18	15	59.613	1.997	282.946	127.796	14.284
$q = 1$	$q = 1$	27.011	48	78.722	-18.92	699.58	131.927	150.33
	A_FB	24.999	19	57.014	0.513	270.651	140.273	24.641
	A_MTL	25.022	17	57.179	0.76	271.681	136.18	23.188
	A_KD	25.053	17	57.165	0.859	270.926	135.297	22.658
$q = 5$	$q = 5$	25.048	46	73.939	-23.702	683.322	129.091	177.272
	A_FB	25.385	18	56.625	1.758	264.374	97.084	8.057
	A_MTL	25.468	16	56.64	2.031	261.138	96.776	8.409
	A_KD	25.437	17	57.116	1.695	267.606	91.34	10.717
$q = 10$	$q = 10$	30.269	39	79.687	-14.844	673.351	111.144	206.995
	A_FB	28.613	12	64.818	2.699	320.729	123.679	15.552
	A_MTL	28.612	14	64.818	2.516	321.3	123.679	15.869
	A_KD	28.563	14	64.645	2.52	320.957	127.618	15.934
$q = 15$	$q = 15$	23.427	56	74.969	-23.34	699.404	145.773	170.676
	A_FB	22.008	17	51.138	1.039	214.214	115.322	7.717
	A_MTL	21.848	21	51.167	0.149	222.45	115.582	9.945
	A_KD	21.935	17	51.467	0.84	216.567	112.101	8.538
H_{KD}	H_{KD}	28.921	16	66.178	1.65	353.698	154.217	16.231
	A_FB	28.916	17	66.605	1.644	350.631	152.111	16.922
	A_MTL	28.871	16	66.605	1.719	350.742	152.111	16.865
	A_KD	28.967	15	66.329	1.869	351.991	150.583	17.124

Table 8: FEMNIST full table of results.