Federated Multi-Task Learning for Competing Constraints

Tian Li
CMU
tianli@cmu.edu

Shengyuan Hu
CMU
shengyua@andrew.cmu.edu

Ahmad Beirami
Facebook AI
beirami@fb.com

Virginia Smith
CMU
smithv@cmu.edu

Abstract

In addition to accuracy, fairness and robustness are two critical concerns for federated learning systems. In this work, we first identify that *robustness* to adversarial training-time attacks and *fairness*, measured as the uniformity of performance across devices, are competing constraints in statistically heterogeneous networks. To address these constraints, we propose employing a simple, general multi-task learning objective, and analyze the ability of the objective to achieve a favorable trade-off between fairness and robustness. We develop a scalable solver for the objective and show that multi-task learning can enable more accurate, robust, and fair models relative to state-of-the-art baselines across a suite of federated datasets.

1 Introduction

Federated learning (FL) aims to collaboratively learn from data that has been generated by, and resides on, a number of remote devices [34]. FL stands to produce highly accurate statistical models by aggregating knowledge from disparate data sources. However, to deploy federated learning in practice, it is necessary for the resulting systems to be not only accurate, but to also satisfy a number of pragmatic constraints, regarding issues such as fairness, robustness, privacy, and security. Simultaneously satisfying these various constraints can be exceptionally difficult.

In this work, we focus specifically on targeting constraints between accuracy, fairness (i.e., uniform performance distribution across the network), and robustness (against training-time attacks).[1] Many prior efforts have separately considered fairness or robustness in federated learning. For instance, strategies to enforce fairness include focusing on the worst-performing devices by solving minimax optimization [21, 35] or reweighting the devices to promote accuracy uniformity and allow for a flexible fairness/accurcy trade-off [27, 28]. Common robust methods include techniques such as gradient clipping [40] or the use of robust aggregators to combine model updates [6, 44].

While these approaches are effective at either promoting fairness or defending against training-time attacks in isolation, we show that the constraints of fairness and robustness can directly compete with one another, and that simultaneously optimizing for accuracy, fairness, and robustness requires careful consideration. For example, as we empirically demonstrate (Section 4), current fairness approaches can render FL systems highly susceptible to training time attacks from malicious devices. On the other hand, robust baselines may filter out rare but informative updates, resulting in unfairness.

While addressing the competing constraints of FL may seem like an insurmountable problem, in this work we identify that *statistical heterogeneity* is a root cause for tension between these constraints, and is key in paving a path forward. Our insight is that we can better address competing constraints by properly modeling and accounting for heterogeneity in federated learning. In particular, we propose to use *multi-task learning* (a framework that learns shared, heterogeneous models) to model federated

[1] We focus on the attacks against the main learning task, as opposed to backdoor attacks [2, 40, 41]. Typically, main-task attacks could compromise the entire model, particularly hurting the predictive power of benign devices. Therefore, we define robustness as the test performance of benign devices.

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

data. Multi-task learning (MTL) has been studied in non-federated settings [e.g., 15], as well as convex settings for federated learning [38]. However, it has not been directly explored as a solution for robustness and fairness, and the area of non-convex federated MTL is generally less explored.

To this end, we propose a simple yet effective multi-task objective to achieve robustness and fairness jointly in both convex and non-convex settings. To solve the objective, we provide a lightweight and scalable solver which accounts for low device participation and local updating for federated settings. Theoretically, we take a first step towards rigorously analyzing the benefits of multi-task learning to handle competing constraints on a toy problem. Practically, through experiments on federated data, across a set of attacks, we demonstrate that our multi-task learning approach is more accurate, robust, and fair compared with strong baselines that handle these constraints separately.

2 Related Work

Robustness in Federated Learning. Training-time attacks (including data poisoning and model poisoning attacks) have been well studied in the machine learning community. In data poisoning attacks [5], an adversary can compromise the model by corrupting a set of carefully-chosen samples [9, 18, 22, 30, 37, 43] or injecting poisoned data points into the training set [14, 22, 41]. In federated settings, Bagdasaryan et al. [2] show that data poisoning by a single adversary is insufficient to compromise the global model. As a result, stronger attack methods have been proposed including scaling malicious model updates [2], collaborative attacking [39], or adding edge-case adversarial training samples [41]. Some other works focus on performing attacks in a defense-aware fashion to make the malicious updates evasive from benign updates [3, 17]. We make different assumptions on the adversaries for different scenarios, and investigate label flipping [3, 4], sending random updates, and scaling malicious updates [2] as attack baselines. While we do not focus on backdoor attacks in this work, exploring MTL for this problem would be an interesting direction of future study.

In terms of defenses against training-time attacks, in distributed settings, robust aggregation [6, 36, 40] is a common strategy to mitigate the effect of malicious updates. However, these robust aggregators usually rely on the assumption of I.I.D. data, which may not be applicable in federated settings. Other defense mechanisms include gradient clipping [40] or normalization [21]. While these works can improve robustness, they could produce *unfair* models by potentially filtering out many informative updates, especially in heterogeneous environments [41]. In this work, we compare multi-task learning with several strong defenses (median, gradient clipping [40], Krum, Multi-Krum [6], gradient-norm based anomaly detector [2], and a new defense proposed herein) and demonstrate that multi-task learning is able to improve both robustness and fairness compared with these methods.

Fairness in Federated Learning. Fairness issues (i.e., uniformity of performance distribution), also known as representation disparity [20], are a major concern in training in heterogeneous networks. Some works propose minimax optimization [35] or alternative approaches to reweighting samples [27, 28] to encourage a more fair quality of service offered to all devices. Other works consider varying notions of fairness (e.g., proportional fairness [46]) in federated learning. Nevertheless, fair methods may not be robust in that they can easily overfit to corrupted devices (Section 4.1). While Hu et al. [21] consider both fairness and robustness in one algorithm, this work combines classical fairness and limited robustness mechanisms (i.e., minimax optimization and gradient normalization), as opposed to the multi-task framework proposed herein to jointly address the constraints in a unified manner.

Personalized Federated Learning. In federated learning, personalization is a natural approach to handling statistical heterogeneity by fitting separate models to the distributed data, while increasing the effective sample size on each local device. Several previous works have proposed different methods to train personalized models. Smith et al. [38] propose a primal-dual multi-task learning framework for federated learning, which only applies to convex settings. Some works use alternative objectives to learn personalized models interpolated between the global and local models [11, 33]. However, the solutions to these objectives can reduce to local minimizers. Another approach is to explicitly regularize the local models, either towards their average [19], or towards a reference point [12]. Another line of work applies meta-learning for personalization in federated settings [8, 16, 23, 24]. Other works enforce hard parameter sharing among local models [1, 29]. However, few of these works explore the benefits of personalization in terms of fairness or robustness defined herein. Wang et al. [42] empirically show that local finetuning can help fairness. Yu et al. [45] use different personalization methods (including multi-task learning) to improve accuracies after applying robust mechanisms. Our work instead argues that multi-task learning itself offers robustness benefits, and can provide both robustness and fairness to benign devices simultaneously. Our objective is also

inspired by the classical mean-regularized multi-task learning objective [15]; but we regularize towards a global model rather than the mean of local models.

3 Federated Multi-Task Learning for Accuracy, Fairness, and Robustness

In this section, we first describe our proposed multi-task objective (Section 3.1), and then present a scalable algorithm to solve this objective in federated settings (Section 3.2).

3.1 Multi-Task Learning Objective

In federated learning, a classical objective is to learn a single global model to fit all data across the network [34]. In particular, we aim to solve:

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

where $F_k(w)$ $(1 \leqslant k \leqslant N)$ is the local loss for device k and p_k is a pre-defined weight such that $\sum_k p_k = 1$. In general, each device may generate data x_k via a distinct distribution $\mathcal{D}_k$, i.e., $F_k(w) := \mathbb{E}_{x_k \sim \mathcal{D}_k}\left[f_k(w; x_k)\right]$ where f_k is the loss on an individual sample. Due to such statistical heterogeneity, the model performance can vary across all devices and be highly non-uniform. In addition, with the presence of malicious devices, training a global model can negatively affect its performance on the benign devices, i.e., lacking robustness. To better account for heterogeneity which could allow us to simultaneously achieve fairness and robustness, we propose to leverage multi-task learning to model federated data.

In particular, we propose the following multi-task learning objective to learn separate models for each device. In order to incorporate global information, we use an L_2 regularizer to enforce local models to be closer to the optimal global model. The goal is to solve

$$\min_{\{w_k\}, 1 \leqslant k \leqslant N} \sum_{k=1}^{N} p_k \left(F_k(w_k) + \frac{\lambda}{2} \left\| w_k - w^* \right\|^2 \right), \tag{2}$$

where w^* is the optimal global model defined as $w^* = \arg\min_w \sum_{k=1}^{N} p_k F_k(w)$. λ is a hyperparameter that controls the interpolation between local and global models. When λ is set to 0, Objective (2) is reduced to training local models; as λ grows large, it recovers global model optimization (Objective (1)). To reason about the benefits of Objective (2), let us consider a simple case where the data are *homogeneous* across devices. Without adversaries, learning a single global model is optimal for generalization. However, in the presence of adversaries, learning globally might introduce lethal corruption, while learning local models may not generalize well due to the limited sample size. Our framework with an appropriate λ offers a trade-off between these two extremes to achieve optimal generalization and robustness even with I.I.D. data. In the heterogeneous case, a finite λ exists to offer optimal robustness and fairness simultaneously. We make rigorous the benefits of multi-task learning in terms of the accuracy-fairness-robustness divide on a toy problem in Appendix B.

Empirically, as we demonstrate in Section 4.2, Objective (2) produces more robust and fair personalized models for benign devices in both convex and non-convex settings.

Other Multi-Task Learning Objectives for Federated Learning. As a first step towards exploring the benefits of multi-task learning to robustness and fairness, we only consider objectives that can be optimized relatively easily in federated settings. There are some prior works using related but different multi-task objectives for personalization [11, 19, 33]. Some approaches interpolate between local and global models [11, 33]. Nevertheless, perhaps surprisingly, they are essentially solving local problems separately. For instance, in strongly convex settings where there is a unique empirical minimizer for each local device, solving the objectives in Mansour et al. [33] and Deng et al. [11] will degenerate into finding the exact local minimizers, as the optimal global model w^* to interpolate with is fixed. To further validate this, we empirically show that the objective proposed in Deng et al. [11] arrives at the same solutions as those obtained by solving local problems (Appendix D.1). Objective (2) is also related to the previous mean-regularized objective [19]. Empirically, their approach does not outperform Objective (2) (Table 4, Appendix). Note that we do not claim that Objective (2) is optimal in terms of the accuracy-fairness-robustness tradeoff. Rather, we propose it as a simple and practical multi-task objective that can improve both robustness and fairness for federated learning.

Algorithm 1 Solver for Multi-Task Learning Objective (2)

Input: K, T, λ, η, w^0, p_k, v_k, $k = 1, \cdots, N$
for $t = 0, \cdots, T-1$ **do**
  Server selects a subset S_t of K devices at random (each device k is chosen with prob. p_k)
  Server sends w^t to all selected devices
  Each selected device k computes the following for some local iterations:

$$w_k^t = w_k^t - \eta \nabla F_k(w_k^t),\ v_k = v_k - \eta(\nabla F_k(v_k) + \lambda(v_k - w_k^t))$$

  Each selected device k sends $\Delta_k^t := w_k^t - w^t$ back to the server
  Server updates w^{t+1} as:

$$w^{t+1} = w^t + \frac{1}{|S_t|} \sum_{k \in S_t} \Delta_k^t$$

end for
return $v_1, v_2, \cdots, v_N$ as personalized models

Other Regularizers. If we intend to enforce the local models to be closer to the optimal global model, there are potentially other choices other than the L_2 term, such as using a Bregman divergence-based regularization or reshaping the L_2 regularization ball using the Fisher information matrix. Under the common logistic loss (which is also what we use for all models in the experiments), the Bregman divergence will reduce to the KL divergence (relative entropy), and its second-order Taylor series expansion will result in an L_2 ball reshaped with the Fisher information matrix. In fact, such regularization is studied in other related contexts like continual learning [25] or multi-task learning [45]. However, in the presence of corrupted data or model poisoning, learning more information from the (potentially corrupted) global model may hurt robustness. As we verify in our experiments (Appendix D.2), incorporating approximate empirical fisher information does not improve the performance, while adding non-trivial computation overheads.

3.2 Solver

To solve Objective (2), there are generally two choices: (i) first solving for w^*, and then for each device k, solving a local objective $\min_{w_k} p_k F_k(w_k) + \frac{\lambda}{2}\|w_k - w^*\|^2$, or (ii) jointly optimizing the global model (to obtain w^*) and solving the local subproblems. Let us not consider the computation aspects for now. Suppose we only care about the final solutions, then these two approaches will arrive at the same solutions in convex cases. In non-convex settings, we observe that there could be additional benefits of using joint optimization—empirically, the updating scheme would guide the optimization trajectory towards a better solution compared with finetuning starting from w^*.

Some previous works argue that early stopping can be provably robust against label noise in training neural networks [26]. Motivated by this, we hypothesize that leveraging the *early* global information (i.e., before the global model converges to w^*) could also be helpful when training personalization models. Intuitively, under data poisoning or model poisoning attacks, the global model may start from a random one and gradually overfit to clean data or corrupted data. Therefore, it might be less appropriate to choose (i). For similar reasons, another natural baseline (finetuning based on local objectives $\min_{w_k} F_k(w_k)$ with w_k starting from w^*) may also underperform the joint optimization method. We compare Algorithm 1 with local finetuning baselines in Section 4.2 and demonstrate that the combination of Objective (2) and its solver is more robust and fair.

4 Experiments

In this section, we first describe our empirical setup, and then we demonstrate the tension between fairness and robustness (Section 4.1). Finally, we compare the performance of the proposed multi-task learning approach in terms of robustness and fairness with several strong baselines (Section 4.2).

Setup. For all experiments, we measure robustness via test accuracies on benign devices, and measure fairness via the test variance (or standard deviation), also across benign devices. We use two federated datasets from a federated learning benchmark [7], and one dataset with a convex model from prior federated learning work [38]. The datasets are provided in Table 2, Appendix C.

4.1 Competing Constraints between Accuracy, Fairness, and Robustness

When training a single global model, fair methods aim to encourage a more uniform performance distribution, but may be highly susceptible to training-time attacks in statistically heterogeneous

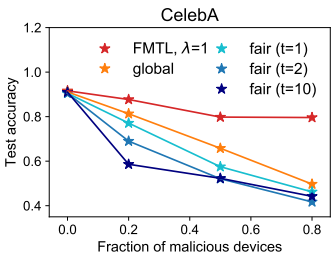

Figure 1: Fair methods can overfit to corrupted devices by imposing more weights on them, thus being particularly susceptible to attacks.

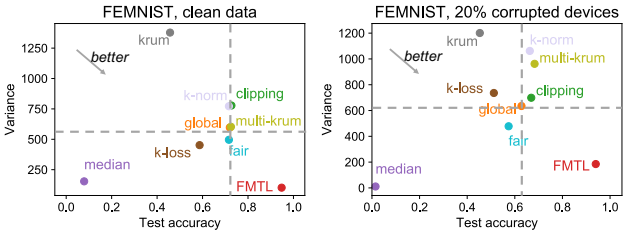

Figure 2: Compared with learning a global model, robust baselines (i.e., the methods listed in the figure excluding 'global' and 'FMTL') are either robust but not fair (with higher accuracy, larger variance), or not even robust (with lower accuracy). Proposed Federated MTL (FMTL) lies at the lower right corner, which is our preferred region.

environments. We investigate the test accuracy on benign devices under three training objectives: training (i) global, (ii) local, and (iii) fair models. The q-FFL [28] objective has been recently proposed for fairness in federated learning; and we use an improved version, TERM [27], as the fair baseline. TERM also recovers AFL [35], another fair FL objective, as a special case. It uses a parameter t to offer different tradeoffs between fairness and accuracy. We take $t = 1, 2, 10$ and perform the simple data poisoning attack of randomly changing training labels on a subset of devices. The results are reported in Figure 1. As the corruption level increases, fitting a global model becomes less robust. Using fair methods will be more susceptible to attacks. When t gets larger, the test accuracy gets lower, which indicates that the fair method overfits more to corrupted devices.

Next, we apply various strong robust methods under the same attack, and explore the robustness/accuracy and fairness tradeoffs. For Krum and Multi-Krum [6], we do our best to favor them—assuming that the central server knows the expected number of malicious devices selected at each communication round. Other robust approaches include: taking the coordinate-wise median of gradients ('median'), gradient clipping ('clipping'), filtering out the gradients with largest norms ('k-norm'), and taking the gradient with the k-th largest loss where k is the number of malicious devices ('k-loss'). From Figure 2, we see that robust baselines are either (i) more robust than global but less fair, or (ii) fail to provide robustness due to heterogeneity.

4.2 Results of the Proposed Federated Multi-Task Learning Objective

We apply three types of attacks to corrupt a randomly-selected subset of the devices. We choose the corruption level until a point where there is a significant performace drop when training a global model. Among all defense mechanisms we compare with, we present the results on three strongest defenses here; and leave the full results to Appendix D.3.

Attack Scenarios. We consider a set of attacks based on different assumptions on the adversaries. **(A1)** We first assume that the corrupted devices do not have access to the training APIs and only the data are poisoned via randomly changing the labels on the training set. For other attacks, we assume byzantine adversaries where the corrupted devices could send arbitrary model updates to the server to attack the global model. **(A2)** One baseline is to send random Gaussian parameters drawn from zero mean with the same variance as the updates of normal training so that it would be more difficult for the server to defend against. **(A3)** Another stronger attack, which is called model replacement in previous work, is to scale the adversarial updates introduced by corrupted data by some constant so that the aggregated model updates will be dominated by the malicious one [2].

Federated MTL is Both Robust and Fair. As shown in Figure 3, the test accuracy on the benign devices under robust baselines can degrade significantly when the number of malicious devices increases, while the proposed objective maintains high accuracies. We also note that some robust methods can outperform the multi-task approach (Table 7 in Appendix). Augmenting multi-task learning with robust aggregators can further improve the robustness performance, which we do not explore in this paper. In Table 1, we compare the proposed objective with global, local, and fair (the TERM objective mentioned before) methods in terms of test accuracies and test standard deviation. When the corruption level is high, 'global' or 'fair' will even fail to converge. Our Federated MTL objective (FMTL) results in more accurate and fair solutions both with and without attacks.

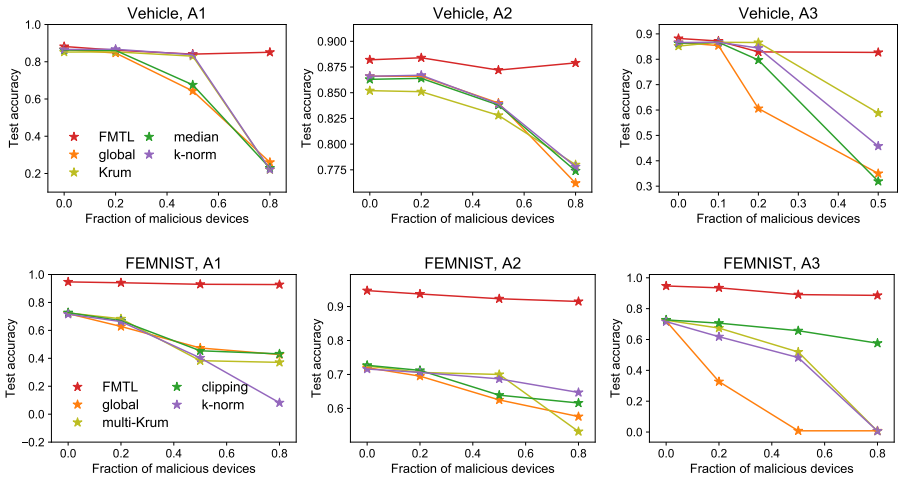

Figure 3: Robustness of multi-task learning on the Vehicle and FEMNIST datasets. We compare multi-task learning with learning a global model and three strongest defense mechanisms among all defense baselines we investigate (see Appendix D.3 for full results on all defense methods), and show that multi-task learning is the most robust under all attacks (i.e., achieving the highest test accuracy on *benign* devices).

Table 1: Test accuracy and standard deviation (numbers in the parentheses) across benign devices on FEMNIST (top) and Vehicle (bottom). The proposed multi-task learning approach is either (i) more fair compared with the baselines of training a global model, or (ii) more accurate than the fair baseline under a set of attacks. We bold the highest worse-cast accuracy (i.e., accuracy mean minus standard deviation) across all methods.

Vehicle		**A1**		**A2**		**A3**	
Methods	clean	20% corrupted	80% corrupted	20% corrupted	80% corrupted	20% corrupted	50% corrupted
global	0.866 (.16)	0.847 (.08)	0.260 (.27)	0.866 (.18)	0.762 (.27)	0.606 (.08)	0.350 (.19)
local	0.836 (.07)	0.835 (.08)	0.857 (.06)	0.835 (.08)	0.857 (.06)	**0.835 (.08)**	**0.840 (.09)**
fair (TERM, t=1)	0.866 (.15)	0.799 (.07)	0.310 (.22)	0.858 (.17)	0.747 (.23)	0.613 (.07)	0.328 (.16)
FMTL	**0.882 (.05)**	**0.862 (.05)**	**0.851 (.06)**	**0.884 (.05)**	**0.879 (.04)**	0.829 (.08)	0.833 (.08)
FEMNIST		**A1**		**A2**		**A3**	
Methods	clean	20% corrupted	80% corrupted	20% corrupted	80% corrupted	10% corrupted	20% corrupted
global	0.720 (.24)	0.628 (.25)	0.427 (.27)	0.695 (.26)	0.576 (.27)	0.327 (.23)	0.008 (.06)
local	0.915 (.18)	0.898 (.18)	0.859 (.23)	0.898 (.19)	**0.859 (.23)**	**0.895 (.15)**	**0.898 (.19)**
fair (TERM, t=1)	0.716 (.22)	0.574 (.22)	0.363 (.24)	0.026 (.06)	0.178 (.14)	0.567 (.24)	0.005 (.11)
FMTL	**0.948 (.10)**	**0.940 (.14)**	**0.933 (.13)**	**0.943 (.13)**	0.822 (.22)	0.794 (.26)	0.752 (.25)

Comparison with Local Finetuning. As mentioned in section 3.2, we consider two local finetuning strategies here: (i) solving $\min_{w_k} F_k(w_k) + \frac{\lambda}{2}\|w_k - w^*\|^2$ for each $k \in [N]$ after solving for w^*, and (ii) directly finetuning on $F_k(w_k)$ for each $k \in [N]$ starting from w^*. The first one is another possible solver for the proposed Objective (2). In realistic federated networks, finetuning can face several practical issues like determining when to stop for each device. Therefore, it is not straightforward to make it scalable and automated. Despite this, in order to obtain the best performance of finetuning, we solve the local problem on each device by running 30 epochs of mini-batch SGD. The results are shown in Figure 4. Both finetuning baselines improve the performance compared with learning a global model, while Objective (2) combined with joint optimization performs the best.

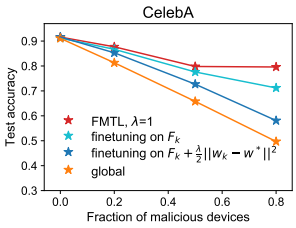

Figure 4: The proposed Objective (2) and Algorithm 1 with joint optimization outperforms two local finetuning baselines.

5 Conclusion and Future Work

In this work, we propose using a multi-task learning objective to address the competing constraints of accuracy, fairness, and robustness in federated learning. In the future, we plan to build on our empirical study and initial theoretical results (Appendix B) to rigorously characterize the benefits multi-task learning in terms of the accuracy-fairness-robustness trade-offs for more general problems.

Acknowledgement

The work of TL, SH, and VS was supported in part by the National Science Foundation grant IIS1838017, a Google Faculty Award, a Carnegie Bosch Institute Research Award, and the CONIX Research Center. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the National Science Foundation or any other funding agency.

References

[1] A. Agarwal, J. Langford, and C.-Y. Wei. Federated residual learning. *arXiv preprint arXiv:2003.12880*, 2020.

[2] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov. How to backdoor federated learning. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[3] A. N. Bhagoji, S. Chakraborty, P. Mittal, and S. Calo. Analyzing federated learning through an adversarial lens. In *International Conference on Machine Learning*, 2019.

[4] B. Biggio, B. Nelson, and P. Laskov. Support vector machines under adversarial label noise. In *Asian Conference on Machine Learning*, 2011.

[5] B. Biggio, B. Nelson, and P. Laskov. Poisoning attacks against support vector machines. In *International Conference on Machine Learning*, 2012.

[6] P. Blanchard, E. M. E. Mhamdi, R. Guerraoui, and J. Stainer. Machine learning with adversaries: Byzantine tolerant gradient descent. In *Advances in Neural Information Processing Systems*, 2017.

[7] S. Caldas, P. Wu, T. Li, J. Konečnỳ, H. B. McMahan, V. Smith, and A. Talwalkar. Leaf: A benchmark for federated settings. *arXiv preprint arXiv:1812.01097*, 2018.

[8] F. Chen, M. Luo, Z. Dong, Z. Li, and X. He. Federated meta-learning with fast convergence and efficient communication. *arXiv preprint arXiv:1802.07876*, 2018.

[9] X. Chen, C. Liu, B. Li, K. Lu, and D. Song. Targeted backdoor attacks on deep learning systems using data poisoning. *arXiv preprint arXiv:1712.05526*, 2017.

[10] G. Cohen, S. Afshar, J. Tapson, and A. van Schaik. Emnist: an extension of mnist to handwritten letters. *arXiv preprint arXiv:1702.05373*, 2017.

[11] Y. Deng, M. M. Kamani, and M. Mahdavi. Adaptive personalized federated learning. *arXiv preprint arXiv:2003.13461*, 2020.

[12] C. T. Dinh, N. H. Tran, and T. D. Nguyen. Personalized federated learning with moreau envelopes. In *Advances in Neural Information Processing Systems*, 2020.

[13] M. F. Duarte and Y. H. Hu. Vehicle classification in distributed sensor networks. *Journal of Parallel and Distributed Computing*, 2004.

[14] J. Dumford and W. Scheirer. Backdooring convolutional neural networks via targeted weight perturbations. *arXiv preprint arXiv:1812.03128*, 2018.

[15] T. Evgeniou and M. Pontil. Regularized multi–task learning. In *International Conference on Knowledge Discovery and Data Mining*, 2004.

[16] A. Fallah, A. Mokhtari, and A. Ozdaglar. Personalized federated learning: A meta-learning approach. In *Advances in Neural Information Processing Systems*, 2020.

[17] M. Fang, X. Cao, J. Jia, and N. Gong. Local model poisoning attacks to byzantine-robust federated learning. In $USENIX$ *Security Symposium*, 2020.

[18] T. Gu, B. Dolan-Gavitt, and S. Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017.

[19] F. Hanzely and P. Richtárik. Federated learning of a mixture of global and local models. *arXiv preprint arXiv:2002.05516*, 2020.

[20] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, 2018.

[21] Z. Hu, K. Shaloudegi, G. Zhang, and Y. Yu. FedMGDA+: Federated learning meets multi-objective optimization. *arXiv preprint arXiv:2006.11489*, 2020.

[22] W. R. Huang, J. Geiping, L. Fowl, G. Taylor, and T. Goldstein. Metapoison: Practical general-purpose clean-label data poisoning. In *Advances in Neural Information Processing Systems*, 2020.

[23] Y. Jiang, J. Konečnỳ, K. Rush, and S. Kannan. Improving federated learning personalization via model agnostic meta learning. *arXiv preprint arXiv:1909.12488*, 2019.

[24] M. Khodak, M.-F. F. Balcan, and A. S. Talwalkar. Adaptive gradient-based meta-learning methods. In *Advances in Neural Information Processing Systems*, 2019.

[25] J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 2017.

[26] M. Li, M. Soltanolkotabi, and S. Oymak. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2020.

[27] T. Li, A. Beirami, M. Sanjabi, and V. Smith. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*, 2020.

[28] T. Li, M. Sanjabi, A. Beirami, and V. Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2020.

[29] P. P. Liang, T. Liu, L. Ziyin, R. Salakhutdinov, and L.-P. Morency. Think locally, act globally: Federated learning with local and global representations. *arXiv preprint arXiv:2001.01523*, 2020.

[30] Y. Liu, S. Ma, Y. Aafer, W. Lee, J. Zhai, W. Wang, and X. Zhang. Trojaning attack on neural networks. In *Network and Distributed System Security Symposium*, 2018.

[31] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, 2015.

[32] H. Mahdavifar, A. Beirami, B. Touri, and J. S. Shamma. Global games with noisy information sharing. *IEEE Transactions on Signal and Information Processing over Networks*, 2017.

[33] Y. Mansour, M. Mohri, J. Ro, and A. T. Suresh. Three approaches for personalization with applications to federated learning. *arXiv preprint arXiv:2002.10619*, 2020.

[34] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2017.

[35] M. Mohri, G. Sivek, and A. T. Suresh. Agnostic federated learning. In *International Conference on Machine Learning*, 2019.

[36] K. Pillutla, S. M. Kakade, and Z. Harchaoui. Robust aggregation for federated learning. *arXiv preprint arXiv:1912.13445*, 2019.

[37] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. In *Advances in Neural Information Processing Systems*, 2018.

[38] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar. Federated multi-task learning. In *Advances in Neural Information Processing Systems*, 2017.

[39] G. Sun, Y. Cong, J. Dong, Q. Wang, and J. Liu. Data poisoning attacks on federated machine learning. *arXiv preprint arXiv:2004.10020*, 2020.

[40] Z. Sun, P. Kairouz, A. T. Suresh, and H. McMahan. Can you really backdoor federated learning? *arXiv preprint arXiv:1911.07963*, 2019.

[41] H. Wang, K. Sreenivasan, S. Rajput, H. Vishwakarma, S. Agarwal, J.-y. Sohn, K. Lee, and D. Papailiopoulos. Attack of the tails: Yes, you really can backdoor federated learning. In *Advances in Neural Information Processing Systems*, 2020.

[42] K. Wang, R. Mathews, C. Kiddon, H. Eichner, F. Beaufays, and D. Ramage. Federated evaluation of on-device personalization. *arXiv preprint arXiv:1910.10252*, 2019.

[43] C. Xie, K. Huang, P.-Y. Chen, and B. Li. DBA: Distributed backdoor attacks against federated learning. In *International Conference on Learning Representations*, 2020.

[44] D. Yin, Y. Chen, R. Kannan, and P. Bartlett. Byzantine-robust distributed learning: Towards optimal statistical rates. In *International Conference on Machine Learning*, 2018.

[45] T. Yu, E. Bagdasaryan, and V. Shmatikov. Salvaging federated learning by local adaptation. *arXiv preprint arXiv:2002.04758*, 2020.

[46] J. Zhang, C. Li, A. Robles-Kelly, and M. Kankanhalli. Hierarchically fair federated learning. *arXiv preprint arXiv:2004.10386*, 2020.

Appendix: Analysis of the Proposed Multi-Task Learning Objective

Here, we provide a theoretical analysis of the proposed multi-task learning objective (Objective (2)), mainly on a simplified problem of federated point estimation. While our analysis in its current form does not cover more general convex or even non-convex functions, on this toy problem, we theoretically investigate the benefits of Objective (2) in terms of test accuracy, fairness, and robustness, which helps to motivate the use of the proposed objective.

We first present some properties regarding Objective (2), which servers as a first step towards understanding the solutions of the proposed objective.

A Properties of the Multi-Task Learning Objective

Let the multi-task learning objective on device k be

$$g_k(w) = f_k(w) + \lambda\psi(w), \tag{3}$$

where f_k is strongly convex, and

$$\psi(w) := \frac{1}{2}\|w - w^*\|^2, \tag{4}$$

$$w^* := \arg\min_w \left\{ \frac{1}{N} \sum_{k\in[N]} f_k(w) \right\}. \tag{5}$$

Let

$$\widehat{w}_k(\lambda) = \arg\min_w g_k(w). \tag{6}$$

Without any distributional assumptions on the tasks, we first characterize the solutions of the objective $g_k(w)$.

Lemma 1. *For all $\lambda \geqslant 0$,*

$$\frac{\partial}{\partial\lambda} f_k(\widehat{w}_k(\lambda)) \geqslant 0, \tag{7}$$

$$\frac{\partial}{\partial\lambda} \psi(\widehat{w}_k(\lambda)) \leqslant 0. \tag{8}$$

In addition, for all k, if $f_k(w^*)$ is finite, then

$$\lim_{\lambda\to\infty} \widehat{w}_k(\lambda) = w^*. \tag{9}$$

Proof. The proof here directly follows the proof in Hanzely and Richtárik [Theorem 3.1, 19]. □

For the proposed multi-task objective, as λ increases, the local training loss on f_k will also increase, and the resulting personalized models will be closer to the global model. Therefore, λ effectively controls how much personalization we impose. Since training loss is minimized when $\lambda = 0$, training separate local models is the most robust and fair when we *do not consider generalization*.

However, in order to obtain the guarantees on the test performance, we need to explicitly model the joint distributions of data on all devices. In the next section, we explore a Bayesian framework on a point estimation problem to examine the generalization, fairness, and robustness of the proposed multi-task objective, all on test data.

B Federated Point Estimation

For the federated point estimation problem, we first examine the case without corrupted devices in Section B.1. We prove that there exists a λ that results in an optimal average test performance *and* optimal fairness across all devices. When there are adversaries, we analyze the robustness benefits of our multi-task learning objective in Section B.2. In particular, we show there exists a λ which leads

to the highest test accuracy across benign devices (i.e., the most robust) and minimizes the variance of the test error across benign devices (i.e., the most fair) jointly.

Before we proceed, we first state a technical lemma that will be used throughout the analysis. The proof can be found in Mahdavifar et al. [32].

Lemma 2 (Lemma 11, Mahdavifar et al. [32])**.** *Let θ be drawn from the non-informative uniform prior on $\mathbb{R}$. Further, let $\{\phi_k\}_{k\in[K]}$ denote noisy observations of θ with additive zero-mean independent Gaussian noises with variances $\{\sigma_k^2\}_{k\in[K]}$. Let*

$$\frac{1}{\sigma_\theta^2} := \sum_{k\in[K]} \frac{1}{\sigma_k^2}. \tag{10}$$

Then, conditioned on $\{\phi_k\}_{k\in[K]}$, we can write θ as

$$\theta = \sigma_\theta^2 \sum_{k\in[K]} \frac{\phi_k}{\sigma_k^2} + z,$$

where z is $\mathcal{N}(0,\sigma_\theta^2)$ which is independent of $\{\phi_k\}_{k\in[K]}$.

B.1 No Adversaries: Multi-Task Learning for Accuracy and Fairness

We consider a Bayesian framework. Let θ be drawn from the non-informative prior on $\mathbb{R}$, i.e., uniformly distributed on $\mathbb{R}$.[2] We assume that K devices have their data distributed with parameters $\{w_k\}_{k\in[K]}$.

$$w_k = \theta + \zeta_k, \tag{11}$$

where $\zeta_k \sim \mathcal{N}(0,\tau^2)$ are i.i.d. τ controls the degree of dependence between the tasks on different devices. If $\tau = 0$, then the data on all devices is distributed according to parameter θ, i.e., the tasks are the same, and if $\tau \to \infty$, the tasks on different devices become completely independent.

Let each device have n data points[3] denoted by $\mathbf{x}_k = \{x_{k,1},\dots,x_{k,n}\}$, such that

$$x_{k,i} = w_k + z_{k,i}, \tag{12}$$

where $z_{k,i} \sim \mathcal{N}(0,\sigma^2)$ and are i.i.d.

Assume that

$$f_k(w) = \frac{1}{2}\left\| w - \frac{1}{n}\sum_{i\in[n]} x_{k,i} \right\|_2^2, \tag{13}$$

and denote by $\widehat{w}_k$ the minimizer of f_k. It is clear that

$$\widehat{w}_k = \frac{1}{n}\sum_{i\in[n]} x_{k,i}. \tag{14}$$

Further, let

$$w^* := \arg\min_w \left\{ \frac{1}{K}\sum_{k\in[K]} f_k(w) \right\}. \tag{15}$$

It is straightforward calculation to verify that

$$w^* = \frac{1}{nK}\sum_{i\in[n]}\sum_{k\in[K]} x_{k,i} = \frac{1}{K}\sum_{k\in[K]} \widehat{w}_k. \tag{16}$$

Let the objective at each device be

$$g_k(w) = f_k(w) + \lambda\psi(w) \tag{17}$$

where

$$\psi(w) := \frac{1}{2}\|w - w^*\|^2, \tag{18}$$

[2] This is an improper prior that makes our calculations simpler and the interpretations nicer.

[3] For ease of notations, we assume that each device has the same number of data points. It is straightforward to extend the current analysis to handle varying number of samples per device.

Lemma 3. *Denote by $\widehat{w}_k(\lambda)$ the minimizer of g_k. Then,*

$$\widehat{w}_k(\lambda) = \frac{\lambda}{1+\lambda} w^* + \frac{1}{1+\lambda}\widehat{w}_k \tag{19}$$

$$= \frac{\lambda}{(1+\lambda)K}\sum_{j\neq k}\widehat{w}_j + \frac{K+\lambda}{(1+\lambda)K}\widehat{w}_k. \tag{20}$$

Let

$$\sigma_n^2 := \frac{\sigma^2}{n}. \tag{21}$$

$$\widehat{w}^{K\setminus k} := \frac{1}{K-1}\sum_{j\neq k}\widehat{w}_j \tag{22}$$

Lemma 4. *Given observations $\widehat{w}^{K\setminus k}$ and $\widehat{w}_k$, w_k is Gaussian distributed and given by*

$$w_k = \frac{\sigma_w^2}{\sigma_n^2}\widehat{w}_k + \frac{(K-1)\sigma_w^2}{K\tau^2+\sigma_n^2}\widehat{w}^{K\setminus k} + \xi, \tag{23}$$

where

$$\frac{1}{\sigma_w^2} = \frac{1}{\sigma_n^2} + \frac{K-1}{K\tau^2+\sigma_n^2}, \tag{24}$$

and

$$\xi \sim \mathcal{N}\left(0, \sigma_w^2\right). \tag{25}$$

Proof. $\widehat{w}_k$ is a noisy observation of w_k with additive zero-mean independent Gaussian noise with variance σ_n^2. Given θ, $\widehat{w}^{K\setminus k}$ is Gaussian with mean θ and variance $\frac{\tau^2+\sigma_n^2}{K-1}$. By symmetry, given observations $\widehat{w}^{K\setminus k}$, θ is Gaussian with variance $\frac{\tau^2+\sigma_n^2}{K-1}$. Therefore, w_k is Gaussian with variance $\frac{\tau^2+\sigma_n^2}{K-1} + \tau^2 = \frac{K\tau^2+\sigma_n^2}{K-1}$. $\widehat{w}^{K\setminus k}$ can be viewed as a noisy observation of w_k with Gaussian noise with variance $\frac{K\tau^2+\sigma_n^2}{K-1}$. The proof then follows by directly invoking Lemma 2. □

Theorem 1. *Let λ^* be the optimal λ that minimizes the test performance, i.e.,*

$$\lambda^* = \arg\min_{\lambda} E\left\{ (w_k - \widehat{w}_k(\lambda))^2 \middle| \widehat{w}^{K\setminus k}, \widehat{w}_k \right\}. \tag{26}$$

Then,

$$\lambda^* = \frac{\sigma_n^2}{\tau^2} = \frac{\sigma^2}{n\tau^2}. \tag{27}$$

Proof. The proof follows by inspecting Lemma 4, and observing that λ^* leads to the MMSE estimator of w_k given $\widehat{w}^{K\setminus k}$, $\widehat{w}_k$, i.e., $\widehat{w}_k(\lambda^*)$ is the MMSE estimator of w_k given the observations $\widehat{w}^{K\setminus k}$ and $\widehat{w}_k$. □

Remark 1. *We note that by using λ^* in Objective* (2)*, we not only achieve the most accurate solution for the objective, but also we achieve the most accurate solution of any possible federated point estimation algorithm in this problem, as Objective* (2) *with λ^* realizes the MMSE estimator for w_k.*

We have derived an optimal $\lambda^* = \frac{\sigma^2}{n\tau^2}$ for multi-task learning in terms of generalization. Recall that we define fairness as the variance of the performance across all devices [20]. Next, we prove that the same λ^* that minimizes the expected MSE also achieves the optimal fairness.

Theorem 2. *Among all possible solutions of Objective* (2) *parameterized by λ, λ^* results in the most fair performance across all devices when there are no adversaries, i.e., it minimizes the variance of test performance (test mean square error) across all devices.*

Proof. Let

$$\widehat{E}_K\left\{a_k\right\} := \frac{1}{K}\sum_{k\in[K]} a_k \tag{28}$$

Denote the variance of test performance across K devices as $\text{var}_K\left\{(w_k - \widehat{w}_k(\lambda))^2\right\}$. Then,

$$\text{var}_K\left\{(w_k - \widehat{w}_k(\lambda))^2\right\} = \widehat{E}_K\left\{(w_k - \widehat{w}_k(\lambda))^4\right\} - \left(\widehat{E}_K\left\{(w_k - \widehat{w}_k(\lambda))^2\right\}\right)^2. \tag{29}$$

Also, notice that

$$w_k - \widehat{w}_k(\lambda) = \xi_k + a_k \tag{30}$$

where

$$a_k = \widehat{w}_k(\lambda) - \widehat{w}_k(\lambda^*), \tag{31}$$

and $\lambda^* = \frac{\sigma^2}{n\tau^2}$, which is the λ yielding the optimal generalization (Equation (27)).

By expanding $\text{var}_K\left\{(w_k - \widehat{w}_k(\lambda))^2\right\}$, we have

$$E\left\{\text{var}_K\left\{(w_k - \widehat{w}_k(\lambda))^2\right\}\middle|\, \widehat{w}^{K\setminus k}, \widehat{w}_k\right\} \tag{32}$$

$$= E\left\{\widehat{E}_K\left\{(w_k - \widehat{w}_k(\lambda))^4\right\} - \left(\widehat{E}_K\left\{(w_k - \widehat{w}_k(\lambda))^2\right\}\right)^2\middle|\, \widehat{w}^{K\setminus k}, \widehat{w}_k\right\} \tag{33}$$

$$= E\left\{\widehat{E}_K\left\{(\xi_k + a_k)^4\right\} - \left(\widehat{E}_K\left\{(\xi_k + a_k)^2\right\}\right)^2\middle|\, \widehat{w}^{K\setminus k}, \widehat{w}_k\right\} \tag{34}$$

$$= E\left\{\widehat{E}_K\left\{\xi_k^4 + 6\xi_k^2 a_k^2 + a_k^4\right\} - \left(\widehat{E}_K\left\{\xi_k^2 + a_k^2\right\}\right)^2\middle|\, \widehat{w}^{K\setminus k}, \widehat{w}_k\right\} \tag{35}$$

$$= E\left\{\widehat{E}_K\left\{\xi_k^4 + 6\xi_k^2 a_k^2 + a_k^4\right\} - \left(\widehat{E}_K\left\{\xi_k^2\right\}\right)^2 - 2\widehat{E}_K\left\{\xi_k^2\right\}\widehat{E}_K\left\{a_k^2\right\} - \left(\widehat{E}_K\left\{a_k^2\right\}\right)^2\middle|\, \widehat{w}^{K\setminus k}, \widehat{w}_k\right\} \tag{36}$$

$$= 3\sigma_w^4 + 6\sigma_w^2\widehat{E}_K\left\{a_k^2\right\} + \widehat{E}_K\left\{a_k^4\right\} - \sigma_w^4 - 2\sigma_w^2\widehat{E}_K\left\{a_k^2\right\} - \left(\widehat{E}_K\left\{a_k^2\right\}\right)^2 \tag{37}$$

$$= 2\sigma_w^4 + 4\sigma_w^2\widehat{E}_K\left\{a_k^2\right\} + \widehat{E}_K\left\{a_k^4\right\} - \left(\widehat{E}_K\left\{a_k^2\right\}\right)^2, \tag{38}$$

where we have used the fact that we can swap expectations, and $E\{\xi_k^4\} = 3\sigma_w^4$, given that ξ is Gaussian distributed.

Inspecting Equation (38), we can see that it is minimized if $\widehat{E}_K\left\{a_k^2\right\} = 0$ or $a_k = 0$ for all $k \in [K]$. Notice that

$$\widehat{E}_K\left\{a_k^4\right\} - \left(\widehat{E}_K\left\{a_k^2\right\}\right)^2 \geqslant 0 \tag{39}$$

with equality if and only if $a_k = a_j$ for all $k, j \in [K]$. Hence, $a_k = 0$ is the minimizer of the variance. □

Observations. From the optimal $\lambda^* = \frac{\sigma^2}{n\tau^2}$ for mean test accuracy and variance of the test accuracy, we have the following observations.

- Test error and variance can be jointly minimized with one λ.
- As $n \to \infty$, $\lambda^* \to 0$, i.e., when each local device has an infinite number of samples, there is no need for federated learning, and training completely local models is optimal in terms of generalization and fairness.
- As $\tau \to \infty$, $\lambda^* \to 0$, i.e., if the data on different devices (the tasks) are unrelated, then training local models is optimal; On the other hand, as $\tau \to 0$, $\lambda^* \to \infty$, i.e., if the data across all devices are identically distributed, or equivalently if the tasks are the same, then training a global model is the best we can achieve.

So far we have proved that the same λ^* achieves the best performance (expected mean square error) for any device k *and* fairness (variance of mean square error) without considering adversaries. In Section B.2 below, we analyze the benefit of multi-task learning for fairness and robustness in the presence of adversaries.

B.2 With Adversaries: Multi-Task Learning for Accuracy, Fairness, and Robustness

To reason about the behavior of multi-task learning for robustness, we make the following assumptions on the adversaries.

Let K_a and $K_b \geqslant 1$ denote the number of adversarial and benign devices, respectively, such that $K = K_a + K_b$.

Definition 1. *We say that a device k is a benign device if $w_k \sim \theta + \mathcal{N}(0, \tau^2)$; and we say a device k is a malicious device (or an adversary) if $w_k \sim \theta + \mathcal{N}(0, \tau_a^2)$ where $\tau_a \geqslant \tau$.*

As mentioned in the introduction, in the presence of adversaries, we measure fairness as the performance variance on *benign* devices, and robustness as the performance mean across *benign* devices. We next characterize the benefits of multi-task learning under such definition.

Lemma 5. *Let w_k be the parameter associated with a benign device. Given observations $\widehat{w}^{K\backslash k} := \frac{1}{K-1}\sum_{j\neq k}\widehat{w}_j$ and $\widehat{w}_k$, w_k is Gaussian distributed and given by*

$$w_k = \frac{\sigma_{w,a}^2}{\sigma_n^2}\widehat{w}_k + \frac{(K-1)\sigma_{w,a}^2}{K\tau^2 + \sigma_n^2 + \frac{K_a}{K-1}(\tau_a^2 - \tau^2)}\widehat{w}^{K\backslash k} + \xi_a, \tag{40}$$

where

$$\frac{1}{\sigma_{w,a}^2} = \frac{1}{\sigma_n^2} + \frac{K-1}{K\tau^2 + \sigma_n^2 + \frac{K_a}{K-1}(\tau_a^2 - \tau^2)}, \tag{41}$$

and

$$\xi_a \sim \mathcal{N}\left(0, \sigma_{w,a}^2\right). \tag{42}$$

Proof. The proof is the same as the proof for Lemma 4, except that we note that the variance of Gaussian distributed θ given observations $\widehat{w}^{K\backslash k}$ is $\frac{(\tau^2+\sigma_n^2)(K-K_a-1)+(\tau_a^2+\sigma_n^2)K_a}{(K-1)^2} = \frac{\tau^2+\sigma_n^2+\frac{K_a}{K-1}(\tau_a^2-\tau^2)}{K-1}$. □

Theorem 3. *Let w_k be a benign device. Let λ_a^* be the optimal λ that minimizes the test performance, i.e.,*

$$\lambda_a^* = \arg\min_{\lambda} E\left\{ (w_k - \widehat{w}_k(\lambda))^2 \middle| \widehat{w}^{K\backslash k}, \widehat{w}_k \right\}. \tag{43}$$

Then,

$$\lambda_a^* = \frac{\sigma^2}{n}\frac{K}{K\tau^2 + \frac{K_a}{K-1}(\tau_a^2 - \tau^2)}. \tag{44}$$

Proof. We obtain λ_a^* following the proof of Theorem 1. □

Theorem 4. *Among all solutions of Objective* (2) *parameterized by λ, λ_a^* results in the most fair performance across all benign devices, i.e., it minimizes the variance of test performance (test mean square error) on benign devices.*

Proof. Similarly, we look at the variance of the test mean square error across benign devices:

$$\text{var}_{K_b}\left\{(w_k - \widehat{w}_k(\lambda))^2\right\} = \widehat{E}_{K_b}\left\{(w_k - \widehat{w}_k(\lambda))^4\right\} - \left(\widehat{E}_{K_b}\left\{(w_k - \widehat{w}_k(\lambda))^2\right\}\right)^2. \tag{45}$$

The rest of the proof is the same as the proof of Theorem 2, except that we set $a_k = \widehat{w}_k(\lambda) - \widehat{w}_k(\lambda_a^*)$. □

Remark 2. *For any benign device k, the solution we obtain by solving Objective* (2) *with λ_a^* is the most robust solution one could obtain among any federated point estimation method given observations $\widehat{w}_k$ and $\widehat{w}^{K\backslash k}$. λ_a^* also results in a most fair model in the solution space of Objective* (2)*.*

Lemma 6. *The expected test error minimized at λ_a^* is $\sigma_{w,a}^2$; and the variance of the test performance minimized at λ_a^* is $2\sigma_{w,a}^4$.*

Proof. For the expected test performance, we note that

$$E\left\{(w_k - \widehat{w}_k(\lambda_a^*))^2 \middle| \widehat{w}^{K\setminus k}, \widehat{w}_k\right\} = E[\xi_a^2] = \sigma_{w,a}^2. \tag{46}$$

For variance, as $a_k = 0$ if $\lambda = \lambda_a^*$, we get

$$\text{var}_{K_b}\left\{(w_k - \widehat{w}_k(\lambda_a^*))^2\right\} = 2\sigma_{w,a}^4 + 4\sigma_{w,a}^2 \widehat{E}_K\left\{a_k^2\right\} + \widehat{E}_K\left\{a_k^4\right\} - \left(\widehat{E}_K\left\{a_k^2\right\}\right)^2 = 2\sigma_{w,a}^4. \tag{47}$$

□

Observations. From λ_a^*, we have the following interesting observations.

- Mean test error (performance, or robustness) and variance of the performance across benign devices (fairness) can still be minimized with the same λ_a in the presence of adversaries.
- As $\tau_a \to \infty$, $\lambda_a^* \to 0$, i.e., training local models is optimal in terms of robustness and fairness when adversary's task may be arbitrarily far from the the task in the benign devices.
- As $\tau \to 0$, if $\tau_a > 0$, $\lambda_a^* < \infty$, which means that learning a global model is *not* optimal even with homogeneous data in the presence of adversaries.
- λ_a^* is a decreasing function of the number (K_a) and the capability (τ_a) of the corrupted devices. In other words, as the attacks become more adversarial, we need more personalization.
- The smallest test error is $\sigma_{w,a}^2$, and the optimal variance is $2\sigma_{w,a}^4$, which are both increasing with K_a (number of adversarial devices) or τ_a (the power of adversary) by inspecting (41). This reveals a fundamental trade-off between fairness and robustness.

Discussions. Through our analysis, we prove that multi-task learning with an appropriate λ is more accurate, robust, and fair compared with training global or local models. We provide closed-form solutions for λ^* across different settings (with and without adversaries), and show that multi-task learning allows to achieve a favorable tradeoff between fairness and robustness. In the future, we plan to generalize the current theoretical framework to linear models and more general convex models.

C Experimental Details

We summarize the datasets, the corresponding models, and tasks in the table below. FEMNIST is Federated EMNIST, which is EMNIST partitioned by the writers of digits/characters created by a previous federated learning benchmark [7].

Table 2: Summary of datasets.

Datasets	**Data Partition**	**Models**	**Tasks**
Vehicle [13] (23 devices)[4]	natural (each device is a vehicle)	linear SVM	binary classification
FEMNIST [7, 10] (100 devices)	synthetic (assign 5 classes to each device)	CNN	62-class classification
CelebA [31] (515 devices)	natural (each device is a celebrity)	CNN	binary classification

D Full Results

D.1 Results in Convex Settings

To sanity check if APFL is doing local minimization for any $\alpha \in (0, 1)$, we test its performance on another downsampled version of Vehicle. The results are shown in Table 3 below.

Table 3: The interpolation-based approach (APFL in Deng et al. [11]) is solving local subproblems for any interpolation parameter $\alpha \neq 0$—generating the same results as the local training baseline. The proposed objective in this work (FMTL) can improve accuracy and fairness under clean data (first column), and is also robust under different attaks.

Vehicle		**A1**			**A3**		
Methods	clean	20% corrupted	50% corrupted	80% corrupted	10% corrupted	15% corrupted	20% corrupted
global	0.821 (.21)	0.773 (.21)	0.803 (.16)	0.214 (.11)	0.275 (.22)	0.237 (.10)	0.086 (.09)
local	0.791 (.14)	0.795 (.14)	0.792 (.12)	**0.829** (.07)	0.795 (.14)	0.792 (.12)	0.829 (.07)
FMTL, λ=0.1	0.814 (.14)	0.812 (.15)	0.798 (.14)	0.757 (.10)	0.742 (.15)	0.751 (.13)	0.757 (.11)
FMTL, λ=1	0.841 (.12)	**0.852** (.15)	**0.849** (.15)	0.814 (.11)	0.803 (.15)	0.786 (.13)	0.828 (.11)
FMTL, λ=2	**0.847** (.13)	0.843 (.13)	0.826 (.15)	0.829 (.09)	**0.812** (.16)	0.775 (.14)	**0.857** (.07)
APFL, α=0.3	0.803 (.14)	0.799 (.14)	0.792 (.14)	0.828 (.07)	0.795 (.14)	0.798 (.12)	0.828 (.08)
APFL, α=0.5	0.798 (.14)	0.790 (.14)	0.786 (.13)	**0.829** (.07)	0.795 (.14)	**0.803** (.13)	0.843 (.10)
APFL, α=0.8	0.785 (.14)	0.795 (.14)	0.798 (.12)	**0.829** (.07)	0.790 (.13)	0.798 (.13)	0.842 (.10)

[4] `http://www.ecs.umass.edu/~mduarte/Software.html`

D.2 Compare with Other Multi-Task Learning Objectives

Table 4: Compare Objective (2) using the L_2 regularizer with other multi-task learning objectives: (i) L2SGD which regularizes local models towards their mean [19], and (ii) using Fisher information when enforcing local models towards the optimal global model [25, 45]. The local objective in such case is $\min_w F_k(w) + \frac{\lambda}{2}\sum_i \mathbf{F}_{ii} \cdot (w[i] - w^*[i])^2$ where $[i]$ denotes the index of parameters and $\mathbf{F}_{ii}$ denotes the i-th diagonal of the Fisher information matrix $\mathbf{F}$. The proposed objective (FMTL with L_2) performs better than both alternatives. In the presence of noisy data, using the Fisher information at the current (possibly corrupted) global model may not improve the performance.

FEMNIST		**A1**
Methods	clean	50% corrupted
global	0.720 (.24)	0.474 (.30)
L2SGD, best λ	0.918 (.15)	0.914 (.17)
EWC, best λ	0.935 (.16)	0.925 (.23)
FMTL, best λ	**0.947** (.15)	**0.930** (.20)

CelebA		**A1**
Methods	clean	50% corrupted
global	0.911 (.19)	0.538 (.28)
L2SGD, best λ	0.899 (.18)	0.725 (.25)
EWC, best λ	0.910 (.18)	0.642 (.26)
FMTL, best λ	**0.921** (.16)	**0.735** (.26)

D.3 Complete Results

In Section 4.2, we present partial results on three strongest attacks, and on a subset of federated data. Here, we provide full results indicating the robustness and fairness of the proposed approach on all attacks, all defenses, and all datasets. The numbers in the parentheses are test accuracy standard deviation across all devices. We bold the numbers with the highest worst-case accuracy (average accuracy minus standard deviation).

Table 5: Full results on Vehicle.

Vehicle		**A1**		**A2**		**A3**	
Methods	clean	20% corrupted	80% corrupted	20% corrupted	80% corrupted	20% corrupted	50% corrupted
global	0.866 (.16)	0.847 (.08)	0.260 (.27)	0.866 (.18)	0.762 (.27)	0.606 (.08)	0.350 (.19)
local	0.836 (.07)	0.835 (.08)	**0.857 (.09)**	0.835 (.08)	0.857 (.09)	0.835 (.08)	**0.840 (.09)**
fair	0.866 (.15)	0.799 (.07)	0.310 (.22)	0.858 (.17)	0.747 (.23)	0.613 (.07)	0.328 (.16)
median	0.863 (.16)	0.861 (.18)	0.229 (.31)	0.864 (.18)	0.774 (.28)	0.797 (.07)	0.319 (.17)
Krum	0.852 (.17)	0.853 (.19)	0.221 (.32)	0.851 (.19)	0.780 (.31)	**0.866 (.18)**	0.588 (.14)
multi-Krum	0.866 (.16)	0.867 (.18)	0.220 (.32)	0.867 (.18)	0.770 (.31)	0.836 (.08)	0.406 (.15)
clipping	0.864 (.16)	0.865 (.17)	0.234 (.30)	0.865 (.18)	0.764 (.27)	0.789 (.07)	0.315 (.17)
k-norm	0.866 (.16)	0.867 (.17)	0.222 (.32)	0.867 (.18)	0.778 (.31)	0.844 (.09)	0.458 (.16)
k-loss	0.850 (.05)	0.755 (.03)	0.217 (.31)	0.852 (.06)	0.825 (.09)	0.692 (.08)	0.328 (.16)
FMTL, λ=0.1	0.845 (.07)	0.841 (.08)	0.851 (.06)	0.844 (.07)	0.866 (.05)	0.829 (.08)	0.827 (.08)
FMTL, λ=1	0.875 (.05)	0.859 (.06)	0.776 (.08)	0.875 (.06)	**0.879 (.04)**	0.813 (.07)	0.757 (.08)
FMTL, λ=2	**0.882 (.05)**	**0.862 (.05)**	0.709 (.12)	**0.884 (.05)**	0.869 (.04)	0.791 (.06)	0.690 (.09)

Table 6: Full results on FEMNIST.

FEMNIST		**A1**		**A2**		**A3**	
Methods	clean	20% corrupted	80% corrupted	20% corrupted	80% corrupted	10% corrupted	20% corrupted
global	0.720 (.24)	0.628 (.25)	0.427 (.27)	0.695 (.26)	0.576 (.27)	0.327 (.23)	0.008 (.06)
local	0.915 (.18)	0.898 (.18)	0.859 (.23)	0.898 (.19)	0.859 (.23)	0.895 (.15)	**0.898 (.19)**
fair	0.716 (.22)	0.574 (.22)	0.363 (.24)	0.026 (.06)	0.178 (.14)	0.567 (.24)	0.005 (.11)
median	0.079 (.12)	0.015 (.03)	0.011 (.05)	0.096 (.13)	0.153 (.13)	0.090 (.10)	0.090 (.10)
Krum	0.457 (.37)	0.453 (.35)	0.003 (.04)	0.459 (.38)	0.088 (.19)	0.290 (.31)	0.532 (.32)
multi-Krum	0.725 (.25)	0.683 (.31)	0.371 (.29)	0.706 (.29)	0.532 (.30)	0.674 (.28)	0.006 (.05)
clipping	0.727 (.28)	0.669 (.26)	0.432 (.29)	0.712 (.28)	0.616 (.26)	0.706 (.28)	0.576 (.29)
k-norm	0.716 (.28)	0.663 (.33)	0.082 (.20)	0.706 (.28)	0.647 (.36)	0.618 (.28)	0.006 (.05)
k-loss	0.587 (.21)	0.512 (.27)	0.004 (.11)	0.545 (.26)	0.343 (.30)	0.003 (.08)	0.006 (.05)
FMTL, λ=0.01	0.947 (.15)	**0.941 (.15)**	0.928 (.20)	0.937 (.17)	**0.915 (.19)**	**0.935 (.15)**	0.886 (.20)
FMTL, λ=0.1	**0.948 (.10)**	**0.940 (.14)**	**0.933 (.13)**	**0.943 (.13)**	0.822 (.22)	0.794 (.26)	0.752 (.25)
FMTL, λ=1	0.902 (.15)	0.895 (.12)	0.854 (.20)	0.891 (.19)	0.591 (.35)	0.672 (.31)	0.609 (.33)

Table 7: Full results on CelebA.

CelebA		**A1**		**A2**		**A3**	
Methods	clean	20% corrupted	80% corrupted	20% corrupted	80% corrupted	10% corrupted	20% corrupted
global	0.911 (.19)	0.813 (.22)	0.497 (.27)	0.901 (.19)	0.847 (.21)	0.537 (.33)	0.539 (.33)
local	0.692 (.27)	0.690 (.27)	0.681 (.26)	0.690 (.27)	0.681 (.26)	0.692 (.27)	0.690 (.27)
fair	0.905 (.17)	0.690 (.26)	0.417 (.26)	0.768 (.25)	0.707 (.27)	0.537 (.33)	0.539 (.33)
median	0.910 (.18)	0.876 (.22)	0.474 (.24)	0.894 (.19)	0.860 (.21)	0.905 (.18)	**0.885 (.20)**
Krum	0.775 (.25)	0.456 (.33)	0.459 (.33)	0.565 (.32)	0.534 (.29)	0.777 (.26)	0.734 (.26)
multi-Krum	0.911 (.18)	**0.898 (.19)**	0.523 (.32)	0.904 (.19)	0.767 (.27)	0.555 (.29)	0.514 (.27)
clipping	0.909 (.18)	0.890 (.17)	0.479 (.27)	0.909 (.18)	0.868 (.23)	**0.908 (.17)**	0.879 (.21)
k-norm	0.908 (.18)	0.898 (.20)	0.534 (.10)	0.907 (.20)	**0.886 (.20)**	0.778 (.24)	0.684 (.26)
k-loss	0.873 (.19)	0.675 (.29)	0.455 (.29)	0.856 (.22)	0.876 (.21)	0.538 (.33)	0.539 (.10)
FMTL, λ=0.1	0.884 (.24)	0.844 (.27)	0.792 (.27)	0.875 (.25)	0.856 (.25)	0.701 (.27)	0.680 (.29)
FMTL, λ=1	0.916 (.17)	0.877 (.21)	**0.796 (.23)**	0.907 (.17)	0.864 (.20)	0.654 (.28)	0.654 (.29)
FMTL, λ=2	**0.921 (.16)**	0.856 (.21)	0.766 (.25)	**0.910 (.17)**	0.852 (.20)	0.612 (.32)	0.598 (.31)