Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach

Cuong Tran †, Ferdinando Fioretto †, Pascal Van Hentenryck‡

†Department of Electrical Engineering and Computer Science, Syracuse University

‡School of Industrial and Systems Engineering, Georgia Tech

Motivation

- Anti-discrimination laws require AI systems to be fair w.r.t gender, races, ages,...
- But due to various reasons (bias on training data, historical bias), some learning models might be discriminative, e.g a credit score system is likely to approve loan applications from men than women.
- Consequently, **fair** learning models have been proposed. To build such models, sensitive group information(gender, age,..) of training data needed to be collect. Then a constrained learning framework can be applied, to ensure fairness.
- But, the privacy issues can arise here. Public **fair** models can reveal these sensitive information of the training data they were trained on!
- Our proposed work *PF-LD* [1] addresses how to keep confidential information (gender, race,..) of training data when training a fair model.

Problem Settings

Given a training data $D = \{X_i, A_i, Y_i\}_{i=1}^n$, X_i is non-sensitive feature, $Y_i \in \{0,1\}$ is binary label, A_i is sensitive information (e.g gender or race info). A classifier $\mathcal{M}_\theta(X)$ can be found by minimizing the following empirical loss:

$$\min_\theta J(\mathcal{M}_\theta, D) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\mathcal{M}_\theta(X_i), Y_i), \tag{1}$$

where $\mathcal{L}(\mathcal{M}_\theta(X_i), Y_i)$ measures the mis-match between model's prediction $\mathcal{M}_\theta(X_i)$ with its ground-truth Y_i.

Fairness Definition

Given input feature X, its sensitive information A, and model prediction $\mathcal{M}(X)$, we consider three group fairness definitions [2].

1. *Demographic Parity*: $\mathcal{M}(X) \perp\!\!\!\perp A$, or $\mathbb{E}[(X)|A=a] = \mathbb{E}[\mathcal{M}(X)]\ \forall a$
2. *Equalized Odds*: $\mathcal{M}(X) \perp\!\!\!\perp A|Y$, or $\mathbb{E}[\mathcal{M}(X)|A=a, Y=y] = \mathbb{E}[\mathcal{M}(X)|Y=y]\ \forall a, y$
3. *Accuracy Parity*: $\mathbb{E}[\mathcal{M}(X) \neq Y|A=a] = \mathbb{E}[\mathcal{M}(X) \neq Y]\ \forall a$

Given one of these fairness definition, the **fair** classifier can be learned constrainting the traditional learning in Equation 1:

$$\operatorname*{argmin}_\theta J(\mathcal{M}_\theta, D) \tag{2a}$$

$$\text{s.t:} \quad \mu(D_P) - \mu(D_G) = \mathbf{0}^\top. \tag{2b}$$

The constraints 2b can capture any of the above three fairness constraints, i.e the group/sub-population statistics $\mu(D_G)$ (e.g women or men) should be similar to the full population statistics $\mu(D_P)$.

Privacy Definition

We consider the notation of differential privacy(DP) here:

Definition 1 (Differential Privacy). *A randomized mechanism $\mathcal{K}: \mathcal{D} \to \mathcal{R}$ is (ϵ, δ)-differentially private (DP) if, for two adjacent datasets D, D', where D' is obtained from D by changing the sensitive information of one user in D, and for any subset of output responses $R \subseteq \mathcal{R}$:*

$$\Pr[\mathcal{K}(D) \in R] \leq e^\epsilon \Pr[\mathcal{K}(D') \in R] + \delta.$$

Intuitively, the DP property guarantees that in the worst case scenario when the adversary knows almost information of dataset D, except the sensitive info of one user X (e.g his/her races), the adversary can not infer his/her race with high probability by looking at the outcome of the mechanism $\mathcal{K}(.)$

ϵ, δ are privacy parameters, the smaller they are the more privacy (usually less utility) a model is.

Our Proposed Work

Key ideas:

- To learn a fair model, we propose to use Lagrangian Dual (LD) method to solve the constrained optimization in Equation 2.
- To protect users' confidential information during training, we extend the primal and dual steps in LD so that they are both differentially private.

The Lagrangian function from Equation 2 is:

$$\mathcal{L}_\lambda(\theta) = J(\mathcal{M}_\theta, D_P) + \lambda^\top |\mu(D_P) - \mu(D_G)|, \tag{3}$$

Traditional LD optimization consists of two iterative steps:

- Primal step, i.e optimize main parameter θ, i.e $\overset{*}{\theta}(\lambda) = \operatorname{argmin}_\theta \mathcal{L}_\lambda(\theta, \lambda)$
- Dual step, i.e optimize the multipliers λ, $\overset{*}{\lambda}(\theta) = \operatorname{argmax}_{\lambda \geq 0} \mathcal{L}_\lambda(\theta, \lambda)$

Note that, the LD method is not differentially private. Our proposed private fair Lagrangian Dual method (PF-LD) will extend a private version for primal and dual step.

- Private primal step, we limit individual contribution to the constraints by gradient clipping, and adding a Gaussian noise to the gradients.

$$\theta \leftarrow \theta - \alpha(\nabla_\theta [J(\mathcal{M}_\theta, B_P)] + \lambda^\top \left|\nabla_\theta \mu(B_P) - \bar{\nabla}_\theta^{C_p} \mu(B_G)\right| + \mathcal{N}(0, \sigma_p^2 \Delta_p^2 \boldsymbol{I})), \tag{4}$$

where B is a random mini-batch from D, and $\bar{\nabla}_\theta^{C_p}(x) = \nabla x / \max(1, \frac{\|\nabla x\|}{C_p})$ denotes the gradients of a given scalar loss x clipped in a C_p-ball, for $C_p > 0$.

- Private dual step, we follow similarly strategy to update privately multipliers λ

$$\lambda_{k+1} \leftarrow \lambda_k + s_k \left(|\mu(D_P) - \bar{\mu}^{C_d}(D_G)| + \mathcal{N}(0, \sigma_d^2 \Delta_d^2 \boldsymbol{I})\right) \tag{5}$$

Privacy Loss Computation

- The Gaussian noise levels σ_p (in primal step) and σ_d (in dual step) determines the privacy parameters ϵ, δ. Larger noise σ_d, σ_p guarantees more privacy but can hurt fairness goals.
- We employ moment accountant techniques to compute ϵ, δ based on σ_p, σ_d.

Results

- Dataset: Bank data, task is to detect client subscriptions, the sensitive info to protect is customers'age.
- Metrics: classification accuracy (higher is better) and fairness violation (lower is better).
- We compare ours (PF-LD) against the previous works, denoted by *CLF*, *A*, *Z*, and *M*

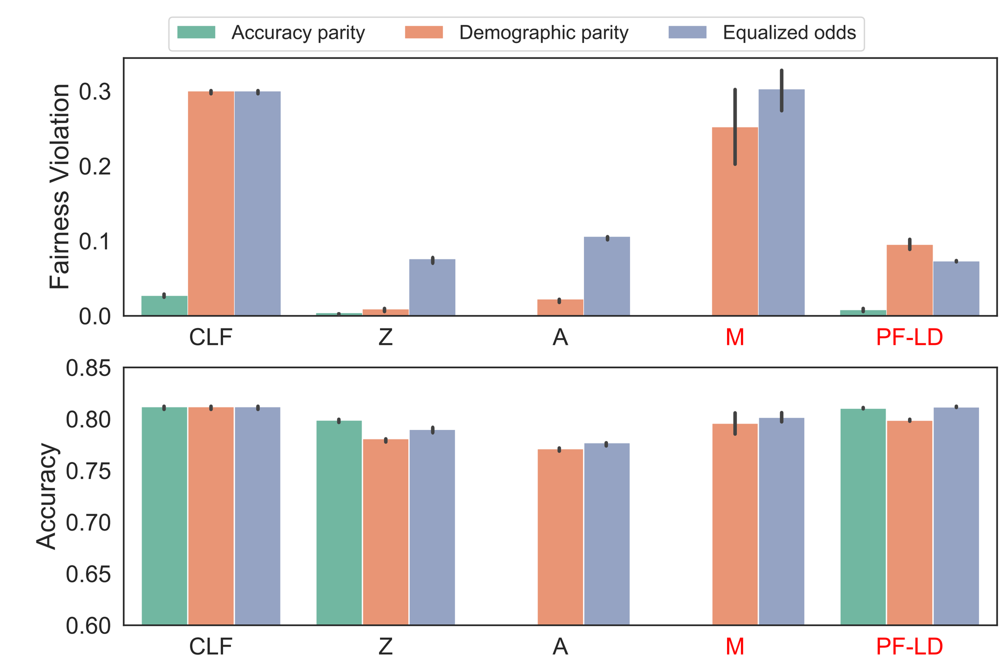

Fig. 1: Accuracy/Fairness comparison.

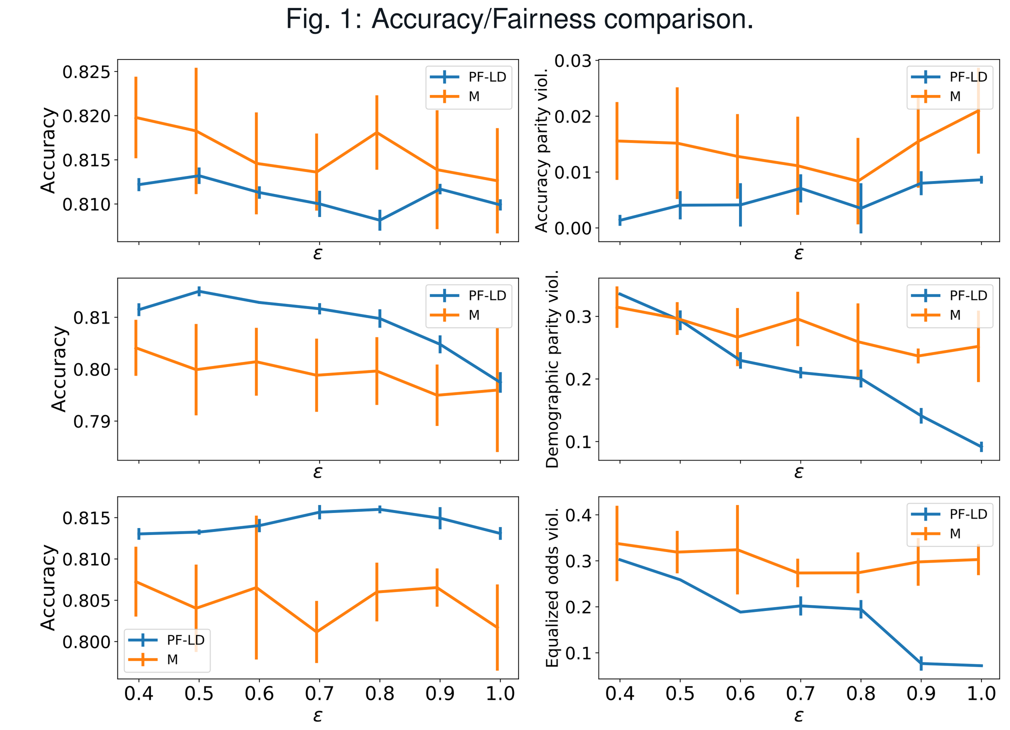

Fig. 2: PF-LD vs M under different privacy parameter ϵ

Conclusion

- We introduced PF-LD, a differentially private and fair Dual Lagrangian framework that can protect users privacy and ensure classifiers' fairness.
- PF-LD is better than the state-of-the art model e.g *M* in term of accuracy/fairness. Our performance is more robust (less variance) than the competitors.

References

[1] Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. "Differentially Private and Fair Deep Learning: A Lagrangian Dual Approach". In: (2020). arXiv: `2009.12562 [cs.LG]`.

[2] Zafar. "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment". In: *WWW*. 2017.