Appendix

Appendix A. Additional Discussion of Related Work

The study of differentially private fair learning algorithms was initiated by Jagielski et al. (2019). Jagielski et al. (2019) considered equalized odds and proposed two DP algorithms: 1) an ϵ-DP post-processing approach derived from Hardt et al. (2016a); and 2) an (ϵ, δ)-DP in-processing approach based on Agarwal et al. (2018). The major drawback of their post-processing approach is the unrealistic requirement that the algorithm have access to the sensitive attributes at test time, which Jagielski et al. (2019) admits "isn't feasible (or legal) in certain applications." Additionally, post-processing approaches are known to suffer from inferior fairness-accuracy tradeoffs compared with in-processing methods. While the in-processing method of Jagielski et al. (2019) does not require access to sensitive attributes at test time, it comes with a different set of disadvantages: 1) it is *limited to binary* classification; 2) its theoretical performance guarantees require the use of the *computationally inefficient* (i.e. exponential-time) exponential mechanism (McSherry and Talwar, 2007); 3) its theoretical performance guarantees require computations on the full training set and *do not permit mini-batch implementations*; 4) it requires the hypothesis class $\mathcal{H}$ to have finite VC dimension. In this work, we propose *the first algorithm that overcomes all of these pitfalls*: our algorithm is amenable to multi-way classification with multiple sensitive attributes, computationally efficient, and comes with convergence guarantees that hold even when mini-batches of $m < n$ samples are used in each iteration of training, and even when $\text{VC}(\mathcal{H}) = \infty$. Furthermore, our framework is flexible enough to accommodate many notions of group fairness besides equalized odds (e.g. demographic parity, accuracy parity).

Following Jagielski et al. (2019), several works have proposed other DP fair learning algorithms. *None of these works have managed to simultaneously address all the shortcomings* of the method of Jagielski et al. (2019). The work of Xu et al. (2019) proposed DP and fair binary logistic regression, but did not provide any theoretical convergence/performance guarantees. The work of Mozannar et al. (2020) combined aspects of both Hardt et al. (2016a) and Agarwal et al. (2018) in a two-step locally differentially private fairness algorithm. Their approach is *limited to binary classification*. Moreover, their algorithm requires $n/2$ samples in each iteration (of their in-processing step), making it *impractical for large-scale problems*. More recently, Tran et al. (2021b) devised another DP in-processing method based on lagrange duality, which covers non-binary classification problems. In a subsequent work, Tran et al. (2021a) studied the effect of DP on accuracy parity in ERM, and proposed using a regularizer to promote fairness. Finally, Tran et al. (2022) provided a semi-supervised fair "Private Aggregation of Teacher Ensembles" framework. A shortcoming of each of these three most recent works is their *lack of theoretical convergence or accuracy guarantees*. In another vein, some works have observed the disparate impact of privacy constraints on demographic subgroups (Bagdasaryan et al., 2019; Tran et al., 2021c).

Appendix B. Equalized Odds Version of ERMI

If equalized odds (Hardt et al., 2016b) is the desired fairness notion, then one should use the following variation of ERMI as a regularizer Lowy et al. (2022):

$$\begin{aligned}\widehat{D}_R(\widehat{Y};S|Y) &:= \mathbb{E}\left\{\frac{\hat{p}_{\widehat{Y},S|Y}(\widehat{Y},S|Y)}{\hat{p}_{\widehat{Y}|Y}(\widehat{Y}|Y)\hat{p}_{S|Y}(S|Y)}\right\}-1\\ &= \sum_{y=1}^{l}\sum_{j=1}^{l}\sum_{r=1}^{k}\frac{\hat{p}_{\widehat{Y},S|Y}(j,r|y)^2}{\hat{p}_{\widehat{Y}|Y}(j|y)\hat{p}_{S|Y}(r|y)}\hat{p}_Y(y)-1. \end{aligned} \tag{4}$$

Here $\hat{p}_{\widehat{Y},S|Y}$ denotes the empirical joint distribution of the predictions and sensitive attributes $(\widehat{Y},S)$ conditional on the true labels Y. In particular, if $D_R(\widehat{Y};S|Y)=0$, then $\widehat{Y}$ and S are conditionally independent given Y (i.e. equalized odds is satisfied).

Appendix C. Complete Version of Theorem 4

Let $\hat{\mathbf{y}}(x_i;\theta)\in\{0,1\}^l$ and $\mathbf{s}_i\in\{0,1\}^k$ be the one-hot encodings of $\hat{y}(x_i,\theta)$ and s_i, respectively: i.e., $\hat{\mathbf{y}}_j(x_i;\theta)=\mathbb{1}_{\{\hat{y}(x_i,\theta)=j\}}$ and $\mathbf{s}_{i,r}=\mathbb{1}_{\{s_i=r\}}$ for $j\in[l], r\in[k]$. Also, denote $\widehat{P}_s=\operatorname{diag}(\hat{p}_S(1),\dots,\hat{p}_S(k))$, where $\hat{p}_S(r):=\frac{1}{n}\sum_{i=1}^n\mathbb{1}_{\{s_i=r\}}\geqslant\rho>0$ is the empirical probability of attribute r ($r\in[k]$). Then we have the following re-formulation of (FERMI obj.) as a min-max problem:

Theorem 6 (Lowy et al. (2022)) (FERMI obj.) *is equivalent to*

$$\min_{\theta}\max_{W\in\mathbb{R}^{k\times l}}\left\{\widehat{F}(\theta,W):=\widehat{\mathcal{L}}(\theta)+\lambda\frac{1}{n}\sum_{i=1}^n\widehat{\psi}_i(\theta,W)\right\}, \tag{5}$$

where

$$\begin{aligned}\widehat{\psi}_i(\theta,W) :=& -\operatorname{Tr}(W\mathbb{E}[\hat{\mathbf{y}}(x_i,\theta)\hat{\mathbf{y}}(x_i,\theta)^T|x_i]W^T)\\ &+2\operatorname{Tr}(W\mathbb{E}[\hat{\mathbf{y}}(x_i;\theta)\mathbf{s}_i^T|x_i,\mathbf{s}_i]\widehat{P}_s^{-1/2})-1,\end{aligned}$$

$\mathbb{E}[\hat{\mathbf{y}}(x_i;\theta)\hat{\mathbf{y}}(x_i;\theta)^T|x_i]=\operatorname{diag}(\mathcal{F}_1(x_i,\theta),\dots,\mathcal{F}_l(x_i,\theta))$, *and* $\mathbb{E}[\hat{\mathbf{y}}(x_i;\theta)\mathbf{s}_i^T|x_i,\mathbf{s}_i]$ *is a* $k\times l$ *matrix with* $\mathbb{E}[\hat{\mathbf{y}}(x_i;\theta)\mathbf{s}_i^T|x_i,\mathbf{s}_i]_{r,j}=\mathbf{s}_{i,r}\mathcal{F}_j(x_i,\theta)$.

Strong concavity of $\widehat{\psi}_i$ is shown in Lowy et al. (2022).

Appendix D. DP-FERMI Algorithm: Privacy

We begin with a routine calculation of the derivatives of $\widehat{\psi}_i$, which follows by elementary matrix calculus:

Lemma 7 *Let* $\widehat{\psi}_i(\theta,W)=-\operatorname{Tr}(W\mathbb{E}[\hat{\mathbf{y}}(x_i,\theta)\hat{\mathbf{y}}(x_i,\theta)^T|x_i]W^T)+2\operatorname{Tr}(W\mathbb{E}[\hat{\mathbf{y}}(x_i;\theta)\mathbf{s}_i^T|x_i,\mathbf{s}_i]\widehat{P}_s^{-1/2})-1$, *where* $\mathbb{E}[\hat{\mathbf{y}}(x_i;\theta)\hat{\mathbf{y}}(x_i;\theta)^T|x_i]=\operatorname{diag}(\mathcal{F}_1(x_i,\theta),\dots,\mathcal{F}_l(x_i,\theta))$ *and* $\mathbb{E}[\hat{\mathbf{y}}(x_i;\theta)\mathbf{s}_i^T|x_i,\mathbf{s}_i]$ *is a* $k\times l$ *matrix with* $\mathbb{E}[\hat{\mathbf{y}}(x_i;\theta)\mathbf{s}_i^T|x_i,\mathbf{s}_i]_{r,j}=\mathbf{s}_{i,r}\mathcal{F}_j(x_i,\theta)$. *Then,*

$$\begin{aligned}\nabla_\theta\widehat{\psi}_i(\theta,W) =& -\nabla_\theta\operatorname{vec}(\mathbb{E}[\hat{\mathbf{y}}(x_i,\theta)\hat{\mathbf{y}}(x_i,\theta)^T|x_i])^T\operatorname{vec}(W^TW)\\ &+2\nabla_\theta\operatorname{vec}(\mathbb{E}[\mathbf{s}_i\hat{\mathbf{y}}(x_i,\theta)^T|x_i,s_i])\operatorname{vec}\left(W^T\left(\widehat{P}_S\right)^{-1/2}\right)\end{aligned}$$

and

$$\nabla_w \widehat{\psi}_i(\theta, W) = -2W\mathbb{E}[\widehat{\mathbf{y}}(x_i,\theta)\widehat{\mathbf{y}}(x_i,\theta)^T|x_i] + 2\widehat{P}_S^{-1/2}\mathbb{E}[\mathbf{s}_i\widehat{\mathbf{y}}(x_i,\theta)^T|x_i,s_i].$$

Using Lemma 7, we can prove that Algorithm 1 is DP:

Theorem 8 (Precise Statement of Privacy Claim in Theorem 5) *Let* $\epsilon \leqslant 2\ln(1/\delta)$, $\delta \in (0,1)$, *and* $T \geqslant \left(n\frac{\sqrt{\epsilon}}{2m}\right)^2$. *Assume* $\mathcal{F}(\cdot,x)$ *is* L_θ*-Lipschitz for all* x, *and* $|(W_t)_{r,j}| \leqslant D$ *for all* $t \in [T], r \in [k], j \in [l]$. *Then, for* $\sigma_w^2 \geqslant \frac{16T\ln(1/\delta)}{\epsilon^2 n^2\rho}$ *and* $\sigma_\theta^2 \geqslant \frac{16L_\theta^2 D^2\ln(1/\delta)T}{\epsilon^2 n^2\rho}$, *Algorithm 1 is* (ϵ,δ)*-DP with respect to the sensitive attributes for all data sets containing at least* ρ*-fraction of minority attributes. Further, if* $\sigma_w^2 \geqslant \frac{32T\ln(1/\delta)}{\epsilon^2 n^2}\left(\frac{1}{\rho}+D^2\right)$ *and* $\sigma_\theta^2 \geqslant \frac{64L_\theta^2 D^2\ln(1/\delta)T}{\epsilon^2 n^2\rho} + \frac{32D^4L_\theta^2 l^2 T\ln(1/\delta)}{\epsilon^2 n^2}$, *then Algorithm 1 is* (ϵ,δ)*-DP (with respect to all features) for all data sets containing at least* ρ*-fraction of minority attributes.*

Proof First consider the case in which only the sensitive attributes are private. By the moments accountant Theorem 1 in Abadi et al. (2016), it suffices to bound the sensitivity of the gradient updates by $\Delta_\theta^2 \leqslant \frac{8D^2L_\theta^2}{m^2\rho}$ and $\Delta_w^2 \leqslant \frac{8}{m^2\rho}$. Here

$$\Delta_\theta^2 = \sup_{Z\sim Z',\theta,W}\left\|\frac{1}{m}\sum_{i\in B_t}\left[\nabla_\theta\widehat{\psi}(\theta,W;z_i) - \nabla_\theta\widehat{\psi}(\theta,W;z_i')\right]\right\|^2$$

and $Z \sim Z'$ means that Z and Z' are two data sets (both with ρ-fraction of minority attributes) that differ in exactly one person's sensitive attributes: i.e. $s_i \neq s_i'$ for some unique $i \in [n]$, but $z_j = z_j'$ for all $j \neq i$ and $(x_i,y_i) = (x_i',y_i')$. Likewise,

$$\Delta_w^2 = \sup_{Z\sim Z',\theta,W}\left\|\frac{1}{m}\sum_{i\in B_t}\left[\nabla_w\widehat{\psi}(\theta,W;z_i) - \nabla_w\widehat{\psi}(\theta,W;z_i')\right]\right\|^2.$$

Now, by Lemma 7,

$$\begin{aligned}\nabla_\theta\widehat{\psi}_i(\theta,W) = &-\nabla_\theta \operatorname{vec}(\mathbb{E}[\widehat{\mathbf{y}}(x_i,\theta)\widehat{\mathbf{y}}(x_i,\theta)^T|x_i])^T\operatorname{vec}(W^TW)\\ &+ 2\nabla_\theta\operatorname{vec}(\mathbb{E}[\mathbf{s}_i\widehat{\mathbf{y}}(x_i,\theta)^T|x_i,s_i])\operatorname{vec}\left(W^T\left(\widehat{P}_S\right)^{-1/2}\right),\end{aligned}$$

and notice that only the second term depends on S. Therefore, we can bound the ℓ_2-sensitivity of the θ-gradient updates by:

$$\begin{aligned}
\Delta_\theta^2 = \sup_{Z\sim Z',W,\theta} \Bigg\| & \frac{1}{m}\sum_{i=1}^{m} 2\nabla_\theta \operatorname{vec}(\mathbb{E}[\mathbf{s}_i\widehat{\mathbf{y}}(x_i,\theta)^T|x_i,s_i])\operatorname{vec}\left(W^T\left(\widehat{P}_S\right)^{-1/2}\right) \\
& - 2\nabla_\theta \operatorname{vec}(\mathbb{E}[\mathbf{s}_i'\widehat{\mathbf{y}}(x_i,\theta)^T|x_i,s_i'])\operatorname{vec}\left(W^T\left(\widehat{P}_{S'}\right)^{-1/2}\right)\Bigg\|^2 \\
\leqslant & \frac{4}{m^2}\sup_{x,\mathbf{s}_i,\mathbf{s}_i',W,\theta}\left[\sum_{r=1}^{k}\sum_{j=1}^{l}\|\nabla_\theta\mathcal{F}_j(\theta,x)\|^2 W_{r,j}^2\left(\frac{s_{i,r}}{\sqrt{\widehat{P}_S(r)}}-\frac{s_{i,r}'}{\sqrt{\widehat{P}_{S'}(r)}}\right)^2\right] \\
\leqslant & \frac{8}{\rho m^2}\sup_{x,W,\theta}\left(\sum_{j=1}^{l}\|\nabla_\theta\mathcal{F}_j(\theta,x)\|^2 W_{r,j}^2\right) \\
\leqslant & \frac{8D^2L_\theta^2}{\rho m^2},
\end{aligned}$$

using Lipschitz continuity of $\mathcal{F}(\cdot,x)$, the assumption that $\mathcal{W}$ has diameter bounded by D, the assumption that the data sets have at least ρ-fraction of sensitive attribute r for all $r\in[k]$. Similarly, for the W-gradients, we have

$$\nabla_w\widehat{\psi}_i(\theta,W) = -2W\mathbb{E}[\widehat{\mathbf{y}}(x_i,\theta)\widehat{\mathbf{y}}(x_i,\theta)^T|x_i] + 2\widehat{P}_S^{-1/2}\mathbb{E}[\mathbf{s}_i\widehat{\mathbf{y}}(x_i,\theta)^T|x_i,s_i]$$

by Lemma 7. Hence

$$\begin{aligned}
\Delta_W^2 = \sup_{\theta,W,\mathbf{s}_i,\mathbf{s}_i'}\frac{4}{m^2}\Bigg\| & -W\operatorname{diag}(\mathcal{F}_1(\theta,x_i),\ldots,\mathcal{F}_l(\theta,x_i)) + \widehat{P}_S^{-1/2}\mathbb{E}[\mathbf{s}_i\widehat{\mathbf{y}}_i(x_i;\theta_t)^T|x_i,s_i] \\
& + W\operatorname{diag}(\mathcal{F}_1(\theta,x_i),\ldots,\mathcal{F}_l(\theta,x_i)) - \widehat{P}_{S'}^{-1/2}\mathbb{E}[\mathbf{s'}_i\widehat{\mathbf{y}}_i(x_i;\theta_t)^T|x_i,s_i']\Bigg\|^2 \\
\leqslant & \frac{4}{m^2}\sup_{\theta,W,\mathbf{s}_i,\mathbf{s}_i'}\sum_{j=1}^{l}\mathcal{F}_j(\theta,x_i)^2\sum_{r=1}^{k}\left(\frac{s_{i,r}}{\sqrt{\widehat{P}_S(r)}}-\frac{s_{i,r}'}{\sqrt{\widehat{P}_{S'}(r)}}\right)^2 \\
\leqslant & \frac{8}{m^2\rho},
\end{aligned}$$

since $\sum_{j=1}^{l}\mathcal{F}_j(\theta,x_i)^2 \leqslant \sum_{j=1}^{l}\mathcal{F}_j(\theta,x_i) = 1$. This establishes the desired privacy guarantee with respect to sensitive attributes for Algorithm 1.

Now consider the case in which all features are private. We aim to bound the sensitivities of the gradient updates to changes in a single sample $z_i = (s_i,x_i,y_i)$. Denote these new sensitivities by

$$\tilde{\Delta}_\theta = \sup_{Z\sim Z',\theta,W}\left\|\frac{1}{m}\sum_{i\in B_t}\left[\nabla_\theta\widehat{\psi}(\theta,W;z_i)-\nabla_\theta\widehat{\psi}(\theta,W;z_i')\right]\right\|,$$

where we now write $Z \sim Z'$ to mean that Z and Z' are two data sets (both with ρ-fraction of minority attributes) that differ in exactly one person's (sensitive and non-sensitive) data: i.e. $z_i \neq z_i'$ for some unique $i \in [n]$. Likewise,

$$\tilde{\Delta}_W = \sup_{Z \sim Z', \theta, W} \left\| \frac{1}{m} \sum_{i \in B_t} \left[\nabla_w \widehat{\psi}(\theta, W; z_i) - \nabla_w \widehat{\psi}(\theta, W; z_i') \right] \right\|.$$

Then

$$\begin{aligned}
\tilde{\Delta}_\theta = \frac{1}{m} \sup_{z_i, z_i', \theta, W, S \sim S'} \Big\| & - \nabla_\theta \operatorname{vec}(\mathbb{E}[\widehat{\mathbf{y}}(x_i, \theta) \widehat{\mathbf{y}}(x_i, \theta)^T | x_i])^T \operatorname{vec}(W^T W) + 2 \nabla_\theta \operatorname{vec}(\mathbb{E}[\mathbf{s}_i \widehat{\mathbf{y}}(x_i, \theta)^T | x_i, s_i]) \\
& \operatorname{vec}\left(W^T \left(\widehat{P}_S \right)^{-1/2} \right) + \nabla_\theta \operatorname{vec}(\mathbb{E}[\widehat{\mathbf{y}}(x_i', \theta) \widehat{\mathbf{y}}(x_i', \theta)^T | x_i'])^T \operatorname{vec}(W^T W) \\
& - 2 \nabla_\theta \operatorname{vec}(\mathbb{E}[\mathbf{s}_i' \widehat{\mathbf{y}}(x_i', \theta)^T | x_i', s_i']) \operatorname{vec}\left(W^T \left(\widehat{P}_{S'} \right)^{-1/2} \right) \Big\| \\
& \leqslant \frac{2 L_\theta l D}{m} + \Delta_\theta.
\end{aligned}$$

Thus, $\tilde{\Delta}_\theta^2 \leqslant \frac{4 L_\theta^2 l^2 D^2}{m^2} + 2\Delta_\theta^2$. Therefore, by the moments accountant, the collection of all θ_t updates in Algorithm 1 is (ϵ, δ)-DP if $\sigma_\theta^2 \geqslant \frac{32 D^2 L_\theta^2 T \ln(1/\delta)}{\rho \epsilon^2 n^2} + \frac{8 D^2 L_\theta^2 l^2 T \ln(1/\delta)}{\epsilon^2 n^2} = \frac{8 L_\theta^2 D^2 T \ln(1/\delta)}{\epsilon^2 n^2} \left(\frac{4}{\rho} + l^2 \right)$.

Next, we bound the sensitivity $\tilde{\Delta}_W$ of the W-gradient updates. We have

$$\begin{aligned}
\tilde{\Delta}_W^2 = \sup_{\theta, W, z_i, z_i'} \frac{4}{m^2} \Big\| & - W \operatorname{diag}(\mathcal{F}_1(\theta, x_i), \ldots, \mathcal{F}_l(\theta, x_i)) + \widehat{P}_S^{-1/2} \mathbb{E}[\mathbf{s}_i \widehat{\mathbf{y}}_i(x_i; \theta_t)^T | x_i, s_i] \\
& + W \operatorname{diag}(\mathcal{F}_1(\theta, x_i'), \ldots, \mathcal{F}_l(\theta, x_i')) - \widehat{P}_{S'}^{-1/2} \mathbb{E}[\mathbf{s'}_i \widehat{\mathbf{y}}_i^T(x_i'; \theta_t) | x_i', s_i'] \Big\|^2 \\
\leqslant 2\Delta_W^2 + \frac{8}{m^2} \sup_{\theta, W, x_i, x_i'} & \left\| W \operatorname{diag}(\mathcal{F}_1(\theta, x_i) - \mathcal{F}_1(\theta, x_i'), \ldots, \mathcal{F}_l(\theta, x_i) - \mathcal{F}_l(\theta, x_i')) \right\|^2 \\
\leqslant 2\Delta_W^2 + \frac{16 D^2}{m^2} & \sup_{\theta, x_i} \sum_{j=1}^{l} \mathcal{F}_j(\theta, x_i)^2 \\
\leqslant 2\Delta_W^2 + \frac{16 D^2}{m^2} & .
\end{aligned}$$

Therefore, by the moments accountant, the collection of all W_t updates in Algorithm 1 is (ϵ, δ)-DP if $\sigma_w^2 \geqslant \frac{32 T \ln(1/\delta)}{\epsilon^2 n^2} \left(\frac{1}{\rho} + D^2 \right)$. This completes the proof. ■

Appendix E. DP-FERMI Algorithm: Utility

To prove the convergence guarantee in Theorem 5, we will first derive a more general result. Namely, in Appendix E.1, we will provide a precise upper bound on the stationarity gap of noisy DP stochastic gradient descent ascent (DP-SGDA).

E.1. Noisy DP-SGDA for Nonconvex-Strongly Concave Min-Max Problems

Consider a generic (smooth) nonconvex-strongly concave min-max ERM problem:

$$\min_{\theta\in\mathbb{R}^{d_\theta}} \max_{w\in\mathcal{W}} \left\{ F(\theta, w) := \frac{1}{n}\sum_{i=1}^{n} f(\theta, w; z_i) \right\}, \tag{6}$$

where $f(\theta, \cdot; z)$ is μ-strongly concave[3] for all θ, z but $f(\cdot, w; z)$ is potentially non-convex. We propose Noisy DP-SGDA[4] (Algorithm 2) for privately solving (6), which is a noisy DP

Algorithm 2 Noisy Differentially Private Stochastic Gradient Descent-Ascent (DP-SGDA)

1: **Input**: data Z, $\theta_0 \in \mathbb{R}^{d_\theta}$, $w_0 \in \mathcal{W}$, step-sizes (η_θ, η_w), privacy noise parameters σ_θ, σ_w, batch size m, iteration number $T \geqslant 1$.
2: **for** $t = 0, 1, \ldots, T-1$ **do**
3: Draw a batch of data points $\{z_i\}_{i=1}^m$ uniformly at random from Z.
4: Update $\theta_{t+1} \leftarrow \theta_t - \eta_\theta \left(\frac{1}{m}\sum_{i=1}^m \nabla_\theta f(\theta_t, w_t; z_i) + u_t\right)$, where $u_t \sim \mathcal{N}(0, \sigma_\theta^2 \mathbf{I}_{d_\theta})$ and $w_{t+1} \leftarrow \Pi_{\mathcal{W}}\left[w_t + \eta_w \left(\frac{1}{m}\sum_{i=1}^m \nabla_w f(\theta_t, w_t; z_i) + v_t\right)\right]$, where $v_t \sim \mathcal{N}(0, \sigma_w^2 \mathbf{I}_{d_w})$.
5: **end for**
6: Draw $\hat{\theta}_T$ uniformly at random from $\{\theta_t\}_{t=1}^T$.
7: **Return:** $\hat{\theta}_T$

variation of two-timescale SGDA (Lin et al., 2020). Now, we provide *the first theoretical convergence guarantee for DP non-convex min-max optimization*:

Theorem 9 (Privacy and Utility of Algorithm 2, Informal Version) *Let $\epsilon \leqslant 2\ln(1/\delta)$, $\delta \in (0,1)$. Assume: $f(\cdot, w; z)$ is L_θ-Lipschitz[5] and $f(\theta, \cdot; z)$ is L_w-Lipschitz for all θ, w, z; and $\mathcal{W} \subset \mathbb{R}^{d_w}$ is a convex, compact set. Denote $\Phi(\theta) = \max_{w\in\mathcal{W}} F(\theta, w)$. Choose $\sigma_w^2 = \frac{8TL_w^2\ln(1/\delta)}{\epsilon^2 n^2}$, $\sigma_\theta^2 = \frac{8TL_\theta^2\ln(1/\delta)}{\epsilon^2 n^2}$, and $T \geqslant \left(n\frac{\sqrt{\epsilon}}{2m}\right)^2$. Then, Algorithm 2 is (ϵ, δ)-DP. Further, if $f(\cdot, \cdot; z)$ has Lipschitz gradients and $f(\theta, \cdot; z)$ is strongly concave, then $\exists\ T, \eta_\theta, \eta_w$ such that*

$$\mathbb{E}\|\nabla\Phi(\hat{\theta}_T)\|^2 = \mathcal{O}\left(\frac{\sqrt{d\ln(1/\delta)}}{\epsilon n}\right),$$

where $d = \max(d_\theta, d_w)$. (The expectation is solely over the algorithm.)

In our DP fair learning application, $f(\theta, W; z_i) = \ell(\theta, x_i, y_i) + \lambda\widehat{\psi}_i(\theta, W)$ and the strong concavity assumption on f in Theorem 9 is automatically satisfied, by Lowy et al. (2022). The Lipschitz and smoothness assumptions on f are standard in optimization literature and are satisfied for loss functions that are typically used in pracdtice. In our application to DP-FERMI, these assumptions hold as long as the loss function ℓ and $\mathcal{F}$ are Lipschitz continuous with Lipschitz gradients. Our next goal is to prove (the precise, scale-invariant version of) Theorem 9. To that end, we require the following notation.

Notation and Assumptions: Let $f : \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_w} \times \mathcal{Z} \to \mathbb{R}$, and $F(\theta, w) = \frac{1}{n}\sum_{i=1}^n f(\theta, w; z_i)$ for fixed training data $Z = (z_1, \cdots, z_n) \in \mathcal{Z}^n$. Let $\mathcal{W} \subset \mathbb{R}^{d_w}$ be a convex, compact set.

3. We say a differentiable function g is μ-strongly concave if $g(\alpha) + \langle\nabla g(\alpha), \alpha' - \alpha\rangle - \frac{\mu}{2}\|\alpha - \alpha'\|^2 \geqslant g(\alpha')$ for all α, α'.
4. DP-SGDA was also used in Yang et al. (2022) for *convex* and *PL* min-max problems.
5. We say function g is L-Lipschitz if $\|g(\alpha) - g(\alpha')\| \leqslant L\|\alpha - \alpha'\|$ for all α, α'.

For any $\theta \in \mathbb{R}^{d_\theta}$, denote $w^*(\theta) \in \operatorname{argmax}_{w\in\mathcal{W}} F(\theta, w)$ and $\widehat{\Phi}(\theta) = \max_{w\in\mathcal{W}} F(\theta, w)$. Let $\Delta_\Phi = \widehat{\Phi}(\theta_0) - \inf_\theta \widehat{\Phi}_Z(\theta)$. Recall that a function h is β-smooth if its derivative ∇h is β-Lipschitz. We write $a \lesssim b$ if there is an absolute constant $C > 0$ such that $a \leqslant Cb$.

Assumption 1

1. *$f(\cdot, w; z)$ is L_θ-Lipschitz and β_θ-smooth for all $w \in \mathcal{W}, z \in \mathcal{Z}$.*
2. *$f(\theta, \cdot; z)$ is L_w-Lipschitz, β_w-smooth, and μ-strongly concave on $\mathcal{W}$ for all $\theta \in \mathbb{R}^{d_\theta}$, $z \in \mathcal{Z}$.*
3. *$\|\nabla_w f(\theta, w; z) - \nabla_w f(\theta', w; z)\| \leqslant \beta_{\theta w}\|\theta - \theta'\|$ and $\|\nabla_\theta f(\theta, w; z) - \nabla_\theta f(\theta, w'; z)\| \leqslant \beta_{\theta w}\|w - w'\|$ for all $\theta, \theta', w, w', z$.*
4. *$\mathcal{W}$ has ℓ_2 diameter bounded by $D \geqslant 0$.*
5. *$\nabla_w F(\theta, w^*(\theta)) = 0$ for all θ, where $w^*(\theta)$ denotes the unconstrained global minimizer of $F(\theta, \cdot)$.*

The first four assumptions are standard in (DP and min-max) optimization. The fifth assumption means that $\mathcal{W}$ contains the *unconstrained* global minimizer $w^*(\theta)$ of $F(\theta, \cdot)$ for all θ. Hence (6) is equivalent to

$$\min_{\theta\in\mathbb{R}^{d_\theta}} \max_{w\in\mathbb{R}^{d_w}} F(\theta, w).$$

This assumption is not actually necessary for our convergence result to hold, but we will need it when we *apply* our results to the DP fairness problem. Moreover, it simplifies the proof of our convergence result. We refer to problems of the form (6) that satisfy Assumption 1 as "(smooth) nonconvex-strongly concave min-max." We denote $\kappa_w := \frac{\beta_w}{\mu}$ and $\kappa_{\theta w} := \frac{\beta_{\theta w}}{\mu}$.

We can now provide the complete, precise version of Theorem 9:

Theorem 10 (Privacy and Utility of Algorithm 2, Formal Version) *Let $\epsilon \leqslant 2\ln(1/\delta)$, $\delta \in (0, 1)$. Grant Assumption 1. Choose $\sigma_w^2 = \frac{8TL_w^2\ln(1/\delta)}{\epsilon^2 n^2}$, $\sigma_\theta^2 = \frac{8TL_\theta^2\ln(1/\delta)}{\epsilon^2 n^2}$, and $T \geqslant \left(n\frac{\sqrt{\epsilon}}{2m}\right)^2$. Then Algorithm 2 is (ϵ, δ)-DP. Further, if we choose $\eta_\theta = \frac{1}{16\kappa_w(\beta_\theta + \beta_{\theta w}\kappa_{\theta w})}$, $\eta_w = \frac{1}{\beta_w}$, and $T \approx \sqrt{\kappa_w[\Delta_\Phi(\beta_\theta + \beta_{\theta w}\kappa_{\theta w}) + \beta_{\theta w}^2 D^2]}\epsilon n \min\left(\frac{1}{L_\theta\sqrt{d_\theta}}, \frac{\beta_w}{\beta_{\theta w}L_w\sqrt{\kappa_w d_w}}\right)$, then*

$$\mathbb{E}\|\nabla\Phi(\hat{\theta}_T)\|^2 \lesssim \sqrt{\Delta_\Phi\left(\beta_\theta + \beta_{\theta w}\kappa_{\theta w})\kappa_w + \kappa_w\beta_{\theta w}^2 D^2\right)}\left[\frac{L_\theta\sqrt{d_\theta\ln(1/\delta)}}{\epsilon n} + \left(\frac{\beta_{\theta w}\sqrt{\kappa_w}}{\beta_w}\right)\frac{L_w\sqrt{d_w\ln(1/\delta)}}{\epsilon n}\right] + \frac{\mathbb{1}_{\{m<n\}}}{m}\left(L_\theta^2 + \frac{\kappa_w\beta_{\theta w}^2 L_w^2}{\beta_w^2}\right).$$

In particular, if $m \geqslant \min\left(\frac{\epsilon n L_\theta}{\sqrt{d_\theta\kappa_w[\Delta_\Phi(\beta_\theta+\beta_{\theta w}\kappa_{\theta w})+\beta_{\theta w}^2 D^2]}}, \frac{\epsilon n L_w\sqrt{\kappa_w}}{\beta_{\theta w}\beta_w\sqrt{d_w\kappa_w[\Delta_\Phi(\beta_\theta+\beta_{\theta w}\kappa_{\theta w})+\beta_{\theta w}^2 D^2]}}\right)$, then

$$\mathbb{E}\|\nabla\Phi(\hat{\theta}_T)\|^2 \lesssim \sqrt{\kappa_w[\Delta_\Phi(\beta_\theta + \beta_{\theta w}\kappa_{\theta w}) + \beta_{\theta w}^2 D^2]}\left(\frac{\sqrt{\ln(1/\delta)}}{\epsilon n}\right)\left(L_\theta\sqrt{d_\theta} + \left(\frac{\beta_{\theta w}\sqrt{\kappa_w}}{\beta_w}\right)L_w\sqrt{d_w}\right).$$

The proof of Theorem 10 will require several technical lemmas. These technical lemmas, in turn, require some preliminary lemmas, which we present below.

We begin with a refinement of Lemma 4.3 from Lin et al. (2020):

Lemma 11 *Grant Assumption 1. Then Φ is $2(\beta_\theta + \beta_{\theta w}\kappa_{\theta w})$-smooth with $\nabla\Phi(\theta) = \nabla_\theta F(\theta, w^*(\theta))$, and $w^*(\cdot)$ is κ_w-Lipschitz.*

Proof The proof follows almost exactly as in the proof of Lemma 4.3 of Lin et al. (2020), using Danskin's theorem, but we carefully track the different smoothness parameters with respect to w and θ (and their units) to obtain the more precise result. ■

Lemma 12 (Lei et al. (2017)) *Let $\{a_l\}_{l\in[n]}$ be an arbitrary collection of vectors such that $\sum_{l=1}^n a_l = 0$. Further, let $\mathcal{S}$ be a uniformly random subset of $[n]$ of size m. Then,*

$$\mathbb{E}\left\|\frac{1}{m}\sum_{l\in\mathcal{S}} a_l\right\|^2 = \frac{n-m}{(n-1)m}\frac{1}{n}\sum_{l=1}^n \|a_l\|^2 \leqslant \frac{\mathbb{1}_{\{m<n\}}}{m\ n}\sum_{l=1}^n \|a_l\|^2.$$

Lemma 13 (Co-coercivity of the gradient) *For any β-smooth and convex function g, we have*

$$\|\nabla g(a) - \nabla g(b)\|^2 \leqslant 2\beta(g(a) - g(b) - \langle g(b), a - b\rangle),$$

for all $a, b \in$ domain(g).

Having recalled the necessary preliminaries, we now provide the novel technical ingredients that we'll need for the proof of Theorem 10. The next lemma quantifies the progress made in minimizing Φ from a single step of noisy stochastic gradient descent in θ (i.e. line 4 of Algorithm 2):

Lemma 14 *For all $t \in [T]$, the iterates of Algorithm 2 satisfy*

$$\begin{aligned}\mathbb{E}\Phi(\theta_t) \leqslant \Phi(\theta_{t-1}) &- \left(\frac{\eta_\theta}{2} - 2(\beta_\theta + \beta_{\theta w}\kappa_{\theta w})\eta_\theta^2\right)\mathbb{E}\|\nabla\Phi(\theta_{t-1})\|^2 \\ &+ \left(\frac{\eta_\theta}{2} + 2\eta_\theta^2(\beta_\theta + \beta_{\theta w}\kappa_{\theta w})\mathbb{E}\|\nabla\Phi(\theta_{t-1}) - \nabla_\theta F(\theta_{t-1}, w_{t-1})\|^2\right) \\ &+ (\beta_\theta\beta_{\theta w}\kappa_{\theta w})\eta_\theta^2\left(d_\theta\sigma_\theta^2 + \frac{4L_\theta^2}{m}\mathbb{1}_{\{m<n\}}\right),\end{aligned}$$

conditional on θ_{t-1}, w_{t-1}.

Proof Let us denote $\widetilde{g} := \frac{1}{m}\sum_{i=1}^m \nabla_\theta f(\theta_{t-1}, w_{t-1}; z_i) + u_{t-1} := g + u_{t-1}$, so $\theta_t = \theta_{t-1} - \eta_\theta\widetilde{g}$. First condition on the randomness due to sampling and Gaussian noise addition. By smoothness of Φ (see Theorem 11), we have

$$\begin{aligned}\Phi(\theta_t) &\leqslant \Phi(\theta_{t-1}) - \eta_\theta\langle\widetilde{g}, \nabla\Phi(\theta_{t-1})\rangle + (\beta_\theta + \beta_{\theta w}\kappa_{\theta w})\eta_\theta^2\|\widetilde{g}\|^2 \\ &= \Phi(\theta_{t-1}) - \eta_\theta\|\nabla\Phi(\theta_{t-1})\|^2 - \eta_\theta\langle\widetilde{g} - \nabla\Phi(\theta_{t-1}), \nabla\Phi(\theta_{t-1})\rangle + (\beta_\theta + \beta_{\theta w}\kappa_{\theta w})\eta_\theta^2\|\widetilde{g}\|^2.\end{aligned}$$

Taking expectation (conditional on θ_{t-1}, w_{t-1}),

$$\begin{aligned}
\mathbb{E}[\Phi(\theta_t)] &\leqslant \Phi(\theta_{t-1}) - \eta_\theta \|\nabla\Phi(\theta_{t-1})\|^2 - \eta_\theta \langle \nabla_\theta F(\theta_{t-1}, w_{t-1}) - \nabla\Phi(\theta_{t-1}), \nabla\Phi(\theta_{t-1})\rangle \\
&\quad + (\beta_\theta + \beta_{\theta w}\kappa_{\theta w})\eta_\theta^2 \left[d_\theta \sigma_\theta^2 + \mathbb{E}\|g - \nabla_\theta F(\theta_{t-1}, w_{t-1})\|^2 + \|\nabla_\theta F(\theta_{t-1}, w_{t-1})\|^2\right] \\
&\leqslant \Phi(\theta_{t-1}) - \eta_\theta \|\nabla\Phi(\theta_{t-1})\|^2 - \eta_\theta \langle \nabla_\theta F(\theta_{t-1}, w_{t-1}) - \nabla\Phi(\theta_{t-1}), \nabla\Phi(\theta_{t-1})\rangle \\
&\quad + (\beta_\theta + \beta_{\theta w}\kappa_{\theta w})\eta_\theta^2 \left[d_\theta \sigma_\theta^2 + \frac{4L_\theta^2}{m}\mathbb{1}_{\{m<n\}} + \|\nabla_\theta F(\theta_{t-1}, w_{t-1})\|^2\right] \\
&\leqslant \Phi(\theta_{t-1}) - \eta_\theta \|\nabla\Phi(\theta_{t-1})\|^2 - \eta_\theta \langle \nabla_\theta F(\theta_{t-1}, w_{t-1}) - \nabla\Phi(\theta_{t-1}), \nabla\Phi(\theta_{t-1})\rangle \\
&\quad + (\beta_\theta + \beta_{\theta w}\kappa_{\theta w})\eta_\theta^2 \left[d_\theta \sigma_\theta^2 + \frac{4L_\theta^2}{m}\mathbb{1}_{\{m<n\}} + 2\|\nabla_\theta F(\theta_{t-1}, w_{t-1}) - \nabla\Phi(\theta_{t-1})\|^2 + 2\|\nabla\Phi(\theta_{t-1})\|^2\right] \\
&\leqslant \Phi(\theta_{t-1}) - \eta_\theta \|\nabla\Phi(\theta_{t-1})\|^2 + \frac{\eta_\theta}{2}\left[\|\nabla\Phi(\theta_{t-1}) - \nabla_\theta F(\theta_{t-1}, w_{t-1})\|^2 + \|\nabla\Phi(\theta_{t-1})\|^2\right] \\
&\quad + (\beta_\theta + \beta_{\theta w}\kappa_{\theta w})\eta_\theta^2 \left[d_\theta \sigma_\theta^2 + \frac{4L_\theta^2}{m}\mathbb{1}_{\{m<n\}} + 2\|\nabla_\theta F(\theta_{t-1}, w_{t-1}) - \nabla\Phi(\theta_{t-1})\|^2 + 2\|\nabla\Phi(\theta_{t-1})\|^2\right] \\
&\leqslant \Phi(\theta_{t-1}) - \left(\frac{\eta_\theta}{2} - 2(\beta_\theta + \beta_{\theta w}\kappa_{\theta w})\eta_\theta^2\right)\|\nabla\Phi(\theta_{t-1})\|^2 \\
&\quad + \left(\frac{\eta_\theta}{2} + 2(\beta_\theta + \beta_{\theta w}\kappa_{\theta w})\eta_\theta^2\right)\|\nabla\Phi(\theta_{t-1}) - \nabla_\theta F(\theta_{t-1}, w_{t-1})\|^2 \\
&\quad + (\beta_\theta + \beta_{\theta w}\kappa_{\theta w})\eta_\theta^2 \left(d_\theta \sigma_\theta^2 + \frac{4L_\theta^2}{m}\mathbb{1}_{\{m<n\}}\right).
\end{aligned}$$

In the first inequality, we used the fact that the Gaussian noise has mean zero and is independent of $(\theta_{t-1}, w_{t-1}, Z)$, plus the fact that $\mathbb{E}g = \nabla_\theta F(\theta_{t-1}, w_{t-1})$. In the second inequality, we used Theorem 12 and Lipschitz continuity of f. In the third and fourth inequalities, we used Young's inequality and Cauchy-Schwartz. ■

For the particular η_θ prescribed in Theorem 10, we obtain:

Lemma 15 *Grant Assumption 1. If $\eta_\theta = \frac{1}{16\kappa_w(\beta_\theta + \beta_{\theta w}\kappa_{\theta w})}$, then the iterates of Algorithm 2 satisfy ($\forall t \geqslant 0$)*

$$\mathbb{E}\Phi(\theta_{t+1}) \leqslant \mathbb{E}\left[\Phi(\theta_t) - \frac{3}{8}\eta_\theta\|\Phi(\theta_t)\|^2 + \frac{5}{8}\eta_\theta\left(\beta_{\theta w}^2\|w^*(\theta_t) - w_t\|^2 + d_\theta\sigma_\theta^2 + \frac{4L_\theta^2}{m}\mathbb{1}_{\{m<n\}}\right)\right].$$

Proof By Theorem 14, we have

$$\begin{aligned}
\mathbb{E}\Phi(\theta_{t+1}) &\leqslant \mathbb{E}\Phi(\theta_t) - \left(\frac{\eta_\theta}{2} - 2(\beta_\theta + \beta_{\theta w}\kappa_{\theta w})\eta_\theta^2\right)\mathbb{E}\|\nabla\Phi(\theta_t)\|^2 \\
&\quad + \left(\frac{\eta_\theta}{2} + 2\eta_\theta^2(\beta_\theta + \beta_{\theta w}\kappa_{\theta w})\mathbb{E}\|\nabla\Phi(\theta_t) - \nabla_\theta F(\theta_t, w_t)\|^2\right) + (\beta_\theta\beta_{\theta w}\kappa_{\theta w})\eta_\theta^2\left(d_\theta\sigma_\theta^2 + \frac{4L_\theta^2}{m}\mathbb{1}_{\{m<n\}}\right) \\
&\leqslant \mathbb{E}\Phi(\theta_t) - \frac{3}{8}\eta_\theta\mathbb{E}\|\nabla\Phi(\theta_t)\|^2 + \frac{5}{8}\eta_\theta\left[\mathbb{E}\|\nabla\Phi(\theta_t) - \nabla_\theta F(\theta_t, w_t)\|^2 + d_\theta\sigma_\theta^2 + \frac{4L_\theta^2}{m}\mathbb{1}_{\{m<n\}}\right] \\
&\leqslant \mathbb{E}\Phi(\theta_t) - \frac{3}{8}\eta_\theta\mathbb{E}\|\nabla\Phi(\theta_t)\|^2 + \frac{5}{8}\eta_\theta\left[\beta_{\theta w}^2\mathbb{E}\|w^*(\theta_t) - w_t\|^2 + d_\theta\sigma_\theta^2 + \frac{4L_\theta^2}{m}\mathbb{1}_{\{m<n\}}\right].
\end{aligned}$$

In the second inequality, we simply used the definition of η_θ. In the third inequality, we used the fact that $\nabla\Phi(\theta_t) = \nabla_\theta F(\theta_t, w^*(\theta_t))$ (see Theorem 11) together with Assumption 1 (part 3). ■

Next, we describe the progress made in the w_t updates:

Lemma 16 *Grant Assumption 1. If $\eta_w = \frac{1}{\beta_w}$, then the iterates of Algorithm 2 satisfy ($\forall t \geqslant 0$)*

$$\mathbb{E}\|w^*(\theta_{t+1}) - w_{t+1}\|^2 \leqslant \left(1 - \frac{1}{2\kappa_w} + 4\kappa_w\kappa_{\theta w}^2\eta_\theta^2\beta_{\theta w}^2\right)\mathbb{E}\|w^*(\theta_t) - w_t\|^2 + \frac{2}{\beta_w^2}\left(\frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}} + d_w\sigma_w^2\right) + 4\kappa_w\kappa_{\theta w}^2\eta_\theta^2\left(\mathbb{E}\|\nabla\Phi(\theta_t)\|^2 + d_\theta\sigma_\theta^2\right).$$

Proof Fix any t and denote $\delta_t := \mathbb{E}\|w^*(\theta_t) - w_t\|^2 := \mathbb{E}\|w^* - w_t\|^2$. We may assume without loss of generality that $f(\theta, \cdot; z)$ is μ-strongly *convex* and that $w_{t+1} = \Pi_{\mathcal{W}}[w_t - \frac{1}{\beta_w}\left(\frac{1}{m}\sum_{i=1}^m \nabla_w f(\theta_t, w_t; z_i) + v_t\right)] := \Pi_{\mathcal{W}}[w_t - \frac{1}{\beta_w}\left(\nabla h(w_t) + v_t\right)] := \Pi_{\mathcal{W}}[w_t - \frac{1}{\beta_w}\nabla\tilde{h}(w_t)]$. Now,

$$\begin{aligned}
\mathbb{E}\|w_{t+1} - w^*\|^2 &= \mathbb{E}\left\|\Pi_{\mathcal{W}}[w_t - \frac{1}{\beta_w}\nabla\tilde{h}(w_t)] - w^*\right\|^2 \leqslant \mathbb{E}\left\|w_t - \frac{1}{\beta_w}\nabla\tilde{h}(w_t) - w^*\right\|^2 \\
&= \mathbb{E}\|w_t - w^*\|^2 + \frac{1}{\beta_w^2}\left[\mathbb{E}\|\nabla h(w_t)\|^2 + d_w\sigma_w^2\right] - \frac{2}{\beta_w}\mathbb{E}\left\langle w_t - w^*, \nabla\tilde{h}(w_t)\right\rangle \\
&\leqslant \mathbb{E}\|w_t - w^*\|^2 + \frac{1}{\beta_w^2}\left[\mathbb{E}\|\nabla h(w_t)\|^2 + d_w\sigma_w^2\right] - \frac{2}{\beta_w}\mathbb{E}\left[F(\theta_t, w_t) - F(\theta_t, w^*) + \frac{\mu}{2}\|w_t - w^*\|^2\right] \\
&\leqslant \delta_t\left(1 - \frac{\mu}{\beta_w}\right) - \frac{2}{\beta_w}\mathbb{E}\left[F(\theta_t, w_t) - F(\theta_t, w^*)\right] + \frac{\mathbb{E}\|\nabla h(w_t)\|^2}{\beta_w^2} + \frac{d_w\sigma_w^2}{\beta_w^2}.
\end{aligned}$$

Further,

$$\begin{aligned}
\mathbb{E}\|\nabla h(w_t)\|^2 &= \mathbb{E}\left[\|\nabla h(w_t) - \nabla_w F(\theta_t, w_t)\|^2 + \|\nabla_w F(\theta_t, w_t)\|^2\right] \\
&\leqslant \frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}} + \mathbb{E}\|\nabla_w F(\theta_t, w_t)\|^2 \\
&\leqslant \frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}} + 2\beta_w[F(\theta_t, w_t) - F(\theta_t, w^*(\theta_t))],
\end{aligned}$$

using independence and Theorem 12 plus Lipschitz continuity of f in the first inequality and Theorem 13 (plus Assumption 1 part 5) in the second inequality. This implies

$$\mathbb{E}\|w_{t+1} - w^*\|^2 \leqslant \delta_t\left(1 - \frac{1}{\kappa_w}\right) + \frac{1}{\beta_w^2}\left[d_w\sigma_w^2 + \frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}}\right]. \tag{7}$$

Therefore,

$$\begin{aligned}
\delta_{t+1} &= \mathbb{E}\|w_{t+1} - w^*(\theta_t) + w^*(\theta_t) - w^*(\theta_{t+1})\|^2 \\
&\leqslant \left(1 + \frac{1}{2\kappa_w - 1}\right) \mathbb{E}\|w_{t+1} - w^*(\theta_t)\|^2 + 2\kappa_w \mathbb{E}\|w^*(\theta_t) - w^*(\theta_{t+1})\|^2 \\
&\leqslant \left(1 + \frac{1}{2\kappa_w - 1}\right)\left(1 - \frac{1}{\kappa_w}\right)\delta_t + \frac{2}{\beta_w^2}\left[d_w\sigma_w^2 + \frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}}\right] + 2\kappa_w \mathbb{E}\|w^*(\theta_t) - w^*(\theta_{t+1})\|^2 \\
&\leqslant \left(1 + \frac{1}{2\kappa_w - 1}\right)\left(1 - \frac{1}{\kappa_w}\right)\delta_t + \frac{2}{\beta_w^2}\left[d_w\sigma_w^2 + \frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}}\right] + 2\kappa_w\kappa_{\theta w}^2 \mathbb{E}\|\theta_t - \theta_{t+1}\|^2 \\
&\leqslant \left(1 + \frac{1}{2\kappa_w - 1}\right)\left(1 - \frac{1}{\kappa_w}\right)\delta_t + \frac{2}{\beta_w^2}\left[d_w\sigma_w^2 + \frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}}\right] \\
&\quad + 4\kappa_w\kappa_{\theta w}^2\eta_\theta^2\left[\mathbb{E}\|\nabla_\theta F(\theta_t, w_t) - \nabla\Phi(\theta_t)\|^2 + \|\nabla\Phi(\theta_t)\|^2 + d_\theta\sigma_t^2\right] \\
&= \left(1 + \frac{1}{2\kappa_w - 1}\right)\left(1 - \frac{1}{\kappa_w}\right)\delta_t + \frac{2}{\beta_w^2}\left[d_w\sigma_w^2 + \frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}}\right] \\
&\quad + 4\kappa_w\kappa_{\theta w}^2\eta_\theta^2\left[\mathbb{E}\|\nabla_\theta F(\theta_t, w_t) - \nabla_\theta F(\theta_t, w^*(\theta_t)\|^2 + \|\nabla\Phi(\theta_t)\|^2 + d_\theta\sigma_t^2\right] \\
&\leqslant \left(1 + \frac{1}{2\kappa_w - 1}\right)\left(1 - \frac{1}{\kappa_w}\right)\delta_t + \frac{2}{\beta_w^2}\left[d_w\sigma_w^2 + \frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}}\right] \\
&\quad + 4\kappa_w\kappa_{\theta w}^2\eta_\theta^2\left[\beta_{\theta w}^2\mathbb{E}\|w_t - w^*(\theta_t)\|^2 + \|\nabla\Phi(\theta_t)\|^2 + d_\theta\sigma_t^2\right],
\end{aligned}$$

by Young's inequality, (7), and Theorem 11. Since $\left(1 + \frac{1}{2\kappa_w - 1}\right)\left(1 - \frac{1}{\kappa_w}\right) \leqslant 1 - \frac{1}{2\kappa_w}$, we obtain

$$\delta_{t+1} \leqslant \left(1 - \frac{1}{2\kappa_w} + 4\kappa_w\kappa_{\theta w}^2\eta_\theta^2\beta_{\theta w}^2\right)\delta_t + \frac{2}{\beta_w^2}\left[d_w\sigma_w^2 + \frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}}\right] + 4\kappa_w\kappa_{\theta w}^2\eta_\theta^2\left[\|\nabla\Phi(\theta_t)\|^2 + d_\theta\sigma_t^2\right],$$

as desired. ■

We are now prepared to prove Theorem 10.

Proof [Proof of Theorem 10] **Privacy:** This is an easy consequence of Theorem 1 in Abadi et al. (2016) (with precise constants obtained from the proof therein, as in Bassily et al. (2019)) applied to both the min (descent in θ) and max (ascent in w) updates. Unlike Abadi et al. (2016), we don't clip the gradients here before adding noise, but the Lipschitz continuity assumptions (Assumption 1 parts 1 and 2) imply that the ℓ_2-sensitivity of the gradient updates in lines 4 and 5 of Algorithm 2 are nevertheless bounded by $2L_\theta/m$ and $2L_w/m$, respectively. Thus, Theorem 1 in Abadi et al. (2016) still applies.
Convergence: Denote $\zeta := 1 - \frac{1}{2\kappa_w} + 4\kappa_w\kappa_{\theta w}^2\eta_\theta^2\beta_{\theta w}^2$, $\delta_t = \mathbb{E}\|w^*(\theta_t) - w_t\|^2$, and

$$C_t := \frac{2}{\beta_w^2}\left(\frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}} + d_w\sigma_w^2\right) + 4\kappa_w\kappa_{\theta w}^2\eta_\theta^2\left(\mathbb{E}\|\nabla\Phi(\theta_t)\|^2 + d_\theta\sigma_\theta^2\right),$$

so that Theorem 16 reads as

$$\delta_t \leqslant \zeta\delta_{t-1} + C_{t-1} \tag{8}$$

for all $t \in [T]$. Applying (8) recursively, we have

$$\begin{aligned}
\delta_t &\leqslant \zeta^t \delta_0 + \sum_{j=0}^{t-1} C_{t-j-1}\zeta^j \\
&\leqslant \zeta^t D^2 + 4\kappa_w \kappa_{\theta w}^2 \eta_\theta^2 \sum_{j=0}^{t-1} \zeta^{t-1-j} \mathbb{E}\|\nabla \Phi(\theta_j)\|^2 \\
&\quad + \left(\sum_{j=0}^{t-1} \zeta^{t-1-j}\right)\left[\frac{2}{\beta_w^2}\left(\frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}} + d_w\sigma_w^2\right) + 4\kappa_w\kappa_{\theta w}^2\eta_\theta^2 d_\theta \sigma_\theta^2\right].
\end{aligned}$$

Combining this inequality with Theorem 15, we get

$$\begin{aligned}
\mathbb{E}\Phi(\theta_t) &\leqslant \mathbb{E}\left[\Phi(\theta_{t-1}) - \frac{3}{8}\eta_\theta \|\nabla\Phi(\theta_{t-1})\|^2\right] + \frac{5}{8}\eta_\theta\left(d_\theta\sigma_\theta^2 + \frac{4L_\theta^2}{m}\mathbb{1}_{\{m<n\}}\right) \\
&\quad + \frac{5}{8}\eta_\theta\beta_{\theta w}^2\Bigg\{\zeta^t D^2 + 4\kappa_w\kappa_{\theta w}^2\eta_\theta^2 \sum_{j=0}^{t-1}\zeta^{t-1-j}\mathbb{E}\|\nabla\Phi(\theta_j)\|^2 \\
&\quad + \left(\sum_{j=0}^{t-1}\zeta^{t-1-j}\right)\left[\frac{2}{\beta_w^2}\left(\frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}} + d_w\sigma_w^2\right) + 4\kappa_w\kappa_{\theta w}^2\eta_\theta^2 d_\theta\sigma_\theta^2\right]\Bigg\}.
\end{aligned}$$

Summing over all $t \in [T]$ and re-arranging terms yields

$$\begin{aligned}
\mathbb{E}\Phi(\theta_T) &\leqslant \Phi(\theta_0) - \frac{3}{8}\eta_\theta \sum_{t=1}^{T}\mathbb{E}\|\nabla\Phi(\theta_{t-1})\|^2 + \frac{5}{8}\eta_\theta T\left(d_\theta\sigma_t^2 + \frac{4L_\theta^2}{m}\mathbb{1}_{\{m<n\}}\right) + \frac{5}{8}\eta_\theta\beta_{\theta w}^2\left(\sum_{t=1}^{T}\zeta^t\right)D^2 \\
&\quad + 4\eta_\theta^3\beta_{\theta w}^2\kappa_w\kappa_{\theta w}^2 \sum_{t=1}^{T}\sum_{j=0}^{t-1}\zeta^{t-1-j}\mathbb{E}\|\nabla\Phi(\theta_j)\|^2 \\
&\quad + \frac{5}{8}\left(\sum_{t=1}^{T}\sum_{j=0}^{t-1}\zeta^{t-1-j}\right)\eta_\theta\beta_{\theta w}^2\left[\frac{2}{\beta_w^2}\left(\frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}} + d_w\sigma_w^2\right) + 4\kappa_w\kappa_{\theta w}^2\eta_\theta^2 d_\theta\sigma_\theta^2\right].
\end{aligned}$$

Now, $\zeta \leqslant 1 - \frac{1}{4\kappa_w}$, which implies that

$$\begin{aligned}
\sum_{t=1}^{T}\zeta^t &\leqslant 4\kappa_w \quad \text{and} \\
\sum_{t=1}^{T}\sum_{j=0}^{t-1}\zeta^{t-1-j} &\leqslant 4\kappa_w T.
\end{aligned}$$

Hence

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla\Phi(\theta_t)\|^2 &\leqslant \frac{3[\Phi(\theta_0) - \mathbb{E}\Phi(\theta_T)]}{\eta_\theta T} + \frac{5}{3}\left(d_\theta\sigma_\theta^2 + \frac{4L_\theta^2}{m}\mathbb{1}_{\{m<n\}}\right) + \frac{7\beta_{\theta w}^2 D^2\kappa_w}{T} \\
&\quad + \frac{48\eta_\theta^2\beta_{\theta w}^2\kappa_w^2\kappa_{\theta w}^2}{T}\left(\sum_{t=1}^{T}\mathbb{E}\|\nabla\Phi(\theta_t)\|^2\right) \\
&\quad + 8\kappa_w\beta_{\theta w}^2\frac{2}{\beta_w^2}\left(\frac{4L_w^2}{m}\mathbb{1}_{\{m<n\}} + d_w\sigma_w^2\right) + 32\beta_{\theta w}^2\kappa_w^2\kappa_{\theta w}^2\eta_\theta^2 d_\theta\sigma_\theta^2.
\end{aligned}$$

Since $\eta_\theta^2\beta_{\theta w}^2\kappa_w^2\kappa_{\theta w}^2 \leqslant \frac{1}{256}$, we obtain

$$\mathbb{E}\|\nabla\Phi(\hat{\theta}_T)\|^2 \lesssim \frac{\Delta_\Phi\kappa_w}{T}\left(\beta_\theta + \beta_{\theta w}\kappa_{\theta w}\right) + \frac{d_\theta L_\theta^2 T\ln(1/\delta)}{\epsilon^2 n^2} + \frac{1}{m}\mathbb{1}_{\{m<n\}}\left(L_\theta^2 + \frac{\kappa_w\beta_{\theta w}^2 L_w^2}{\beta_w^2}\right) + \frac{\kappa_w\beta_{\theta w}^2 L_w^2 d_w T\ln(1/\delta)}{\beta_w^2\epsilon^2 n^2} + \frac{\beta_{\theta w}^2 D^2\kappa_w}{T}.$$

Our choice of T then implies

$$\mathbb{E}\|\nabla\Phi(\hat{\theta}_T)\|^2 \lesssim \sqrt{\Delta_\Phi\left(\beta_\theta + \beta_{\theta w}\kappa_{\theta w}\right)\kappa_w + \kappa_w\beta_{\theta w}^2 D^2)}\left[\frac{L_\theta\sqrt{d_\theta\ln(1/\delta)}}{\epsilon n} + \left(\frac{\beta_{\theta w}\sqrt{\kappa_w}}{\beta_w}\right)\frac{L_w\sqrt{d_w\ln(1/\delta)}}{\epsilon n}\right] + \frac{\mathbb{1}_{\{m<n\}}}{m}\left(L_\theta^2 + \frac{\kappa_w\beta_{\theta w}^2 L_w^2}{\beta_w^2}\right).$$

Finally, our choice of sufficiently large m yields the last claim in Theorem 10. ■

E.2. Proof of Theorem 5

Theorem 5 is an easy consequence of Theorem 9, which we proved in the above subsection:

Theorem 17 (Re-statement of Theorem 5) *Assume the loss function $\ell(\cdot, x, y)$ and $\mathcal{F}(x,\cdot)$ are Lipschitz continuous with Lipschitz gradient for all (x,y), and $\widehat{P}_S(r) \geqslant \rho > 0\ \forall\ r \in [k]$. In Algorithm 1, choose $\mathcal{W}$ to be a sufficiently large ball that contains $W^*(\theta) := \operatorname{argmax}_W \widehat{F}(\theta, W)$ for every θ in some neighborhood of $\theta^* \in \operatorname{argmin}_\theta \max_W \widehat{F}(\theta, W)$. Then there exist algorithmic parameters such that the (ϵ,δ)-DP Algorithm 1 returns $\hat{\theta}_T$ with*

$$\mathbb{E}\|\nabla FERMI(\hat{\theta}_T)\|^2 = \mathcal{O}\left(\frac{\sqrt{\max(d_\theta, kl)\ln(1/\delta)}}{\epsilon n}\right),$$

treating $D = diameter(\mathcal{W})$, λ, ρ, l, and the Lipschitz and smoothness parameters of ℓ and $\mathcal{F}$ as constants.

Proof By Theorem 9, it suffices to show that $f(\theta, W; z_i) := \ell(\theta, x_i, y_i) + \lambda\widehat{\psi}_i(\theta, W)$ is Lipschitz continuous with Lipschitz gradient in both the θ and W variables for any $z_i = (x_i, y_i, s_i)$, and that $f(\theta,\cdot; z_i)$ is strongly concave. We assumed $\ell(\cdot, x_i, y_i)$ is Lipschitz continuous with Lipschitz gradient. Further, the work of Lowy et al. (2022) showed that $f(\theta,\cdot; z_i)$ is strongly concave. Thus, it suffices to show that $\widehat{\psi}_i(\theta, W)$ is Lipschitz continuous with Lipschitz gradient. This clearly holds by Lemma 7, since $\mathcal{F}(x,\cdot)$ is Lipschitz continuous with Lipschitz gradient and $W \in \mathcal{W}$ is bounded. ■

Appendix F. Numerical Experiments: Additional Details and Results

F.1. Measuring Demographic Parity and Equalized Odds Violation

We used the expressions given in (9) and (10) to measure the demographic parity violation and the equalized odds violation respectively. We denote $\mathcal{Y}$ to be the set of all possible

output classes and $\mathcal{S}$ to be the classes of the sensitive attribute. $P[E]$ denotes the empirical probability of the occurrence of an event E.

$$\max_{y' \in \mathcal{Y}, s_1, s_2 \in \mathcal{S}} \left| P[\widehat{y} = y' | s = s_1] - P[\widehat{y} = y' | s = s_2] \right| \tag{9}$$

$$\begin{aligned} \max_{y' \in \mathcal{Y}, s_1, s_2 \in \mathcal{S}} \max(& \left| P[\widehat{y} = y' | s = s_1, y = y'] - P[\widehat{y} = y' | s = s_2, y = y'] \right|, \\ & \left| P[\widehat{y} = y' | s = s_1, y \neq y'] - P[\widehat{y} = y' | s = s_2, y \neq y'] \right|) \end{aligned} \tag{10}$$

F.2. Tabular Datasets

F.2.1. Model Description and Experimental Details

Demographic Parity: We split each dataset in a 3:1 train:test ratio. We preprocess the data similar to Hardt et al. (2016a) and use a simple logistic regression model with a sigmoid output $O = \sigma(Wx+b)$ which we treat as conditional probabilities $p(\widehat{y} = i|x)$. The predicted variables and sensitive attributes are both binary in this case across all the datasets. We analyze fairness-accuracy trade-offs with four different values of $\epsilon \in \{0.5, 1, 3, 9\}$ for each dataset. We compare against state-of-the-art algorithms proposed in Tran et al. (2021a) and (the demographic parity objective of) Tran et al. (2021b). The tradeoff curves for DP-FERMI were generated by sweeping across different values for $\lambda \in [0, 2.5]$. The learning rates for the descent and ascent, η_θ and η_w, remained constant during the optimization process and were chosen from $[0.005, 0.01]$. Batch size was 1024. We tuned the ℓ_2 diameter of the projection set $\mathcal{W}$ and θ-gradient clipping threshold in $[1, 5]$ in order to generate stable results with high privacy (i.e. low ϵ). Each model was trained for 200 epochs. The results displayed are averages over 15 trials (random seeds) for each value of ϵ.
Equalized Odds: We replicated the experimental setup described above, but we took ℓ_2 diameter of $\mathcal{W}$ and the value of gradient clipping for θ to be in $[1, 2]$. Also, we only tested three values of $\epsilon \in \{0.5, 1, 3\}$.

F.2.2. Description of Datasets

Adult Income Dataset: This dataset contains the census information about the individuals. The classification task is to predict whether the person earns more than 50k every year or not. We followed a preprocessing approach similar to Lowy et al. (2022). After preprocessing, there were a total of 102 input features from this dataset. The sensitive attribute for this work in this dataset was taken to be gender. This dataset consists of around 48,000 entries spanning across two CSV files, which we combine and then we take the train-test split of 3:1.
Retired Adult Income Dataset: The Retired Adult Income Dataset proposed by Ding et al. (2021) is essentially a superset of the Adult Income Dataset which attempts to counter some caveats of the Adult dataset. The input and the output attributes for this dataset is the same as that of the Adult Dataset and the sensitive attribute considered in this work is the same as that of the Adult. This dataset contains around 45,000 entries.
Parkinsons Dataset: In the Parkinsons dataset, we use the part of the dataset which had the UPRDS scores along with some of the features of the recordings obtained from

individuals affected and not affected with the Parkinsons disease. The classification task was to predict from the features whether the UPDRS score was greater than the median score or not. After preprocessing, there were a total of 19 input features from this dataset and the sensitive attribute for this dataset was taken to be gender. This dataset contains around 5800 entries in total. We took a train-test split of 3:1.

Credit Card Dataset: This dataset contains the financial data of users in a bank in Taiwan consisting of their gender, education level, age, marital status, previous bills, and payments. The assigned classification task is to predict whether the person defaults their credit card bills or not, essentially making the task if the clients were credible or not. We followed a preprocessing approach similar to Lowy et al. (2022). After preprocessing, there were a total of 85 input features from this dataset. The sensitive attribute for this dataset was taken to be gender. This dataset consists of around 30,000 entries from which we take the train-test split of 3:1.

UTK-Face Dataset: This dataset is a large scale image dataset containing with an age span from 0 to 116. The dataset consists of over 20,000 face images with details of age, gender, and ethnicity and covers large variation in pose, facial expression, illumination, occlusion, resolution. We consider the age classification task with 9 age groups similar to the experimental setup in Tran et al. (2022). We consider the sensitive attribute to be the ethnicity which consists of 5 different classes.

F.2.3. Demographic Parity

Adult Results: See Fig. 6 for complete results on Adult data set.

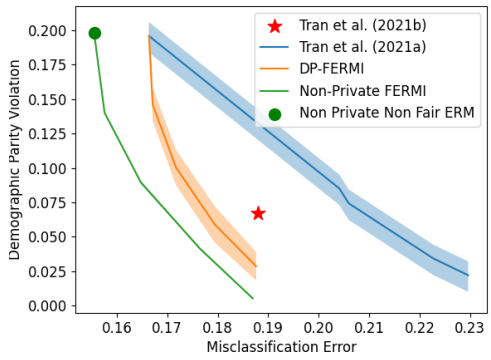
Retired Adult Results: See Fig. 7 for our results on Retired Adult Dataset. The results are qualitatively similar to the reusults reported in the main body: our algorithm (DP-FERMI) achieves the most favorable fairness-accuracy tradeoffs across all privacy levels.

Credit Card Results: See Fig. 8 for our results on Credit Card Dataset. DP-FERMI offers superior fairness-accuracy-privacy profile compared to all applicable baselines.

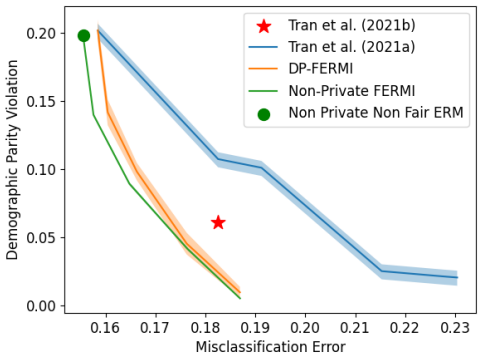
Additional Results for Parkinsons Dataset: More results for Parkinsons are shown in Fig. 9. DP-FERMI offers the best performance.

F.2.4. Equalized Odds

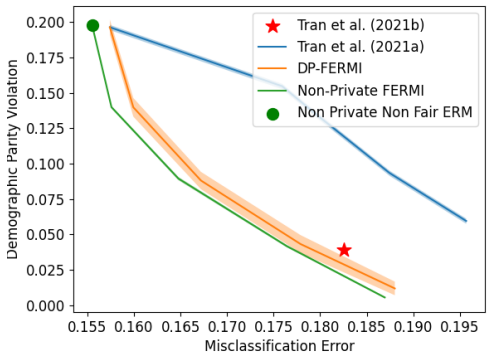
Equalized Odds Variation of DP-FERMI Algorithm: The (FERMI obj.) minimizes the Exponential Renyi Mutual Information (ERMI) between the output and the sensitive attributes which essentially leads to a reduced demographic parity violation. The equalized-odds condition is more constrained and enforces the demographic parity condition for data grouped according to labels. For the equalized-odds, the ERMI between the predicted and the sensitive attributes is minimized conditional to each of the label present in the output variable of the dataset. So, the FERMI regularizer is split into as many parts as the number of labels in the output. This enforces each part of the FERMI regularizer to minimize the ERMI while an output label is given/constant. Each part has its own unique W that is maximized in order to create a stochastic estimator for the ERMI with respect to each of the output labels.



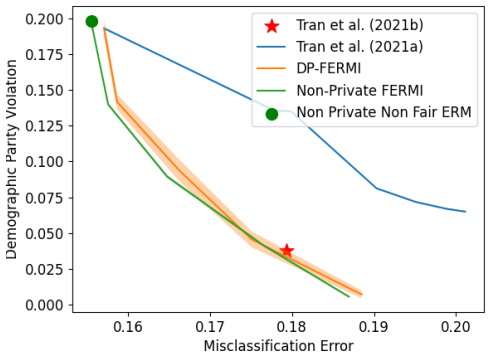
(a) $\epsilon = 0.5$



(b) $\epsilon = 1$



(c) $\epsilon = 3$



(d) $\epsilon = 9$

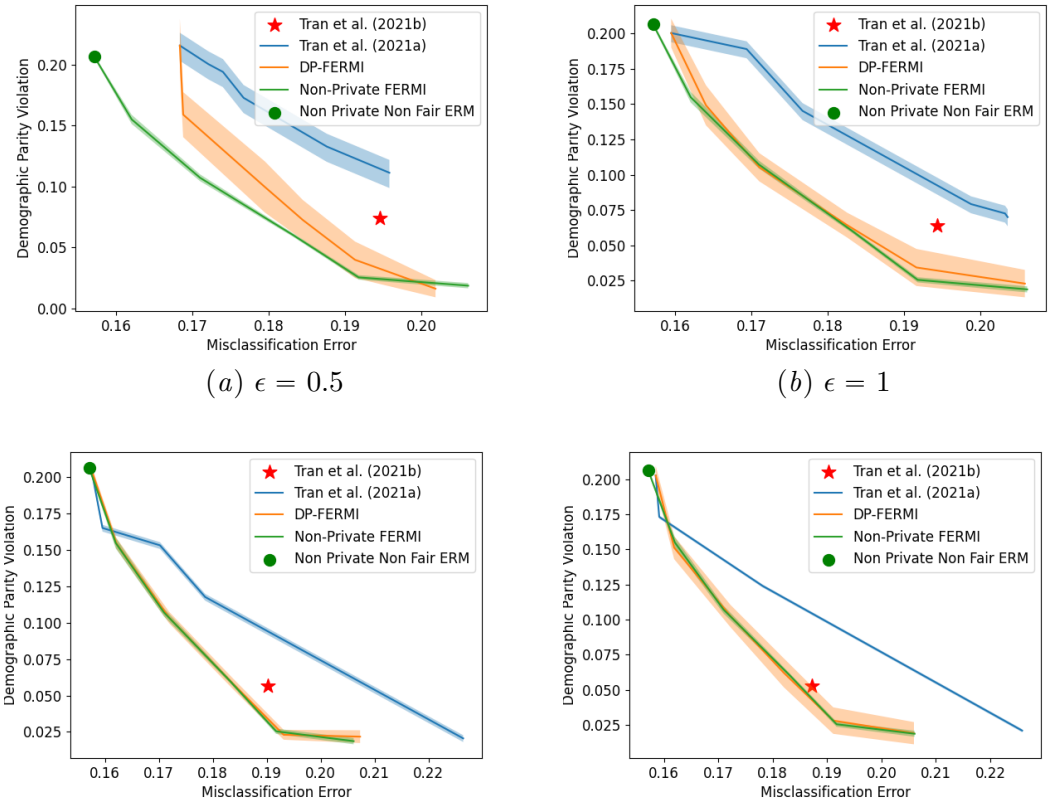
Figure 6: Private, Fair (Demographic Parity) logistic regression on Adult Dataset.

Adult Results: Results for the equalized odds version of DP-FERMI on Adult dataset are shown in Fig. 10. Our approach outperforms the previous state-of-the-art methods.

Retired Adult Results: (Initial) Results for the equalized odds version of DP-FERMI on the retired-adult dataset are shown in Fig. 11. Our approach outperforms Tran et al. (2021b) and we are currently tuning our non-private and/or non-fair versions of our models along with Jagielski et al. (2019).

F.3. Image Dataset (UTK-Face)

We split the dataset in a 3:1 train:test ratio. Batch size was 64. 128 x 128 normalized images were used as input for our model. We tuned the ℓ_2 diameter of $\mathcal{W}$ and the value of gradient clipping for θ to be in $[1, 2]$ and learning rates for the descent and ascent, η_θ and η_w, remained constant during the optimization process and were chosen as 0.001 and 0.005 respectively. We analyze the fairness-accuracy trade-offs with five different values of $\epsilon \in \{10, 25, 50, 100, 200\}$. The results displayed were averaged over observations obtained from 5 different randomly chosen seeds on each configuration of ϵ and a dataset. Each model was trained for 150 epochs. The tradeoff curves for this set of experiments were obtained by sweeping across different values for $\lambda \in [0, 500]$. Complete results are given in Fig. 12.



(a) $\epsilon = 0.5$ (b) $\epsilon = 1$

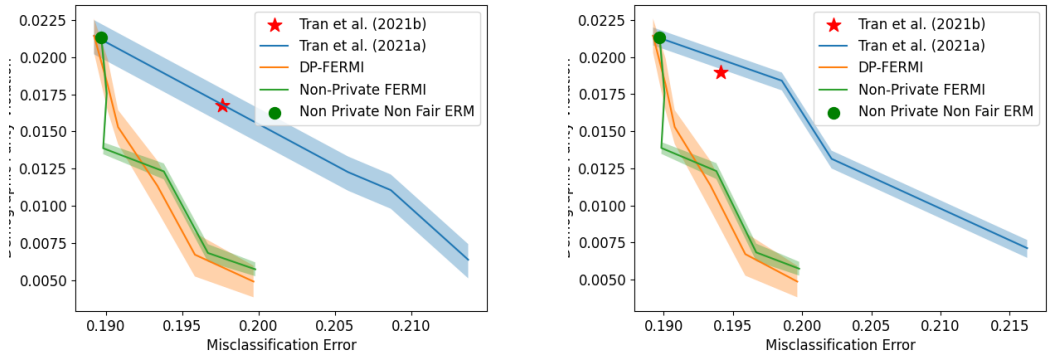
(c) $\epsilon = 3$ (d) $\epsilon = 9$

Figure 7: Private, fair logistic regression on the Retired Adult Dataset

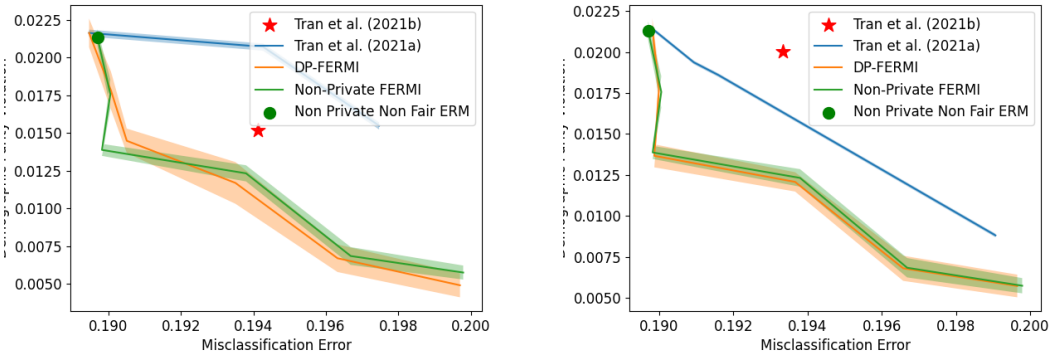
Appendix G. Societal Impacts

In this paper, we considered the socially consequential problem of *privately* learning *fair* models from sensitive data. Motivated by the lack of *scalable* private fair learning methods in the literature, e developed the first differentially private (DP) fair learning algorithm that is guaranteed to converge with small batches (*stochastic optimization*). We hope that our method will be used to help companies, governments, and other organizations to responsibly use sensitive, private data. Specifically, we hope that our DP-FERMI algorithm will be useful in reducing discrimination in algorithmic decision-making while simultaneously preventing leakage of sensitive user data. The stochastic nature of our algorithm might be especially appealing to companies that are using very large models and datasets. On the other hand, there are also some important limitations of our method that need to be considered before deployment.

One caveat of our work is that we have assumed that the given data set contains fair and accurate labels. For example, if gender is the sensitive attribute and "likelihood of repaying a loan" is the target, then we assume that the training data *accurately* describes everyone's financial history without discrimination. If training data is biased against a certain demographic group, then it is possible that our algorithm could *amplify* (rather

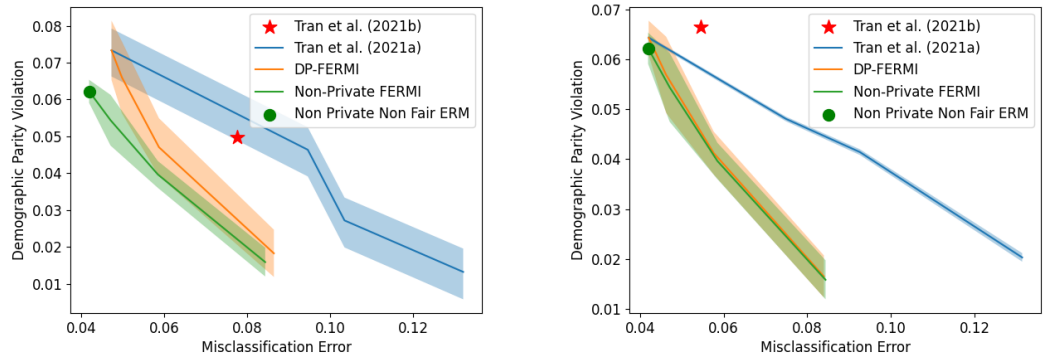


(a) $\epsilon = 0.5$ (b) $\epsilon = 1$



(c) $\epsilon = 3$ (d) $\epsilon = 9$

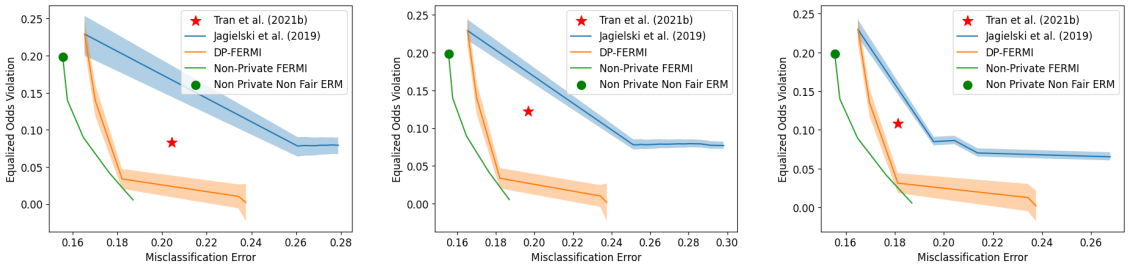
Figure 8: Private, fair (demographic parity) logistic regression on the Credit Card Dataset



(a) $\epsilon = 0.5$ (b) $\epsilon = 9$

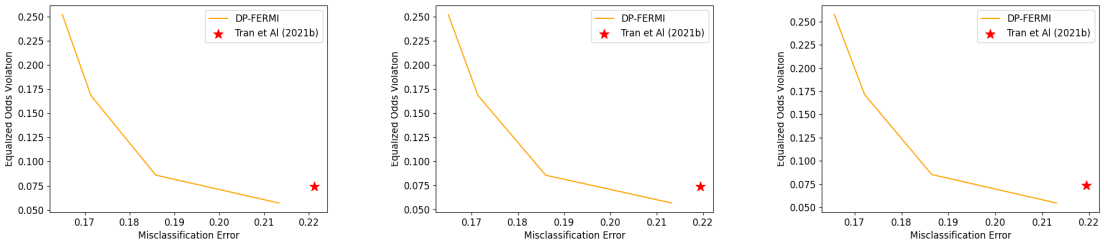
Figure 9: Private, Fair (Demogrpahic Parity) Logistic regression on Parkinsons Dataset

than mitigate) unfairness. See e.g. Kilbertus et al. (2020); Bechavod et al. (2019) for further discussion.



(a) $\epsilon = 0.5$ (b) $\epsilon = 1$ (c) $\epsilon = 3$

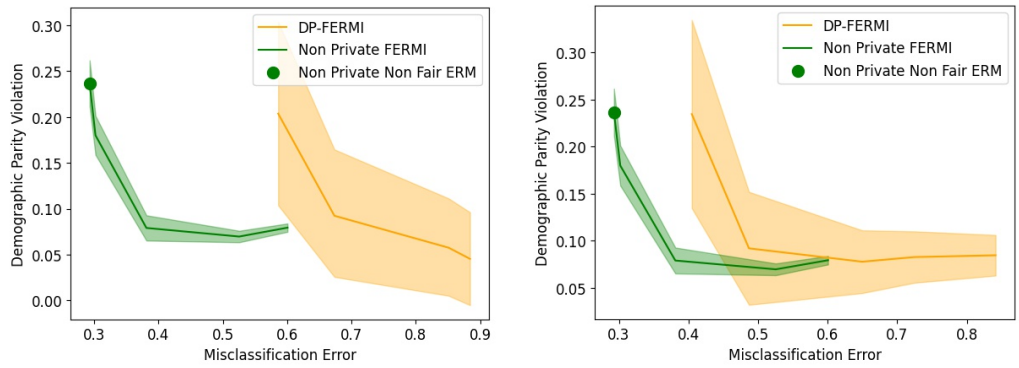
Figure 10: Results obtained for applying DP-FERMI with equalized odds violation on logistic regression on the Adult Dataset



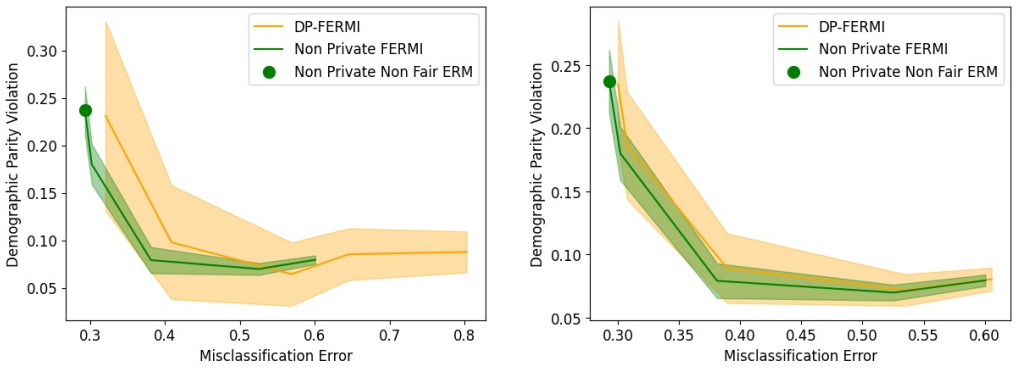
(a) $\epsilon = 0.5$ (b) $\epsilon = 1$ (c) $\epsilon = 3$

Figure 11: Results obtained for applying DP-FERMI with equalized odds violation on logistic regression on the Retired Adult Dataset

Another important practical consideration is how to weigh/value the different desiderata (privacy, fairness, and accuracy) when deploying our method. As shown in prior works (e.g., Cummings et al. (2019)) and re-enforced in the present work, there are fundamental tradeoffs between fairness, accuracy, and privacy: improvements in one generally come at a cost to the other two. Determining the relative importance of each of these three desiderata is a critical question that lacks a clear or general answer. Depending on the application, one might be seriously concerned with either discrimination or privacy attacks, and should calibrate ϵ and λ accordingly. Or, perhaps very high accuracy is necessary for a particular task, with privacy and/or fairness as an afterthought. In such a case, one might want to start with very large ϵ and small λ to ensure high accuracy, and then gradually shrink ϵ and/or increase λ to improve privacy/fairness until training accuracy dips below a critical threshold. A thorough and rigorous exploration of these issues could be an interesting direction for future work.



(a) $\epsilon = 10$ (b) $\epsilon = 25$



(c) $\epsilon = 50$ (d) $\epsilon = 100$

Figure 12: DP-FERMI on a Deep CNN for Image Classification on UTK-Face