A Missing Proofs

Theorem 1. *Let ℓ be a twice differentiable and convex loss function and consider the output perturbation mechanism described above. Then, the excessive risk gap for group $a \in \mathcal{A}$ is approximated by:*

$$\xi_a \approx \frac{1}{2}\Delta_\ell^2\sigma^2 \left|\mathrm{Tr}(\boldsymbol{H}_\ell^a) - \mathrm{Tr}(\boldsymbol{H}_\ell)\right|, \tag{3}$$

where $\boldsymbol{H}_\ell^a = \nabla^2_{\boldsymbol{\theta}^*} \sum_{(X,A,Y)\in D_a} \ell(f_{\boldsymbol{\theta}^*}(X), Y)$ is the Hessian matrix of the loss function at the optimal parameters vector $\boldsymbol{\theta}^*$, computed using the group data D_a, $\boldsymbol{H}^\ell$ is the analogous Hessian computed using the population data D, and $\mathrm{Tr}(\cdot)$ denotes the trace of a matrix.

Proof. Recall that the output perturbation mechanism adds Gaussian noise directly to the non-private model parameters $\boldsymbol{\theta}^*$ to obtain the private parameters $\tilde{\boldsymbol{\theta}}$. Denote $\psi \sim \mathcal{N}(0, \mathbf{I}\Delta_\ell^2\sigma^2)$ the random noise vector with the same size as $\boldsymbol{\theta}^*$. Then $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta}^* + \psi$. Using a second order Taylor expansion around $\boldsymbol{\theta}^*$ the private risk function for group $a \in \mathcal{A}$ is approximated as follows:

$$\mathcal{L}(\tilde{\boldsymbol{\theta}}, D_a) = \mathcal{L}(\boldsymbol{\theta}^* + \psi, D_a) \approx \mathcal{L}(\boldsymbol{\theta}^*, D_a) + \psi^T \nabla_{\boldsymbol{\theta}^*}\mathcal{L}(\boldsymbol{\theta}^*, D_a) + \frac{1}{2}\psi^T \boldsymbol{H}_\ell^a \psi. \tag{7}$$

Taking the expectation with respect to ψ on both sides of the above equation results in:

$$\mathbb{E}\left[\mathcal{L}(\tilde{\boldsymbol{\theta}}, D_a)\right] \approx \mathcal{L}(\boldsymbol{\theta}^*, D_a) + \mathbb{E}\left[\psi^T \nabla_{\boldsymbol{\theta}^*}\mathcal{L}(\boldsymbol{\theta}^*, D_a)\right] + \frac{1}{2}\mathbb{E}\left[\psi^T \boldsymbol{H}_\ell^a \psi\right] \tag{8a}$$

$$= \mathcal{L}(\boldsymbol{\theta}^*, D_a) + \frac{1}{2}\mathbb{E}\left[\psi^T \boldsymbol{H}_\ell^a \psi\right] \tag{8b}$$

$$= \mathcal{L}(\boldsymbol{\theta}^*, D_a) + \frac{1}{2}\sum_{i,j}\mathbb{E}\left[\psi_i (\boldsymbol{H}_\ell^a)_{ij}\psi_j\right] \tag{8c}$$

$$= \mathcal{L}(\boldsymbol{\theta}^*, D_a) + \frac{1}{2}\sum_{i}\mathbb{E}\left[\psi_i^2\right](\boldsymbol{H}_\ell^a)_{ii} \tag{8d}$$

$$= \mathcal{L}(\boldsymbol{\theta}^*, D_a) + \frac{1}{2}\Delta_\ell^2\sigma^2\,\mathrm{Tr}\left(\boldsymbol{H}_\ell^a\right), \tag{8e}$$

where equation (8b) follows from linearity of expectation, by observing that $\nabla_{\boldsymbol{\theta}^*}\mathcal{L}(\boldsymbol{\theta}^*, D_a)$ is a constant term, and that ψ is a 0-mean noise variable, thus, $\mathbb{E}[\psi] = \mathbf{0}^T \times \nabla_{\boldsymbol{\theta}^*}\mathcal{L}(\boldsymbol{\theta}^*, D_a) = \mathbf{0}^T$. Equation (8c) follows by definition of Hessian matrix, where $(H_\ell^a)_{ij}$ denotes the entry with indices i and j of the matrix. Equation (8d) follows from that $\psi_i \perp \psi_j$, for all $i \neq j$, and Equation (8e) from that for a random variable X, $\mathbb{E}[X^2] = (\mathbb{E}[X])^2 + \mathrm{Var}[X]$, and $\mathrm{Var}[\psi_i] = \Delta_\ell^2\sigma^2\ \forall i$ and definition of Trace of a matrix.

Therefore, the group and population excessive risks are approximated as:

$$R_a(\boldsymbol{\theta}) = \mathbb{E}\left[\mathcal{L}(\tilde{\boldsymbol{\theta}}, D_a)\right] - \mathcal{L}(\boldsymbol{\theta}^*, D_a) \approx \frac{1}{2}\Delta_\ell^2\sigma^2\,\mathrm{Tr}\left(\boldsymbol{H}_\ell^a\right) \tag{9}$$

$$R(\boldsymbol{\theta}) = \mathbb{E}\left[\mathcal{L}(\tilde{\boldsymbol{\theta}}, D)\right] - \mathcal{L}(\boldsymbol{\theta}^*, D) \approx \frac{1}{2}\Delta_\ell^2\sigma^2\,\mathrm{Tr}\left(\boldsymbol{H}_\ell\right). \tag{10}$$

The claim follows by definition of excessive risk gap (Equation 2) subtracting Equation (9) from (10) in absolute values. □

Corollary 1. *Consider the ERM problem for a linear model $f_{\boldsymbol{\theta}}(X) \stackrel{def}{=} \boldsymbol{\theta}^T X$, with L_2 loss function i.e., $\ell(f_{\boldsymbol{\theta}}(X), Y) = (f_{\boldsymbol{\theta}}(X) - Y)^2$. Then, output perturbation does not guarantee pure fairness.*

Proof. First, notice that for an L_2 loss function the trace of Hessian loss for a group $a \in \mathcal{A}$ is:

$$\mathrm{Tr}(\boldsymbol{H}_\ell^a) = \mathbb{E}_{x\sim D_a}\|X\|.$$

Therefore, from Theorem 1, the excessive risk gap ξ_a for group a is:

$$\xi_a \approx \frac{1}{2}\Delta_\ell^2\sigma^2 \left|\mathbb{E}_{x\sim D_a}\|X\| - \mathbb{E}_{x\sim D}\|X\|\right|. \tag{11}$$

Notice that ξ_a is larger than zero only if the average input norm of group a is different with that of the population one. Since this condition cannot be guaranteed in general, the output perturbation mechanism for a linear ERM model under the L_2 loss does not guarantee pure fairness. □

Corollary 2. *If for any two groups $a, b \in \mathcal{A}$ their average group norms $\mathbb{E}_{X_a \sim D_a}\|X_a\| = \mathbb{E}_{X_b \sim D_b}\|X_b\|$ have identical values, then output perturbation with L_2 loss function provides pure fairness.*

Proof. The above follows directly by observing that, when the average norms of any two groups have identical values, $\xi_a \approx 0$ for any group $a \in \mathcal{A}$ (see Equation (11)), and thus the average norm of each group also coincide with that of the population. □

The above indicates that as long as the average group norm is invariant across different groups, then output perturbation mechanism provides pure fairness.

Theorem 2. *Consider the ERM problem* (L) *with loss ℓ twice differentiable with respect to the model parameters. The expected loss $\mathbb{E}[\mathcal{L}(\boldsymbol{\theta}_{t+1}; D_a)]$ of group $a \in \mathcal{A}$ at iteration $t+1$, is approximated as:*

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1}; D_a)\right] \approx \underbrace{\mathcal{L}(\boldsymbol{\theta}_t; D_a) - \eta \langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D \rangle + \frac{\eta^2}{2}\mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B\right]}_{\textit{non-private term}} \tag{4}$$

$$+ \underbrace{\eta\left(\langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D \rangle - \langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_D \rangle\right) + \frac{\eta^2}{2}\left(\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] - \mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B\right]\right)}_{\textit{private term due to clipping}} \tag{R_a^{clip}}$$

$$+ \underbrace{\frac{\eta^2}{2}\mathrm{Tr}(\boldsymbol{H}_\ell^a) C^2 \sigma^2}_{\textit{private term due to noise}} \tag{R_a^{noise}}$$

where the expectation is taken over the randomness of the private noise and the mini-batch selection, and the terms $\boldsymbol{g}_Z$ and $\bar{\boldsymbol{g}}_Z$ denote, respectively, the average non-private and private gradients over subset Z of D at iteration t (the iteration number is dropped for ease of notation).

Proof. The proof of Theorem 2 relies on the following two second order Taylor approximations: **(1)** The first approximates the ERM loss at iteration $t+1$ under non-private training, i.e., $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta \boldsymbol{g}_B$, where $B \subseteq D$ denotes the minibatch. **(2)** The second approximates expected ERM loss under private-training, i.e $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta(\bar{\boldsymbol{g}}_B + \psi)$ where $\psi \sim \mathcal{N}(0, \boldsymbol{I} C^2 \sigma^2)$. Finally, the result is obtained by taking the difference of these approximations under private and non-private training.

1. Non-private term. The non private term of Theorem 2 can be derived using second order Taylor approximation as follows:

$$\mathcal{L}(\theta_{t+1}, D_a) = \mathcal{L}(\theta_t - \eta \boldsymbol{g}_B, D_a) \approx \mathcal{L}(\theta_t, D_a) - \eta \langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_B \rangle + \frac{\eta^2}{2} \boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B \tag{12}$$

Taking the expectation with respect to the randomness of the mini-batch B selection on both sides of the above approximation, and noting that $\mathbb{E}[\boldsymbol{g}_B] = \boldsymbol{g}_D$ (as B is selected randomly from dataset D), it follows:

$$\mathbb{E}[\mathcal{L}(\theta_{t+1}, D_a)] \approx \mathcal{L}(\theta_t, D_a) - \eta \mathbb{E}[\langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_B \rangle] + \frac{\eta^2}{2}\mathbb{E}[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B] \tag{13a}$$

$$= \mathcal{L}(\theta_t, D_a) - \eta \langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D \rangle + \frac{\eta^2}{2}\mathbb{E}[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B]. \tag{13b}$$

2. Private term (due to both clipping and noise). Consider the private update in DP-SGD, i.e., $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta(\bar{\boldsymbol{g}}_B + \psi)$. Again, applying a second order Taylor approximation around $\boldsymbol{\theta}_t$ allows us to estimate the expected private loss at iteration $t+1$ as:

$$\mathcal{L}(\boldsymbol{\theta}_{t+1}, D_a) = \mathcal{L}\left(\boldsymbol{\theta}_t - \eta(\bar{\boldsymbol{g}}_B + \psi), D_a\right)$$

$$\approx \mathcal{L}\left(\boldsymbol{\theta}_t, D_a\right) - \eta \left\langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_B + \psi \right\rangle + \frac{\eta^2}{2}\left(\bar{\boldsymbol{g}}_B + \psi\right)^T \boldsymbol{H}_\ell^a \left(\bar{\boldsymbol{g}}_B + \psi\right) \tag{14a}$$

$$= \mathcal{L}\left(\boldsymbol{\theta}_t, D_a\right) - \eta \left\langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_B \right\rangle - \eta \left\langle \boldsymbol{g}_{D_a}, \psi \right\rangle + \frac{\eta^2}{2} \bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B \tag{14b}$$

$$+ \frac{\eta^2}{2}\left(\psi^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B + \bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \psi + \psi^T \boldsymbol{H}_\ell^a \psi\right)$$

Taking the expectation with respect to the randomness of the mini-batch B selection and with respect to the randomness of noise ψ on both sides of the above equation gives:

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1}, D_a)\right] \approx \mathbb{E}\Big[\mathcal{L}(\boldsymbol{\theta}_t, D_a) - \eta\left\langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_B\right\rangle - \eta\left\langle \boldsymbol{g}_{D_a}, \psi\right\rangle + \frac{\eta^2}{2}\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B \tag{15a}$$

$$+ \frac{\eta^2}{2}\left(\psi^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B + \bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \psi + \psi^T \boldsymbol{H}_\ell^a \psi\right)\Big]$$

$$= \mathcal{L}(\boldsymbol{\theta}_t, D_a) - \eta\left\langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_B\right\rangle - \eta\left\langle \boldsymbol{g}_{D_a}, \mathbb{E}[\psi]\right\rangle + \frac{\eta^2}{2}\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] \tag{15b}$$

$$+ \frac{\eta^2}{2}\left(\mathbb{E}\left[\psi\right]^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B + \bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \mathbb{E}[\psi] + \mathbb{E}\left[\psi^T \boldsymbol{H}_\ell^a \psi\right]\right)$$

$$= \mathcal{L}(\boldsymbol{\theta}_t, D_a) - \eta\left\langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_B\right\rangle + \frac{\eta^2}{2}\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] + \frac{\eta^2}{2}\mathbb{E}\left[\psi^T \boldsymbol{H}_\ell^a \psi\right] \tag{15c}$$

$$= \mathcal{L}(\boldsymbol{\theta}_t, D_a) - \eta\left\langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_B\right\rangle + \frac{\eta^2}{2}\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] + \frac{\eta^2}{2}\operatorname{Tr}\left(\boldsymbol{H}_\ell^a\right) C^2\sigma^2, \tag{15d}$$

where (15b), and (15c) follow from linearity of expectation and from that $\mathbb{E}[\psi] = 0$, since ψ is a 0-mean noise variable. Equation (15d) follows from that,

$$\mathbb{E}\left[\psi^T \boldsymbol{H}_\ell^a \psi\right] = \mathbb{E}\left[\sum_{i,j} \psi_i (\boldsymbol{H}_\ell^a)_{i,j} \psi_j\right] = \sum_i \mathbb{E}\left[\psi_i^2 (\boldsymbol{H}_\ell^a)_{i,i}\right] = \operatorname{Tr}\left(\boldsymbol{H}_\ell^a\right) C^2\sigma^2,$$

since $\mathbb{E}[\psi^2] = \mathbb{E}[\psi]^2 + \mathrm{Var}[\psi]$ and $\mathbb{E}[\psi] = 0$ while $\mathrm{Var}[\psi] = C^2\sigma^2$.

Note that in the above approximation (Equation (15)), the component

$$\mathcal{L}(\boldsymbol{\theta}_t, D_a) - \eta\langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_B\rangle + \frac{\eta^2}{2}\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] \tag{16}$$

is associated to the SGD update step in which gradients have been clipped to the clipping bound value C, i.e. $\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \eta(\bar{\boldsymbol{g}}_B)$.

Next, the component

$$\frac{\eta^2}{2}\operatorname{Tr}\left(\boldsymbol{H}_\ell^a\right) C^2\sigma^2 \tag{17}$$

is associated to the SGD update step in which the noise ψ is added to the gradients.

If we take the difference between the approximation associated with the non-private loss term, obtained in Equation 13b, with that associated with the private loss term, obtained in Equation 15d, we can derive the effect of a single step of (private) DP-SGD compared to its non-private counterpart:

$$\mathbb{E}\left[\mathcal{L}(\boldsymbol{\theta}_{t+1}; D_a)\right] \approx \mathcal{L}(\boldsymbol{\theta}_t; D_a) - \eta\left\langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D\right\rangle + \frac{\eta^2}{2}\mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B\right] \tag{18a}$$

$$+ \eta\left(\left\langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D\right\rangle - \left\langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_D\right\rangle\right) + \frac{\eta^2}{2}\left(\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B\right] - \mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B\right]\right) \tag{18b}$$

$$+ \frac{\eta^2}{2}\operatorname{Tr}\left(\boldsymbol{H}_\ell^a\right) C^2\sigma^2. \tag{18c}$$

In the above,

- The components in Equation (18a) are associated with the loss under non-private training (see again Equation 13b);
- The components in Equation (18b) is associated with for excessive risk due to gradient clipping;
- Finally, the components in Equation (18c) is associated with the excessive risk due to noise addition.

□

Next, the paper proves Theorem 3. This result is based on the following assumptions.

Assumption 1. *[Convexity and Smoothness assumption] For a group $a \in \mathcal{A}$, its empirical loss function $\mathcal{L}(\boldsymbol{\theta}, D_a)$ is convex and β_a-smooth.*

Assumption 2. *Let $B \subseteq D$ be a subset of the dataset D, and consider a constant $\varepsilon \geq 0$. Then, the variance associated with the gradient norms of a random mini-batch B, $\sigma_B^2 = \mathrm{Var}\left[\|\boldsymbol{g}_B\|\right] \leq \varepsilon$ as well as that associated with its clipped counterpart, $\bar{\sigma}_B^2 = \mathrm{Var}\left[\|\bar{\boldsymbol{g}}_B\|\right] \leq \varepsilon$.*

The assumption above can be satisfied when the mini-batch size is large enough. For example, the variance is 0 when $|B| = |D|$.

Assumption 3. *The learning rate used in DP-SDG η is upper bounded by quantity $1/\max_{z \in \mathcal{A}} \beta_z$.*

Theorem 3. *Let $p_z = |D_z|/|D|$ be the fraction of training samples in group $z \in \mathcal{A}$. For groups $a, b \in \mathcal{A}$, $R_a^{clip} > R_b^{clip}$ whenever:*

$$\left\|\boldsymbol{g}_{D_a}\right\|\left(p_a - \frac{p_a^2}{2}\right) \geq \frac{5}{2}C + \left\|\boldsymbol{g}_{D_b}\right\|\left(1 + p_b + \frac{p_b^2}{2}\right). \tag{5}$$

To ease notation, the statement of the theorem above uses $\varepsilon = 0$ (See Assumption 2) but the theorem can be generalized to any $\varepsilon \geq 0$.

The following Lemmas are introduced to aid the proof of Theorem 3.

Lemma 1. *Consider the ERM problem* (L) *solved with DP-SGD with clipping value C. The following average clipped per-sample gradients $\bar{\boldsymbol{g}}_Z$, where $Z \subseteq D$, has norm at most C.*

Proof. The result follows by triangle inequality:

$$\begin{aligned}
\left\|\bar{\boldsymbol{g}}_{D_Z}\right\| &= \left\|\frac{1}{|D_Z|}\sum_{i \in D_Z} \bar{\boldsymbol{g}}_i\right\| \\
&\leq \frac{1}{|D_Z|}\sum_{i \in D_Z} \|\bar{\boldsymbol{g}}_i\| \\
&= \frac{1}{|D_Z|}\sum_{i \in D_Z} \left\|\boldsymbol{g}_i \min\left(1, \frac{C}{\|\boldsymbol{g}_i\|}\right)\right\| \\
&\leq \frac{1}{|D_Z|}\sum_{i \in D_Z} C = C.
\end{aligned}$$

□

The next Lemma derives a lower and an upper bound for the component $\mathbb{E}[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B] - \mathbb{E}[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B]$, which appears in the excessive risk term due to clipping R_a^{clip} for some group $a \in \mathcal{A}$.

Lemma 2. *Consider the ERM problem* (L) *with loss ℓ, solved with DP-SGD with clipping value C. Further, let $\varepsilon = 0$ (see Assumption 2). For any group $a \in \mathcal{A}$, the following inequality holds:*

$$-\beta_a \|\boldsymbol{g}_D\|^2 \leq \mathbb{E}[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B] - \mathbb{E}[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B] \leq \beta_a C^2 \tag{19}$$

Proof. Consider a group $a \in \mathcal{A}$. By the convexity assumption of the loss function, the Hessian $\boldsymbol{H}_\ell^a$ is a positive semi-definite matrix, i.e., for all real vectors of appropriate dimensions $\boldsymbol{v}$, it follows that $\boldsymbol{v}^T \boldsymbol{H}_\ell^a \boldsymbol{v} \geq 0$.

Therefore, for a subset $B \subseteq D$ the following inequalities hold:

- $\bar{\boldsymbol{g}}_B \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B \geq 0$,
- $\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B \geq 0$.

Additionally their expectations $\mathbb{E}[\bar{\boldsymbol{g}}_B \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B]$ and $\mathbb{E}[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B]$ are non-negative.

By the smoothness property of the loss function, $\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B \leq \beta_a \|\bar{\boldsymbol{g}}_B\|^2$, thus:

$$\begin{aligned}
\mathbb{E}[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B] &\leq \beta_a \mathbb{E}\left[\|\bar{\boldsymbol{g}}_B\|^2\right] && \text{(20a)} \\
&= \beta_a (\mathbb{E}[\|\bar{\boldsymbol{g}}_B\|]^2 + \mathrm{Var}\left[\|\bar{\boldsymbol{g}}_B\|\right]) && \text{(20b)} \\
&\leq \beta_a (C^2 + \bar{\sigma}_B^2) && \text{(20c)} \\
&\leq \beta_a (C^2 + \varepsilon), && \text{(20d)}
\end{aligned}$$

where Equation (20b) follows from that $\mathbb{E}[X^2] = (\mathbb{E}[X])^2 + \text{Var}[X]$, Equation (20c) is due to Lemma 1, and finally, the last inequality is due to Assumption 2.

Therefore, since $\varepsilon = 0$ by assumption of the Lemma, the following upper bound holds:

$$\mathbb{E}[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B] - \mathbb{E}[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B] \leq \beta_a C^2. \tag{21}$$

Next, notice that

$$\begin{aligned}
\mathbb{E}[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B] - \mathbb{E}[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B] &\geq -\mathbb{E}[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B] && \text{(22a)}\\
&\geq -\mathbb{E}[\beta_a \|\boldsymbol{g}_B\|^2] && \text{(22b)}\\
&= -\beta_a \left(\mathbb{E}[\|\boldsymbol{g}_B\|]^2 + \text{Var}\left[\|\boldsymbol{g}_B\|\right]\right) && \text{(22c)}\\
&= -\beta_a \|\boldsymbol{g}_D\|^2, && \text{(22d)}
\end{aligned}$$

where the inequality in Equation (22a) follows since both terms on the left hand side of the Equation are non negative. Equation (22b) follows by smoothness assumption of the loss function. Equation (22c) follows by definition of expectation of a random variable, since $\mathbb{E}[X]^2 = \mathbb{E}[X^2] + \text{Var}[X]$. Finally, Equation (22d) follows from that $\text{Var}[\boldsymbol{g}_B] \leq \varepsilon = 0$ by Assumption 2, and that $\varepsilon = 0$ by assumption of the Lemma, and thus the norms $\|g_B\| = \|g_D\|$ and, thus, $\mathbb{E}[\boldsymbol{g}_B] = \boldsymbol{g}_D$. Therefore if follows:

$$-\beta_a \|\boldsymbol{g}_D\|^2 \leq \mathbb{E}[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B] - \mathbb{E}[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B]. \tag{23}$$

which concludes the proof. □

Again, the above uses $\varepsilon = 0$ to simplify notation, but the results generalize to the case when $\varepsilon > 0$. In such a case, the bounds require slight modifications to involve the term ε.

Lemma 3. *Let $a, b \in \mathcal{A}$ be two groups. Consider the ERM problem* (L) *solved with DP-SGD with clipping value C and learning rate $\eta \leq 1/\max_{a\in\mathcal{A}} \beta_a$. Then, the difference on the excessive risk due to clipping $R^a_{clip} - R^b_{clip}$ is lower bounded as:*

$$R^a_{clip} - R^b_{clip} \geq \eta \left(\langle \boldsymbol{g}_{D_a} - \boldsymbol{g}_{D_b}, \boldsymbol{g}_D - \bar{\boldsymbol{g}}_D \rangle - \frac{1}{2}(\|\boldsymbol{g}_D\|^2 + C^2) \right). \tag{24}$$

Proof. Recall that $B \subseteq D$ is the mini-batch during the resolution of DP-SGD. Using the lower and upper bounds obtained from Lemma 2, it follows:

$$\begin{aligned}
R^a_{clip} - R^b_{clip} &= \eta \left(\langle \boldsymbol{g}_{D_a}, \boldsymbol{g}_D \rangle - \langle \boldsymbol{g}_{D_a}, \bar{\boldsymbol{g}}_D \rangle \right) + \frac{\eta^2}{2} \left(\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B \right] - \mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B \right] \right) && \text{(25a)}\\
&\quad - \eta \left(\langle \boldsymbol{g}_{D_b}, \boldsymbol{g}_D \rangle - \langle \boldsymbol{g}_{D_b}, \bar{\boldsymbol{g}}_D \rangle \right) - \frac{\eta^2}{2} \left(\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^b \bar{\boldsymbol{g}}_B \right] - \mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^b \boldsymbol{g}_B \right] \right)\\
&= \eta \langle \boldsymbol{g}_{D_a} - \boldsymbol{g}_{D_b}, \boldsymbol{g}_D - \bar{\boldsymbol{g}}_D \rangle + \frac{\eta^2}{2} \left(\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^a \bar{\boldsymbol{g}}_B \right] - \mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^a \boldsymbol{g}_B \right] \right) && \text{(25b)}\\
&\quad - \frac{\eta^2}{2} \left(\mathbb{E}\left[\bar{\boldsymbol{g}}_B^T \boldsymbol{H}_\ell^b \bar{\boldsymbol{g}}_B \right] - \mathbb{E}\left[\boldsymbol{g}_B^T \boldsymbol{H}_\ell^b \boldsymbol{g}_B \right] \right)\\
&\geq \eta \langle \boldsymbol{g}_{D_a} - \boldsymbol{g}_{D_b}, \boldsymbol{g}_D - \bar{\boldsymbol{g}}_D \rangle - \frac{\eta^2}{2} \beta_a \|\boldsymbol{g}_D\|^2 - \frac{\eta^2}{2} \beta_b C^2 && \text{(25c)}\\
&\geq \eta \langle \boldsymbol{g}_{D_a} - \boldsymbol{g}_{D_b}, \boldsymbol{g}_D - \bar{\boldsymbol{g}}_D \rangle - \frac{\eta^2}{2} \max_{z\in\mathcal{A}} \beta_z (\|\boldsymbol{g}_D\|^2 + C^2) && \text{(25d)}\\
&\geq \eta \left(\langle \boldsymbol{g}_{D_a} - \boldsymbol{g}_{D_b}, \boldsymbol{g}_D - \bar{\boldsymbol{g}}_D \rangle - \frac{1}{2}(\|\boldsymbol{g}_D\|^2 + C^2) \right), && \text{(25e)}
\end{aligned}$$

where the inequality (25c) follows as a consequence of Lemma 2, and the inequality (25e) since $\eta \leq \frac{1}{\max_{a\in\mathcal{A}} \beta_a}$. □

Proof of Theorem 3. We want to show that $R^a_{clip} > R^b_{clip}$ given Equation (5). Since, by Lemma 3 the difference $R^a_{clip} - R^b_{clip}$ is lower bounded – see Equation (24), the following shows that the right hand side of Equation (24) is positive, that is:

$$\langle \boldsymbol{g}_{D_a} - \boldsymbol{g}_{D_b}, \boldsymbol{g}_D - \bar{\boldsymbol{g}}_D \rangle - \frac{1}{2} \left(\|\boldsymbol{g}_D\|^2 + C^2 \right) > 0. \tag{26}$$

First, observe that the gradients at the population level can be expressed as a combination of the gradients of the two groups a and b in the dataset: $\boldsymbol{g}_D = p_a \boldsymbol{g}_{D_a} + p_b \boldsymbol{g}_{D_b}$ and $\bar{\boldsymbol{g}} = p_a \bar{\boldsymbol{g}}_{D_a} + p_b \bar{\boldsymbol{g}}_{D_b}$.

By algebraic manipulation, and the above, Equation (26) can thus be expressed as:

$$(26) = \langle \boldsymbol{g}_{D_a} - \boldsymbol{g}_{D_b}, p_a \boldsymbol{g}_{D_a} + p_b \boldsymbol{g}_{D_b} - p_a \bar{\boldsymbol{g}}_{D_a} - p_b \bar{\boldsymbol{g}}_{D_b} \rangle - \frac{1}{2}(\|\boldsymbol{g}_{D_a} p_a + \boldsymbol{g}_{D_b} p_b\|^2 + C^2) \tag{27a}$$

$$= (p_a \|\boldsymbol{g}_{D_a}\|^2 + p_b \boldsymbol{g}_{D_a}^T \boldsymbol{g}_{D_b} - p_a \boldsymbol{g}_{D_a}^T \bar{\boldsymbol{g}}_{D_a} - p_b \boldsymbol{g}_{D_a}^T \bar{\boldsymbol{g}}_{D_b} - p_a \boldsymbol{g}_{D_b}^T \boldsymbol{g}_{D_a} - p_b \|\boldsymbol{g}_{D_b}\|^2 \tag{27b}$$

$$+ p_a \boldsymbol{g}_{D_b}^T \bar{\boldsymbol{g}}_{D_a} + p_b \boldsymbol{g}_{D_b}^T \bar{\boldsymbol{g}}_{D_b} - \frac{1}{2}(p_a^2 \|\boldsymbol{g}_{D_a}\|^2 + 2 p_a p_b \boldsymbol{g}_{D_a} \boldsymbol{g}_{D_b} + p_b^2 \|\boldsymbol{g}_{D_b}\|^2 + C^2).$$

Noting that for any vector $\boldsymbol{x}, \boldsymbol{y}$ the following inequality hold: $\boldsymbol{x}^T \boldsymbol{y} \geq -\|\boldsymbol{x}\|\|\boldsymbol{y}\|$, all the inner products in the above expression can be replaced by their lower bounds:

$$(26) \geq \|\boldsymbol{g}_{D_a}\| \left(\|\boldsymbol{g}_{D_a}\| p_a (1 - \frac{p_a}{2}) - p_b \|\boldsymbol{g}_{D_b}\| - p_a C - p_b C - p_a \|\boldsymbol{g}_{D_b}\| \right) \tag{28a}$$

$$- \|\boldsymbol{g}_{D_b}\| \left(\|\boldsymbol{g}_{D_b}\| p_b (1 + \frac{p_b}{2}) + p_a C + p_b C \right) - \frac{1}{2} C^2$$

$$= \|\boldsymbol{g}_{D_a}\| \left(\|\boldsymbol{g}_{D_a}\| p_a (1 - \frac{p_a}{2}) - (p_b + p_a)(\|\boldsymbol{g}_{D_b}\| + C) \right) \tag{28b}$$

$$- \|\boldsymbol{g}_{D_b}\| \left(\|\boldsymbol{g}_{D_b}\| p_b (1 + \frac{p_b}{2}) + (p_a + p_b) C \right) - \frac{1}{2} C^2$$

$$= \|\boldsymbol{g}_{D_a}\| \left(\|\boldsymbol{g}_{D_a}\| p_a (1 - \frac{p_a}{2}) - \|\boldsymbol{g}_{D_b}\| - C) \right) - \|\boldsymbol{g}_{D_b}\| \left(\|\boldsymbol{g}_{D_b}\| p_b (1 + \frac{p_b}{2}) + C \right) - \frac{1}{2} C^2 \tag{28c}$$

where the last equality is because $p_a + p_b = 1$, by assumption of the dataset having exactly two groups.

By theorem assumption, $\|\boldsymbol{g}_{D_a}\| p_a (1 - \frac{p_a}{2}) \geq \frac{5}{2} C + \|\boldsymbol{g}_{D_b}\| (1 + p_b + \frac{p_b^2}{2})$. It follows that $\|\boldsymbol{g}_{D_a}\| > \|\boldsymbol{g}_{D_b}\|$ and $\|\boldsymbol{g}_{D_a}\| > C$. Combined with Equation (28c) it follows that:

$$(28c) = \|\boldsymbol{g}_{D_a}\| \left(\|\boldsymbol{g}_{D_a}\| p_a (1 - \frac{p_a}{2}) - \|\boldsymbol{g}_{D_b}\| - C - \|\boldsymbol{g}_{D_b}\| p_b (1 + \frac{p_b}{2}) - C \right) - \frac{1}{2} C^2 \tag{29a}$$

$$= \|\boldsymbol{g}_{D_a}\| \left(\|\boldsymbol{g}_{D_a}\| p_a (1 - \frac{p_a}{2}) - 2C - \|\boldsymbol{g}_{D_b}\| (1 + p_b + \frac{p_b^2}{2}) \right) - \frac{1}{2} C^2 \tag{29b}$$

$$\geq \|\boldsymbol{g}_{D_a}\| \frac{C}{2} - \frac{1}{2} C^2 \tag{29c}$$

$$> 0, \tag{29d}$$

where the last equality is because $\|\boldsymbol{g}_{D_a}\| > C$. □

Theorem 4. *For groups $a, b \in \mathcal{A}$, $R_a^{noise} > R_b^{noise}$ whenever*

$$\mathrm{Tr}(\boldsymbol{H}_\ell^a) > \mathrm{Tr}(\boldsymbol{H}_\ell^b).$$

Proof. Suppose $\mathrm{Tr}(\boldsymbol{H}_\ell^a) > \mathrm{Tr}(\boldsymbol{H}_\ell^b)$. By definition of R_a^{noise} and R_b^{noise} from Theorem 2 it follows that:

$$R_a^{noise} = \frac{\eta^2}{2} \mathrm{Tr}(\boldsymbol{H}_\ell^a) C^2 \sigma^2 > \frac{\eta^2}{2} \mathrm{Tr}(\boldsymbol{H}_\ell^b) C^2 \sigma^2 = R_b^{noise},$$

which concludes the proof.

□

Theorem 5. *Consider a K-class classifier $\boldsymbol{f}_{\boldsymbol{\theta},k}$ ($k \in [K]$). For a given sample $X \sim D$, the term $\left(1 - \sum_{k=1}^K \boldsymbol{f}_{\boldsymbol{\theta},k}^2(X)\right)$ is maximized when $\boldsymbol{f}_{\boldsymbol{\theta},k}(X) = 1/K$ and minimized when $\exists k \in [K]$ s.t. $\boldsymbol{f}_{\boldsymbol{\theta},k}(X) = 1$ and $\boldsymbol{f}_{\boldsymbol{\theta},k'} = 0 \; \forall k' \in [K], k' \neq k$.*

Proof. Fix an input X of D and denote $y_k = \boldsymbol{f}_{\boldsymbol{\theta},k}(X) \in [0,1]$. Recall that y_k represents the likelihood of the prediction of input X to be associated with label k.

Note that, by Cauchy–Schwarz inequality

$$1 - \sum_{k=1}^{K} y_k^2 \leq 1 - K\left(\frac{\sum_i^K y_k}{K}\right)^2 \tag{30a}$$

$$= 1 - \frac{1}{K}, \tag{30b}$$

where Equation (30b) follows since $\sum_i^K y_k(X) = 1$. The above expression is maximized when

$$y_k = \boldsymbol{f}_{\theta,k}(X) = \frac{1}{K}.$$

Additionally, since $y_k \in [0,1]$ it follows that $y_k^2 \leq y_k$. Hence,

$$1 - \sum_{k=1}^{K} y_k^2 \geq 1 - \sum_{i=1}^{K} y_k = 0. \tag{31}$$

To hold, the equality above, it must exists $k \in [K]$ such that $y_k = \boldsymbol{f}_{\theta,k}(X) = 1$ and for any other $k' \in [K]$ with $k' \neq k$, $y_{k'} = \boldsymbol{f}_{\theta,k'} = 0$. □

Given the connection of the term $1 - \sum_{k=1}^{K}(1 - f_{\theta,k}^2(X))$ and the associated (trace of the) Hessian loss $\boldsymbol{H}_f$, the result above suggests that the trace of the Hessian is minimized (maximized) when the classifier is very confident (uncertain) about the prediction of $X \sim D$, i.e., when X is far (close) to the decision boundary.

B Experimental settings

Datasets The paper uses the following UCI datasets to support its claims:

1. **Adult** (Income) dataset, where the task is to predict if an individual has low or high income, and the group labels are defined by race: *White* vs *Non-White* [6].
2. **Bank** dataset, where the task is to predict if a user subscribes a term deposit or not and the group labels are defined by age: *people whose age is less than 60 years old vs the rest* [20].
3. **Wine** dataset, where the task is to predict if a given wine is of good quality, and the group labels are defined by wine color: *red vs white* [6].
4. **Abalone** dataset, where the task is to predict if a given abalone ring exceeds the median value, and the group labels are defined by gender: *female vs male* [6].
5. **Parkinsons** dataset, where the task is to predict if a patient has total UPDRS score that exceeds the median value, and the group labels are defined by gender: *female vs male* [19].
6. **Churn** dataset, where the task is to predict if a customer churned or not. The group labels are defined by on gender: *female vs male* [11].
7. **Credit Card** dataset, where the task is to predict if a customer defaults a loan or not. The group labels are defined by gender: *female vs male* [26].
8. **Stroke** dataset, where the task is to predict if a patient have had a stroke based on their physical conditions. The group labels are defined by gender: *female vs male* [1].

All datasets were processed by standardization so each feature has zero mean and unit variance.

Settings For output perturbation, the paper uses a Logistic regression model to obtain the optimal model parameters (we set the regularization parameter $\lambda = 1$) and add Gaussian noise to achieve privacy. The standard deviation of the noise required to the mechanism is determined following Balle and Wang [4].