

A APPENDIX

A.1 MPTP

Algorithm 1 Multi-party Training Process w.r.t. D (MPTP)

Insurer Input: data: $\{X_i, Y_i\}_{i=1}^n$, hypothesis class: \mathcal{H} (if obtain T via supervised learning)

Insurer Output: $\{\tilde{X}_i\}_{i=1}^n$

TTP Input: data: $\{\tilde{X}_i, Y_i, D_i\}$, hypothesis class: \mathcal{F} , risk function: $\mathcal{R}(f_1, \dots, f_{|\mathcal{D}|})$ (Eq. (1))

repeat

 train $f_1, \dots, f_{|\mathcal{D}|}$ by minimizing Eq. (1)

until Convergence

 compute $h^*(\tilde{X})$ using Eq. (2)

return $f_1, \dots, f_{|\mathcal{D}|}, h^*(\tilde{X})$

TTP Output: $f_1, \dots, f_{|\mathcal{D}|}, h^*(\tilde{X})$

A.2 MPTP-LDP

Algorithm 2 Multi-party Training Process w.r.t. S (MPTP-LDP)

Insurer Input: data: $\{X_i, Y_i\}_{i=1}^n$, hypothesis class: \mathcal{H} (if obtain T via supervised learning), hypothesis class: \mathcal{K} (if obtain X^* via supervised learning)

Insurer Output: $\{\tilde{X}_i\}_{i=1}^n, \{X_i^*\}_{i=1}^n$

TTP Input: data: $\{\tilde{X}_i, Y_i, S_i\}_{i=1}^n, \{X_i^*, S_i\}_{i=1}^n$, hypothesis class \mathcal{G} , risk function: $\forall k \in [n_1], \mathcal{R}(g_k) = \sum_{j=1}^m L(g_k(X_{k,j}^*, S_{k,j}))$ (see Lemma 4.4), hypothesis class: \mathcal{F} , risk function: $\mathcal{R}(f_1, \dots, f_{|\mathcal{D}|})$ (Eq. (6)),

if scenario 2 ($\pi, \bar{\pi}$ unknown) **then**

 compute $\hat{\pi}_k, \hat{\pi}_k, k \in [n_1]$ (by applying Lemma 4.4)

 compute \hat{C}_1 using $\hat{\pi}_k, \hat{\pi}_k, k \in [n_1]$ (by C_1 estimation procedure 4.3)

 compute $\hat{\pi}, \hat{\pi}$ using \hat{C}_1

 compute $\hat{\Pi}^{-1}$ using $\hat{\pi}, \hat{\pi}$

else

 compute Π^{-1} using $\pi, \bar{\pi}$

end if

repeat

 train $f_1, \dots, f_{|\mathcal{D}|}$ by minimizing Eq. (6)

until convergence

 compute $h^*(\tilde{X})$ using Eq. (2)

return $f_1, \dots, f_{|\mathcal{D}|}, h^*(\tilde{X})$

TTP Output: $f_1, \dots, f_{|\mathcal{D}|}, h^*(\tilde{X})$

B DEFERRED DISCUSSION ON ASSUMPTIONS

B.1 RESTRICTIONS ON ASSUMPTION A

The restriction of Assumption A relies on the type of generator (which will influence the tail distribution of $\hat{\pi}$) and the number of data within each group (which will influence the accuracy of $\hat{\pi}$). The condition in Assumption A is equivalent to:

$$\mathbb{P} \left(\frac{\left(1 - \frac{1}{|\mathcal{D}|}\right)^2}{t} > \left| \hat{\pi} - \frac{1}{|\mathcal{D}|} \right| \right) \leq \exp\left(\frac{-t}{K}\right),$$

when $K > 0$ is a constant.

Generally speaking, this assumption holds if $\hat{\pi}$ is inverse exponential distributed with a translation of $\frac{1}{|\mathcal{D}|}$, or having a lighter tail than the inverse exponential distribution that is

$$f_{\hat{\pi}}(t) \leq \frac{1}{K(t - \frac{1}{|\mathcal{D}|})^2} \exp\left(-\frac{1}{K|t - \frac{1}{|\mathcal{D}|}|}\right),$$

when t is close to $\frac{1}{|\mathcal{D}|}$, where $f_{\hat{\pi}}(t)$ is the pdf of $\hat{\pi}$. Especially, since a bounded distribution is also sub-exponential, if $|\hat{\pi} - \frac{1}{|\mathcal{D}|}| > \epsilon$, for some $\epsilon > 0$ condition is also satisfied. This will happen when the number of data within groups (m) is sufficiently large and $\pi - \frac{1}{|\mathcal{D}|}$ is large enough.

B.2 A GENERAL DISCUSSION

It is imperative to note that the support of our assumptions and theorem requires the availability of multiple independent datasets. when the observations are naturally organized in this manner, the application of the following assumptions is direct. Notice that the same concept applies to non-independent datasets with a mixing property (such as α -mixing). In such cases, we only need to use the Bernstein inequality under strong mixing conditions (Bousquet & Bousquet (2009), Chen & Louis (2008)).

According to the form of $\hat{C}_{1,k}$, the tail of this estimator is equivalent to the distribution of $\hat{\pi}_k = \max_{i \in [m]} \hat{g}_k(X_{k,i}^*)$ near $\frac{1}{|\mathcal{D}|}$. If \hat{g}_k is a good estimator as well as m is large enough, $\hat{\pi}_k$ will be concentrated near $\pi > \frac{1}{|\mathcal{D}|}$, which is guaranteed by Lemma 4.4. Especially when $\pi - \frac{1}{|\mathcal{D}|}$ is relatively large, it is reasonable to expect that $\hat{\pi}_k$ has a sparse distribution near $\frac{1}{|\mathcal{D}|}$, which implies that $\hat{C}_{1,k}$ has a sub-exponential tail (or even bounded). For Assumption B 4.3 notice that within every group $k \in [n_1]$, $\hat{\pi}_k$ are estimators for π , and thus $\hat{C}_{1,k}$ are plug-in estimator for C_1 . Since $\hat{C}_{1,k}$ are i.i.d., it is reasonable to assume $\hat{C}_{1,k}$ are “nearly” unbiased.

C DEFERRED PROOFS

C.1 LEMMA 4.2 PROOF

Lemma 4.2 Given the privacy parameter ϵ , minimizing the following risk (Risk-LDP) Eq. (6) under ϵ -LDP w.r.t. privatized sensitive attributes S is equivalent of minimizing Eq. (1) w.r.t. true sensitive attributes D at the population level:

$$\mathcal{R}^{LDP}(f_1, \dots, f_k) = \sum_{k=1}^{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} \left(\mathbf{\Pi}_{kj}^{-1} \mathbb{E}_{Y, \tilde{X} | S=j} [L(f_k(\tilde{X}), Y)] \cdot \sum_{l=1}^{|\mathcal{D}|} \mathbf{T}_{kl}^{-1} \mathbb{P}(S=l) \right), \quad (7)$$

where $\mathbf{\Pi}^{-1}$ and \mathbf{T}^{-1} are $|\mathcal{D}| \times |\mathcal{D}|$ row-stochastic matrices.

Proof. Step 1:

Since the ϵ -LDP randomization mechanism is independent of X, Y , therefore, the distribution of S is fully characterized by the privacy parameter ϵ and the distribution of D . Therefore, the distribution of S is deterministic once the privacy parameter ϵ and the distribution of D is given.

Step 2: Recover distributions w.r.t. D

Inspired by proposition 1 in Mozannar et al. (2020). Let $\mathcal{E}_1, \mathcal{E}_2$ be two probability events defined with respect to (\tilde{X}, Y, \hat{Y}) , then consider the following probability:

$$\begin{aligned} & \mathbb{P}(\mathcal{E}_1, \mathcal{E}_2 \mid S = d) \\ &= \sum_{d' \in D} \mathbb{P}(\mathcal{E}_1, \mathcal{E}_2 \mid S = d, D = d') \mathbb{P}(D = d' \mid S = d) \\ &= \sum_{d' \in D} \mathbb{P}(\mathcal{E}_1, \mathcal{E}_2 \mid D = d') \mathbb{P}(D = d' \mid S = d) \\ &= \sum_{d' \in D} \mathbb{P}(\mathcal{E}_1, \mathcal{E}_2 \mid D = d') \frac{\mathbb{P}(S = d \mid D = d') \mathbb{P}(D = d')}{\sum_{d'' \in D} \mathbb{P}(S = d \mid D = d'') \mathbb{P}(D = d'')} \\ &= \mathbb{P}(\mathcal{E}_1, \mathcal{E}_2 \mid D = d) \frac{\pi \mathbb{P}(D = d)}{\pi \mathbb{P}(D = d) + \sum_{d'' \setminus d} \bar{\pi} \mathbb{P}(D = d'')} + \sum_{d' \setminus d} \mathbb{P}(\mathcal{E}_1, \mathcal{E}_2 \mid D = d') \frac{\bar{\pi} \mathbb{P}(D = d')}{\pi \mathbb{P}(D = d) + \sum_{d'' \setminus d} \bar{\pi} \mathbb{P}(D = d'')}. \end{aligned}$$

Then, let $\mathcal{E}_1 = Y, \mathcal{E}_2 = \tilde{X}$, we obtain the following:

$$\begin{aligned} & \mathbb{P}(Y, \tilde{X} \mid S = d) \\ &= \sum_{d' \in D} \mathbb{P}(Y, \tilde{X} \mid S = d, D = d') \mathbb{P}(D = d' \mid S = d) \\ &= \sum_{d' \in D} \mathbb{P}(Y, \tilde{X} \mid D = d') \mathbb{P}(D = d' \mid S = d) \\ &= \sum_{d' \in D} \mathbb{P}(Y, \tilde{X} \mid D = d') \frac{\mathbb{P}(S = d \mid D = d') \mathbb{P}(D = d')}{\sum_{d'' \in D} \mathbb{P}(S = d \mid D = d'') \mathbb{P}(D = d'')} \\ &= \mathbb{P}(Y, \tilde{X} \mid D = d) \frac{\pi \mathbb{P}(D = d)}{\pi \mathbb{P}(D = d) + \sum_{d'' \setminus d} \bar{\pi} \mathbb{P}(D = d'')} + \sum_{d' \setminus d} \mathbb{P}(Y, \tilde{X} \mid D = d') \frac{\bar{\pi} \mathbb{P}(D = d')}{\pi \mathbb{P}(D = d) + \sum_{d'' \setminus d} \bar{\pi} \mathbb{P}(D = d'')}. \end{aligned}$$

Denote $p_d = \mathbb{P}(D = d)$, then let $\mathbf{\Pi}$ be the following $|\mathcal{D}| \times |\mathcal{D}|$ matrix with the following entries:

$$\begin{cases} \mathbf{\Pi}_{i,i} = \frac{\pi p_i}{\pi p_i + \sum_{d'' \setminus i} \bar{\pi} p_{d''}}, \text{ for } i \in D \\ \mathbf{\Pi}_{i,j} = \frac{\bar{\pi} p_j}{\pi p_i + \sum_{d'' \setminus i} \bar{\pi} p_{d''}}, \text{ for } i, j \in D \text{ s.t. } i \neq j \end{cases},$$

then we have the following system of linear equations:

$$\begin{bmatrix} \mathbb{P}(Y, \tilde{X} \mid S = 1) \\ \vdots \\ \mathbb{P}(Y, \tilde{X} \mid S = |\mathcal{D}|) \end{bmatrix} = \mathbf{\Pi} \begin{bmatrix} \mathbb{P}(Y, \tilde{X} \mid D = 1) \\ \vdots \\ \mathbb{P}(Y, \tilde{X} \mid D = |\mathcal{D}|) \end{bmatrix},$$

denote as $\mathbf{s}_1 = \mathbf{\Pi} \mathbf{d}_1$, where $\mathbf{s}_1 = \mathbb{P}(Y, \tilde{X} \mid S)$, $\mathbf{d}_1 = \mathbb{P}(Y, \tilde{X} \mid D)$.

Since $\mathbf{\Pi}$ is row-stochastic and invertible, we show that the entries of $\mathbf{\Pi}^{-1}$ take the following forms:

$$\begin{cases} \mathbf{\Pi}_{i,i}^{-1} = \frac{\pi + |\mathcal{D}| - 2}{|\mathcal{D}| \pi - 1} \frac{\pi p_i + \sum_{d'' \setminus i} \bar{\pi} p_{d''}}{p_i}, \text{ for } i \in D \\ \mathbf{\Pi}_{i,j}^{-1} = \frac{\pi - 1}{|\mathcal{D}| \pi - 1} \frac{\bar{\pi} p_i + \sum_{d'' \setminus i} \pi p_{d''}}{p_i}, \text{ for } i, j \in D \text{ s.t. } i \neq j \end{cases},$$

multiplying $\mathbf{\Pi}^{-1}$ on both side, we recovered

$$\begin{aligned} \mathbb{P}(Y, \tilde{X} \mid D = k) &= \sum_{j=1}^{|\mathcal{D}|} \mathbf{\Pi}_{kj}^{-1} \mathbb{P}(Y, \tilde{X} \mid S = j) \\ &= \mathbf{\Pi}_{k \cdot}^{-1} \mathbb{P}(Y, \tilde{X} \mid S) \end{aligned}$$

where $\mathbf{\Pi}_{k \cdot}^{-1}$ denotes the k^{th} row of $\mathbf{\Pi}^{-1}$.

However, there is still one component that we do need to estimate in order to recover the population distribution of $\mathbb{P}(Y, \tilde{X} \mid D)$. We need to further estimate $\mathbb{P}(D = d)$. Using the same technique, to estimate $\mathbb{P}(D = d)$, first write $P(S = d)$ in terms of the conditional probability of S given D as:

$$\begin{aligned} \mathbb{P}(S = d) &= \sum_{d' \in D} \mathbb{P}(S = d \mid D = d') \mathbb{P}(D = d') \\ &= \mathbb{P}(S = d \mid D = d) \mathbb{P}(D = d) + \sum_{d' \setminus d} \mathbb{P}(S = d \mid D = d') \mathbb{P}(D = d') \\ &= \pi p_d + \sum_{d' \setminus d} \bar{\pi} p_{d'}. \end{aligned}$$

Then we write the above expression in terms of a system of linear equations. Let \mathbf{T} be an $|\mathcal{D}| \times |\mathcal{D}|$ matrix with the following entries:

$$\begin{cases} \mathbf{T}_{i,i} = \pi, \text{ for } i \in D \\ \mathbf{T}_{i,j} = \bar{\pi}, \text{ for } i, j \in D \text{ s.t. } i \neq j \end{cases},$$

then we have the following system of linear equations:

$$\begin{bmatrix} \mathbb{P}(S = 1) \\ \vdots \\ \mathbb{P}(S = |\mathcal{D}|) \end{bmatrix} = \mathbf{T} \begin{bmatrix} \mathbb{P}(D = 1) \\ \vdots \\ \mathbb{P}(D = |\mathcal{D}|) \end{bmatrix},$$

denote as $\mathbf{s}_2 = \mathbf{T} \mathbf{d}_2$, where $\mathbf{s}_2 = \mathbb{P}(S)$ and $\mathbf{d}_2 = \mathbb{P}(D)$.

It follows the same argument that \mathbf{T} is row-stochastic and invertible and it is easy to verify that \mathbf{T}^{-1} takes the following form:

$$\begin{cases} \mathbf{T}_{i,i}^{-1} = \frac{\pi + |\mathcal{D}| - 2}{|\mathcal{D}| \pi - 1}, \text{ for } i \in D \\ \mathbf{T}_{i,j}^{-1} = \frac{\pi - 1}{|\mathcal{D}| \pi - 1}, \text{ for } i, j \in D \text{ s.t. } i \neq j \end{cases},$$

by multiplying \mathbf{T}^{-1} on both side, we obtain:

$$\begin{aligned} \mathbb{P}(D = k) &= \sum_{j=1}^{|\mathcal{D}|} \mathbf{T}_{kj}^{-1} \mathbb{P}(S = j) \\ &= \mathbf{T}_{k \cdot}^{-1} \mathbb{P}(S). \end{aligned}$$

Step 3: Recover the loss w.r.t. D

At the population level, we have recovered that:

$$\mathbb{P}(Y, \tilde{X} \mid D = k) = \mathbf{\Pi}_k^{-1} \mathbb{P}(Y, \tilde{X} \mid S),$$

where $\mathbb{P}(D = k) = \mathbf{T}_k^{-1} \mathbb{P}(S)$ is used in calculation of $\mathbf{\Pi}_k^{-1}$.

Hence, we recover the population equivalent of Eq. (1):

$$\begin{aligned} & \sum_{k=1}^{|\mathcal{D}|} \left(\mathbb{E}_{Y, \tilde{X} \mid D=k} \left[L(Y, f_k(\tilde{X})) \right] \cdot \mathbb{P}(D = k) \right) \\ &= \sum_{k=1}^{|\mathcal{D}|} \left(\int_Y \int_{\tilde{X}} \mathbb{P}(Y, \tilde{X} \mid D = k) L(Y, f_k(\tilde{X})) d\tilde{X} dY \cdot \mathbb{P}(D = k) \right) \\ &= \sum_{k=1}^{|\mathcal{D}|} \left(\left[\int_Y \int_{\tilde{X}} \sum_{j=1}^{|\mathcal{D}|} \mathbf{\Pi}_{kj}^{-1} \mathbb{P}(Y, \tilde{X} \mid S = j) L(Y, f_k(\tilde{X})) d\tilde{X} dY \right] \cdot \sum_{l=1}^{|\mathcal{D}|} \mathbf{T}_{kl}^{-1} \mathbb{P}(S = l) \right) \\ &= \sum_{k=1}^{|\mathcal{D}|} \left(\left[\sum_{j=1}^{|\mathcal{D}|} \int_Y \int_{\tilde{X}} \mathbf{\Pi}_{kj}^{-1} \mathbb{P}(Y, \tilde{X} \mid S = j) L(Y, f_k(\tilde{X})) d\tilde{X} dY \right] \cdot \sum_{l=1}^{|\mathcal{D}|} \mathbf{T}_{kl}^{-1} \mathbb{P}(S = l) \right) \\ &= \sum_{k=1}^{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} \left(\mathbf{\Pi}_{kj}^{-1} \mathbb{E}_{Y, \tilde{X} \mid S=j} \left[L(f_k(\tilde{X}), Y) \right] \cdot \sum_{l=1}^{|\mathcal{D}|} \mathbf{T}_{kl}^{-1} \mathbb{P}(S = l) \right). \end{aligned}$$

Therefore, we conclude that it is equivalent to minimizing:

$$(f_{1*}, \dots, f_{k*}) \leftarrow \arg \min_{f_1, \dots, f_k} \sum_{k=1}^{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} \left(\mathbf{\Pi}_{kj}^{-1} \mathbb{E}_{Y, \tilde{X} \mid S=j} \left[L(f_k(\tilde{X}), Y) \right] \cdot \sum_{l=1}^{|\mathcal{D}|} \mathbf{T}_{kl}^{-1} \mathbb{P}(S = l) \right)$$

This completes the proof. \square

C.2 THEOREM 4.3 PROOF

Theorem 4.3 For any $\delta \in (0, \frac{1}{2})$, $C_1 = \frac{\pi + |\mathcal{D}| - 2}{|\mathcal{D}| \pi - 1}$, denote $VC(\mathcal{F})$ as the VC-dimension of the hypothesis class \mathcal{F} , and K be some constant that depends on $VC(\mathcal{F})$. Let $f = \{f_k\}_{k=1}^{|\mathcal{D}|}$ where $f_k \in \mathcal{F}$ and let $L : Y \times Y \rightarrow \mathbb{R}_+$ be a loss function bounded by some constant M . Denote $k^* \leftarrow \arg \max_k |\hat{\mathcal{R}}^{LDP}(f_k) \hat{\mathbb{P}}(D = k) - \mathcal{R}^{LDP}(f_k) \mathbb{P}(D = k)|$, if $n \geq \frac{8 \ln(\frac{|\mathcal{D}|}{\delta})}{\min_k \mathbb{P}(S=k)}$ then with probability $1 - 2\delta$:

$$\hat{\mathcal{R}}^{LDP}(f) \leq \mathcal{R}(f^*) + K \sqrt{\frac{VC(\mathcal{F}) + \ln(\frac{\delta}{2})}{2n} \frac{2C_1 M |\mathcal{D}|}{\mathbb{P}(S = k^*)}}.$$

Proof. **Step 1:** simplify the objective

Denote $\mathcal{R}(f)$ as the expected risk of $(f_1, \dots, f_{|\mathcal{D}|})$ defined in Lemma 4.2, $\mathcal{R}(f_k)$ as the expected risk of f_k , and $\hat{\mathcal{R}}(f_k)$ as the empirical risk of f_k that depends on the data set given, then we start with

$$\begin{aligned} & \mathbb{P} \left(\left| \hat{\mathcal{R}}^{LDP}(f) \hat{\mathbb{P}}(D = k) - \mathcal{R}(f) \mathbb{P}(D = k) \right| > \epsilon \right) \\ &= \mathbb{P} \left(\left| \hat{\mathcal{R}}^{LDP}(f) \hat{\mathbb{P}}(D = k) + \mathcal{R}^{LDP}(f) \mathbb{P}(D = k) - \mathcal{R}^{LDP}(f) \mathbb{P}(D = k) - \mathcal{R}(f) \mathbb{P}(D = k) \right| > \epsilon \right) \\ &\leq \mathbb{P} \left(\left| \hat{\mathcal{R}}^{LDP}(f) \hat{\mathbb{P}}(D = k) - \mathcal{R}^{LDP}(f) \mathbb{P}(D = k) \right| + \left| \mathcal{R}^{LDP}(f) \mathbb{P}(D = k) - \mathcal{R}(f) \mathbb{P}(D = k) \right| > \epsilon \right) \\ &\stackrel{(a)}{=} \mathbb{P} \left(\left| \hat{\mathcal{R}}^{LDP}(f) \hat{\mathbb{P}}(D = k) - \mathcal{R}^{LDP}(f) \mathbb{P}(D = k) \right| \geq \epsilon \right) \\ &= \mathbb{P} \left(\left| \sum_{k=1}^{|\mathcal{D}|} \hat{\mathcal{R}}^{LDP}(f_k) \hat{\mathbb{P}}(D = k) - \sum_{k=1}^{|\mathcal{D}|} \mathcal{R}^{LDP}(f_k) \mathbb{P}(D = k) \right| > \epsilon \right) \\ &\leq \mathbb{P} \left(\sum_{k=1}^{|\mathcal{D}|} \left| \hat{\mathcal{R}}^{LDP}(f_k) \hat{\mathbb{P}}(D = k) - \mathcal{R}^{LDP}(f_k) \mathbb{P}(D = k) \right| > \epsilon \right) \\ &\stackrel{(b)}{\leq} \mathbb{P} \left(\max_k \left| \hat{\mathcal{R}}^{LDP}(f_k) \hat{\mathbb{P}}(D = k) - \mathcal{R}^{LDP}(f_k) \mathbb{P}(D = k) \right| > \frac{\epsilon}{|\mathcal{D}|} \right) \\ &\stackrel{(c)}{=} \mathbb{P} \left(\left| \sum_{j=1}^{|\mathcal{D}|} \hat{\Pi}_{k^*j}^{-1} \frac{1}{n_j} \sum_{i:S_i=j} L(Y_i, f_{k^*}(\tilde{X}_i)) \sum_{l=1}^{|\mathcal{D}|} \hat{T}_{k^*l}^{-1} \frac{n_l}{n} - \Pi_{k^*}^{-1} \mathbb{E}_{Y, \tilde{X}|S} [L(Y, f_{k^*}(\tilde{X}))] T_{k^*}^{-1} \mathbb{P}(S) \right| > \frac{\epsilon}{|\mathcal{D}|} \right), \end{aligned}$$

where $k^* \leftarrow \arg \max_k \left| \hat{\mathcal{R}}^{LDP}(f_k) \hat{\mathbb{P}}(D = k) - \mathcal{R}^{LDP}(f_k) \mathbb{P}(D = k) \right|$ and $\hat{\mathbb{P}}(S = k) = \frac{n_k}{n}$.

T^{-1} and Π^{-1} are as introduced in Lemma 4.2

(a) is obtained from the population equivalence of two risks from Lemma 4.2

(b) is followed by for two events A, B , if A implies B then $P(A) < P(B)$.

(c) is obtained by expanding $\hat{\mathcal{R}}^{LDP}(f_{k^*}) \hat{\mathbb{P}}(D = k^*)$ and $\mathcal{R}^{LDP}(f_{k^*}) \mathbb{P}(D = k^*)$ respectively.

Step 2: concentration of the empirical risk under Risk-LDP

Denote $n_{y\tilde{x}s}^N = \sum_i \mathbf{1}(y_i = y, \tilde{x}_i = \tilde{x}, s_i = s)$, $\mathbf{Q}_{y\tilde{x}s} = \mathbb{P}(Y = y, \tilde{X} = \tilde{x}, S = s)$, and define the random variable $N_{y\tilde{x}s} = \{i \mid y_i = y, \tilde{x}_i = \tilde{x}, s_i = s\}$. We can deduce $n_s^N = \sum_{\tilde{x} \in \tilde{X}, y \in Y} \mathbf{1}(y_i = y, \tilde{x}_i = \tilde{x}, s_i = s)$. Then, we have $\mathbb{E}[\hat{\mathcal{R}}^{LDP}(f_{k^*}) \hat{\mathbb{P}}(D = k^*) \mid N_{Y\tilde{X}S}] = \mathcal{R}^{LDP}(f_{k^*}) \mathbb{P}(D = k^*)$, where $N_{Y\tilde{X}S}$ denotes all possible $N_{y\tilde{x}s}$. Using similar approach of Lemma 2 in Mozannar et al.

(2020), we can write:

$$\begin{aligned}
& \mathbb{P} \left(\hat{\mathcal{R}}^{LDP}(f_{k^*})^N \hat{\mathbb{P}}(D = k^*)^N - \mathcal{R}^{LDP}(f_{k^*}) \mathbb{P}(D = k^*) > \frac{\epsilon}{|\mathcal{D}|} \right) \\
& \stackrel{(a)}{=} \sum_{N_Y \tilde{X} S} \mathbb{P} \left(\hat{\mathcal{R}}^{LDP}(f_{k^*})^N \hat{\mathbb{P}}(D = k^*)^N - \mathcal{R}^{LDP}(f_{k^*}) \mathbb{P}(D = k^*) > \frac{\epsilon}{|\mathcal{D}|} \middle| N_Y \tilde{X} S \right) \cdot \mathbb{P}(N_Y \tilde{X} S) \\
& \stackrel{(b)}{\leq} \mathbb{P} \left(\bigcup_{\tilde{x} \in \tilde{X}, y \in Y, s \in S} \left\{ n_s^N < \frac{n \sum_{\tilde{x} \in \tilde{X}, y \in Y} \mathcal{Q}_{y\tilde{x}s}}{2} \right\} \right) \\
& \quad + \sum_{\forall \tilde{x}, y, N_{y\tilde{x}s}: n_s^N \geq \frac{n \sum_{\tilde{x} \in \tilde{X}, y \in Y} \mathcal{Q}_{y\tilde{x}s}}{2}} \mathbb{P} \left(\hat{\mathcal{R}}^{LDP}(f_{k^*})^N \hat{\mathbb{P}}(D = k^*)^N - \mathcal{R}^{LDP}(f_{k^*}) \mathbb{P}(D = k^*) > \frac{\epsilon}{|\mathcal{D}|} \middle| N_Y \tilde{X} S \right) \cdot \mathbb{P}(N_Y \tilde{X} S) \\
& \stackrel{(c)}{\leq} |\mathcal{D}| \exp \left\{ - \frac{\min_s n \sum_{\tilde{x} \in \tilde{X}, y \in Y} \mathcal{Q}_{y\tilde{x}s}}{8} \right\} \\
& \quad + \sum_{\forall \tilde{x}, y, N_{y\tilde{x}s}: n_s^N \geq \frac{n \sum_{\tilde{x} \in \tilde{X}, y \in Y} \mathcal{Q}_{y\tilde{x}s}}{2}} \mathbb{P} \left(\hat{\mathcal{R}}^{LDP}(f_{k^*})^N \hat{\mathbb{P}}(D = k^*)^N - \mathcal{R}^{LDP}(f_{k^*}) \mathbb{P}(D = k^*) > \frac{\epsilon}{|\mathcal{D}|} \middle| N_Y \tilde{X} S \right) \cdot \mathbb{P}(N_Y \tilde{X} S),
\end{aligned}$$

where (a) follows by conditioning over all $2^n |\tilde{X}|^n |\mathcal{D}|^n$ possible configurations of $N_{y\tilde{x}s} \subset [n]$.

(b) is obtained by splitting the configurations where $\forall \tilde{x}, y, N_{y\tilde{x}s} : n_s^N \geq \frac{n \sum_{\tilde{x} \in \tilde{X}, y \in Y} \mathcal{Q}_{y\tilde{x}s}}{2}$ and the complement of the event and upper bound the complement of the event by the probability that $\exists s$ s.t. $n_s^N < \frac{n \sum_{\tilde{x} \in \tilde{X}, y \in Y} \mathcal{Q}_{y\tilde{x}s}}{2}$. (c) is obtained by the union bound and we know $n_s^N \sim \text{Binomial} \left(n, \sum_{\tilde{x} \in \tilde{X}, y \in Y} \mathcal{Q}_{y\tilde{x}s} \right)$ and apply the Chernoff bound on $n_{y\tilde{x}s}^N$.

Notice that $f_k(\cdot)$ only takes $\tilde{X} \in T(\mathcal{X})$ as input, therefore, we will be able to apply the McDiarmid Inequality (McDiarmid (1989)). Let $X^n = (X_1, \dots, X_n) \in \mathcal{X}^n$ be n independent random variables and let $g : \mathcal{X}^n \rightarrow \mathbb{R}$, if there exists constants c_1, \dots, c_n s.t.

$$\sup_{x_1, \dots, x_i, x'_i, \dots, x_n} |g(x_1, \dots, x_i, \dots, x_n) - g(x_1, \dots, x'_i, \dots, x_n)| \leq c_i, i = 1, \dots, n,$$

then $\forall \epsilon > 0$:

$$\mathbb{P}(|g(x_1, \dots, x_i, \dots, x_n) - \mathbb{E}[g(x_1, \dots, x'_i, \dots, x_n)]| > \epsilon) \leq 2 \exp \left(- \frac{2\epsilon^2}{\sum_{i=1}^n c_i^2} \right).$$

Since by conditioning on $N_Y \tilde{X} S$, then for $\hat{\mathcal{R}}^{LDP}(f_{k^*})$, everything else is now deterministic except for f_{k^*} , in other words, by conditioning on $N_Y \tilde{X} S$, the value of $\hat{\mathcal{R}}^{LDP}(f_{k^*})$ only depends on f_{k^*} . Then, for two datasets N, N' where they only differ by one value of $f_{k^*}(\tilde{X}_i)$, we try to bound how much f_{k^*} can change.

Recall from Lemma 4.2 we computed the entries of Π^{-1} takes the following form:

$$\begin{cases} \Pi_{i,i}^{-1} = \frac{\pi + |\mathcal{D}| - 2}{|\mathcal{D}| \pi - 1} \frac{\pi p_i + \sum_{d'' \in \mathcal{D} \setminus i} \pi p_{d''}}{p_i}, \text{ for } i \in \mathcal{D} \\ \Pi_{i,j}^{-1} = \frac{\pi - 1}{|\mathcal{D}| \pi - 1} \frac{\pi p_i + \sum_{d'' \in \mathcal{D} \setminus i} \pi p_{d''}}{p_i}, \text{ for } i, j \in \mathcal{D} \text{ s.t. } i \neq j \end{cases}$$

For simplicity, let $C_1 = \frac{\pi + |\mathcal{D}| - 2}{|\mathcal{D}| \pi - 1}$, $C_2 = \frac{\pi - 1}{|\mathcal{D}| \pi - 1}$, since we do not have access to D , therefore we can not directly observe p_d , hence we write Π^{-1} in terms of P_s where $P_s = \mathbb{P}(S)$:

$$\begin{cases} \Pi_{i,i}^{-1} = C_1 \frac{\pi T_{i,i}^{-1} P_s + \sum_{l \in \mathcal{D} \setminus i} \pi T_{l,i}^{-1} P_s}{T_{i,i}^{-1} P_s}, \text{ for } i \in \mathcal{D} \\ \Pi_{i,j}^{-1} = C_2 \frac{\pi T_{i,i}^{-1} P_s + \sum_{l \in \mathcal{D} \setminus i} \pi T_{l,i}^{-1} P_s}{T_{i,i}^{-1} P_s}, \text{ for } i, j \in \mathcal{D} \text{ s.t. } i \neq j \end{cases},$$

we also computed T^{-1} as:

$$\begin{cases} T_{i,i}^{-1} = \frac{\pi + |\mathcal{D}| - 2}{|\mathcal{D}| \pi - 1}, \text{ for } i \in \mathcal{D} \\ T_{i,j}^{-1} = \frac{\pi - 1}{|\mathcal{D}| \pi - 1}, \text{ for } i, j \in \mathcal{D} \text{ s.t. } i \neq j \end{cases},$$

then we have

$$\begin{aligned}
& \sup_{N, N'} \left| \hat{\mathcal{R}}^{LDP}(f_{k^*})^N \hat{\mathbb{P}}(D = k^*)^N - \hat{\mathcal{R}}^{LDP}(f_{k^*})^{N'} \hat{\mathbb{P}}(D = k^*)^{N'} \right| \\
& \stackrel{(a)}{\leq} \sup_{N, N'} \left| \hat{\mathcal{R}}^{LDP}(f_{k^*})^N - \hat{\mathcal{R}}^{LDP}(f_{k^*})^{N'} \right| \\
& = \left| C_1 \frac{\pi \mathbf{T}_{k^*}^{-1} \mathbf{P}_s^N + \sum_{l \setminus k} \bar{\pi} \mathbf{T}_l^{-1} \mathbf{P}_s^N}{\mathbf{T}_{k^*}^{-1} \mathbf{P}_s^N} \hat{\mathcal{R}}^{LDP}(f_{k^*})^N + \sum_{j \setminus k} C_2 \frac{\bar{\pi} \mathbf{T}_{k^*}^{-1} \mathbf{P}_s^N + \sum_{l \setminus k} \pi \mathbf{T}_l^{-1} \mathbf{P}_s^N}{\mathbf{T}_{k^*}^{-1} \mathbf{P}_s^N} \hat{\mathcal{R}}^{LDP}(f_{k^*})^N \right. \\
& \quad \left. - C_1 \frac{\pi \mathbf{T}_{k^*}^{-1} \mathbf{P}_s^N + \sum_{l \setminus k} \bar{\pi} \mathbf{T}_l^{-1} \mathbf{P}_s^N}{\mathbf{T}_{k^*}^{-1} \mathbf{P}_s^N} \hat{\mathcal{R}}^{LDP}(f_{k^*})^{N'} + \sum_{j \setminus k} C_2 \frac{\bar{\pi} \mathbf{T}_{k^*}^{-1} \mathbf{P}_s^N + \sum_{l \setminus k} \pi \mathbf{T}_l^{-1} \mathbf{P}_s^N}{\mathbf{T}_{k^*}^{-1} \mathbf{P}_s^N} \hat{\mathcal{R}}^{LDP}(f_{k^*})^{N'} \right| \\
& = \left| C_1 \frac{\pi \mathbf{T}_{k^*}^{-1} \mathbf{P}_s^N + \sum_{l \setminus k} \bar{\pi} \mathbf{T}_l^{-1} \mathbf{P}_s^N}{\mathbf{T}_{k^*}^{-1} \mathbf{P}_s^N} \left(\frac{\sum_{i \in N, \tilde{x} \in \tilde{X}, y \in Y, S=k} L(y_i, f_{k^*}(\tilde{x}_i))}{n_{k^*}} - \frac{\sum_{i \in N', \tilde{x} \in \tilde{X}, y \in Y, S=k} L(y_i, f_{k^*}(\tilde{x}_i))}{n_{k^*}} \right) \right. \\
& \quad \left. + \sum_{j \setminus k} C_2 \frac{\bar{\pi} \mathbf{T}_{k^*}^{-1} \mathbf{P}_s^N + \sum_{l \setminus k} \pi \mathbf{T}_l^{-1} \mathbf{P}_s^N}{\mathbf{T}_{k^*}^{-1} \mathbf{P}_s^N} \left(\frac{\sum_{i \in N, \tilde{x} \in \tilde{X}, y \in Y, S=j} L(y_i, f_{k^*}(\tilde{x}_i))}{n_{k^*}} - \frac{\sum_{i \in N', \tilde{x} \in \tilde{X}, y \in Y, S=j} L(y_i, f_{k^*}(\tilde{x}_i))}{n_{k^*}} \right) \right| \\
& \stackrel{(b)}{\leq} \left| C_1 \frac{\pi \max_{m \in [|\mathcal{D}|]} \mathbf{T}_m^{-1} \mathbf{P}_s^N + (|\mathcal{D}| - 1) \bar{\pi} \max_{m \in [|\mathcal{D}|]} \mathbf{T}_m^{-1} \mathbf{P}_s^N}{\mathbf{T}_{k^*}^{-1} \mathbf{P}_s^N} \cdot \frac{M}{n_{k^*}} \right| \\
& = \left| C_1 (\pi + \bar{\pi}(|\mathcal{D}| - 1)) \cdot \frac{M}{n_{k^*}} \right| \\
& \stackrel{(c)}{=} \left| \frac{C_1 M}{n_{k^*}} \right|,
\end{aligned}$$

where (a) is followed by the fact that we only consider the change in \tilde{X} , and $\hat{\mathbb{P}}(D = k^*) \leq 1$. (b) is obtained by $C_2 \leq 0, \forall \pi \in (\frac{1}{|\mathcal{D}|}, 1]$. (c) is followed by the fact that $\pi + \bar{\pi}(|\mathcal{D}| - 1) = 1$.

Now, we are ready to apply the McDiarmid Inequality:

$$\begin{aligned}
& \sum_{\forall \tilde{x}, y, N_{y\tilde{x}s}: n_s^N \geq \frac{n \sum_{\tilde{x} \in \tilde{X}, y \in Y} \mathbf{Q}_{y\tilde{x}s}}{2}} \mathbb{P} \left(\hat{\mathcal{R}}^{LDP}(f_{k^*})^N - \mathcal{R}^{LDP}(f_{k^*}) > \frac{\epsilon}{|\mathcal{D}|} \middle| N_{Y\tilde{X}S} \right) \cdot \mathbb{P}(N_{Y\tilde{X}S}) \\
& \leq \sum_{\forall x, y, N_{y\tilde{x}s}: n_s^N \geq \frac{n \sum_{\tilde{x} \in \tilde{X}, y \in Y} \mathbf{Q}_{y\tilde{x}s}}{2}} 2 \exp \left\{ - \frac{\frac{2\epsilon^2}{|\mathcal{D}|^2}}{n \cdot \left(\frac{C_1 M}{n_{k^*}} \right)^2} \right\} \cdot \mathbb{P}(N_{Y\tilde{X}S}) \\
& \stackrel{(a)}{\leq} 2 \exp \left\{ - 2n\epsilon^2 \left(\frac{\mathbb{P}(S = k^*)}{2C_1 M |\mathcal{D}|} \right)^2 \right\},
\end{aligned}$$

where (a) is obtained since when $n_{k^*} = \frac{n \sum_{\tilde{x} \in \tilde{X}, y \in Y} \mathbf{Q}_{y\tilde{x}k^*}}{2} = \frac{n \mathbb{P}(S=k^*)}{2}$, the quantity is maximized.

Now, we have:

$$\mathbb{P} \left(\left| \hat{\mathcal{R}}^{LDP}(f) - \mathcal{R}(f) \right| > \epsilon \right) \leq |\mathcal{D}| \exp \left\{ - \frac{\min_k \mathbb{P}(S = k)}{8} \right\} + 2 \exp \left\{ - 2n\epsilon^2 \left(\frac{\mathbb{P}(S = k^*)}{2C_1 M |\mathcal{D}|} \right)^2 \right\},$$

solve for δ , we now have, for any $\delta \in (0, \frac{1}{2})$, $\epsilon \geq \sqrt{\frac{\ln(\frac{2}{\delta})}{2n} \frac{2C_1 M |\mathcal{D}|}{\mathbb{P}(S=k^*)}}$, if $n \geq \frac{8 \ln(\frac{|\mathcal{D}|}{\delta})}{\min_k \mathbb{P}(S=k)}$, then

$$\mathbb{P} \left(\left| \hat{\mathcal{R}}^{LDP}(f) - \mathcal{R}(f) \right| > \epsilon \right) \leq 2\delta$$

Step 3: Obtain the final result

Recall that one can easily show

$$\hat{\mathcal{R}}^{LDP}(f) - \mathcal{R}(f^*) \leq 2 \sup_{f \in \mathcal{F}} \left| \mathcal{R}^{LDP}(f) - \hat{\mathcal{R}}^{LDP}(f) \right|,$$

but we have already established similar results for one single hypothesis in **Step 2**. Therefore, what remains is to extend the previous result that bounds the generalization error between any single hypothesis and the optimal hypothesis in the entire hypothesis class. And this can be done easily by introducing the VC-dimension of the hypothesis \mathcal{F} . Denote the VC-dimension of our hypothesis class \mathcal{F} as $VC(\mathcal{F})$, then with some constant K and for any $\delta \in (0, \frac{1}{2})$, if $n \geq \frac{8 \ln(\frac{|\mathcal{D}|}{\delta})}{\min_k \mathbb{P}(S=k)}$, we have:

$$\hat{\mathcal{R}}^{LDP}(f) \leq \mathcal{R}(f^*) + K \sqrt{\frac{VC(\mathcal{F}) + \ln(\frac{\delta}{2})}{2n} \frac{2C_1 M |\mathcal{D}|}{\mathbb{P}(S = k^*)}}.$$

This completes the proof. □

C.3 LEMMA 4.4 PROOF

Lemma 4.4 Consider ϵ -LDP setting with $\pi \in (\frac{1}{|\mathcal{D}|}, 1]$ and $\bar{\pi} \in [0, \frac{1}{|\mathcal{D}|})$. For some transformation of X , denoted by $X^* = \tilde{T}(X)$, assume there exists at least one anchor point X_{anchor}^* in the dataset s.t. $\mathbb{P}(D = j^* | X_{\text{anchor}}^*) = 1$ for some $j^* \in [|\mathcal{D}|]$. Then $\pi = \mathbb{P}(S = j^* | X_{\text{anchor}}^*)$. Empirically, for a dataset with n observation, let $\eta_{j^*}^n(X^*) = (\hat{\mathbb{P}}(S = j^* | X_1^*), \dots, \hat{\mathbb{P}}(S = j^* | X_n^*))$, then $\hat{\pi} = \|\eta_{j^*}^n(X^*)\|_\infty$.

Proof. Notice that $\pi \in (\frac{1}{|\mathcal{D}|}, 1]$, $\bar{\pi} \in [0, \frac{1}{|\mathcal{D}|})$ and consequently we have $\pi > \bar{\pi}$. Hence, by Theorem 5 of Zhang et al. (2021), we are in a good position to apply the noise rate estimation method (Theorem 3) in Patrini et al. (2017) to estimate $\pi, \bar{\pi}$. Our ϵ -LDP setting can be considered as a special case of CCN (class conditional noise) where the flip probability is the same across all groups in \mathcal{D} . Consider

$$\begin{aligned} \mathbb{P}(S = j^* | X_{\text{anchor}}^*) &= \sum_{k=1}^{|\mathcal{D}|} \mathbb{P}(S = j^* | D = k) \cdot \mathbb{P}(D = k | X_{\text{anchor}}^*) \\ &\stackrel{(a)}{=} \sum_{k=1}^{|\mathcal{D}|} \mathbb{P}(S = j^* | D = k) \cdot \mathbf{1}\{j^* = k\} \\ &= \pi, \end{aligned}$$

(a) is followed by the definition of anchor point

$$\mathbb{P}(D = j^* | X_{\text{anchor}}^*) = 1 \implies \mathbb{P}(D = k | X_{\text{anchor}}^*) = 0, \forall k \neq j^*, k, j^* \in [|\mathcal{D}|].$$

Then one can easily see that $\mathbb{P}(S = j^* | X_i^*)$ attains its maximum when $\mathbb{P}(D = j^* | X_i^*) = 1$, since we know

$$\begin{cases} \mathbb{P}(S = j^* | D = k) = \pi, & \text{if } j^* = k \\ \mathbb{P}(S = j^* | D = k) = \bar{\pi}, & \text{if } j^* \neq k, \end{cases}$$

hence we know $\mathbb{P}(S = j^* | X_i^*)$ is actually a weighted sum of π and $\bar{\pi}$, where the weights are simply $\{\mathbb{P}(D = k | X_i^*)\}_{k=1}^{|\mathcal{D}|}$. But we also know that $\pi > \bar{\pi}$. Hence, for empirical estimation, $\eta_{j^*}^n = (\hat{\mathbb{P}}(S = j^* | X_1^*), \dots, \hat{\mathbb{P}}(S = j^* | X_n^*))$, then $\hat{\pi} = \|\eta_{j^*}^n\|_\infty$.

This completes the proof \square

C.4 THEOREM 4.5 PROOF

Theorem 4.5 For any $\delta \in (0, \frac{1}{3})$, $C_1 = \frac{\pi + |\mathcal{D}| - 2}{|\mathcal{D}| \pi - 1}$, denote $VC(\mathcal{F})$ as the VC-dimension of the hypothesis class \mathcal{F} , and K be some constant that depends on $VC(\mathcal{F})$. If Assumption A, B, and Lemma 4.4 hold, let $f = \{f_k\}_{k=1}^{|\mathcal{D}|}$ where $f_k \in \mathcal{F}$ and let $L : Y \times Y \rightarrow \mathbb{R}_+$ be a loss function bounded by some constant M . Denote $k^* \leftarrow \arg \max_k |\hat{\mathcal{R}}^{LDP}(f_k) \mathbb{P}(D = k^*) - \mathcal{R}^{LDP}(f_k) \mathbb{P}(D = k^*)|$, if $n \geq \frac{8 \ln(\frac{|\mathcal{D}|}{\delta})}{\min_k \mathbb{P}(S=k)}$, $n_1 \geq \frac{1}{c(\tilde{\epsilon} - \theta)^2} (M_g + \frac{C_1 + \theta}{\ln 2})^2 \ln(\frac{2}{\delta})$, and $M_g + \frac{C_1 + \theta}{\ln 2} > \tilde{\epsilon} > \theta$, where c is an absolute constant, then with probability $1 - 3\delta$:

$$\hat{\mathcal{R}}^{LDP}(f) \leq \mathcal{R}(f^*) + K \sqrt{\frac{VC(\mathcal{F}) + \ln(\frac{\delta}{2})}{2n} \frac{2(C_1 + \tilde{\epsilon})M|\mathcal{D}|}{\mathbb{P}(S = k^*)}}.$$

Proof. We will first introduce some preliminaries that will be used in the proof. We will first introduce how we obtain \hat{C}_1 and then state the assumptions used for the proof.

Step 1: Grouping: we evenly divide $\{X_i^*, S_i\}_{i=1}^n$ into n_1 groups, with $m = \frac{n}{n_1}$ samples each.

Step 2: Estimating within groups: for any $k \in [n_1]$, within each group $\{X_{k,j}^*, S_{k,j}\}_{j=1}^m$, we can derive an m -dimension vector $\boldsymbol{\eta}_{j^*,k}^m(X_{k,\cdot}^*) = (\hat{\mathbb{P}}_k(S = j^* | X_{k,1}^*), \dots, \hat{\mathbb{P}}_k(S = j^* | X_{k,m}^*))$ and $\hat{\pi}_k = \|\boldsymbol{\eta}_{j^*,k}^m(X_{k,\cdot}^*)\|_\infty$, as defined in Lemma 4.4. Then, by a simple plug in to get $\hat{C}_{1,k} = \frac{\hat{\pi}_k + |\mathcal{D}| - 2}{|\mathcal{D}| \hat{\pi}_k - 1}$.

Step 3: Averaging: \hat{C}_1 is our estimator for C_1 , computed as $\hat{C}_1 = \frac{1}{n_1} \sum_{k=1}^{n_1} \hat{C}_{1,k}$, $\hat{C}_{1,k}$, $k \in [n_1]$.

Next, we state two assumptions used to derive Theorem 4.5 (noise rate is estimated from the data).

Assumption A: (Sub-exponentiality) For all $k \in [n_1]$, define $\hat{g}_k(X^*) = \hat{\mathbb{P}}_k(S = j^* | X^*)$. There exists a constant $M_g > 0$, such that $\|\hat{C}_{1,k}\|_{\psi_1} = \|\min_{i \in [m]} \frac{\hat{g}_k(X_{k,i}^*) + |\mathcal{D}| - 2}{|\mathcal{D}| \hat{g}_k(X_{k,i}^*) - 1}\|_{\psi_1} \leq M_g$ for all $k \in [n_1]$, where $\|\cdot\|_{\psi_1}$ is the sub-exponential norm: $\|X\|_{\psi_1} = \inf\{t > 0 | \mathbb{E}[e^{X/t}] \leq 2\}$.

Assumption B: (Nearly Unbiasedness) For all $k \in [n_1]$, $\hat{C}_{1,k}$ is a “nearly” unbiased estimator of C_1 , namely $|\mathbb{E}[\hat{C}_{1,k}] - C_1| < \theta$ for all $k \in [n_1]$, where $\theta > 0$.

Now, we begin the proof.

First, we will prove a concentration inequality with regard to \hat{C}_1 and C_1 .

Since for any constant L , we have

$$\begin{aligned} \|L\|_{\psi_1} &= \inf\{t > 0 | \mathbb{E}[e^{|L|/t}] \leq 2\} \\ &= \inf\{t > 0 | \mathbb{E}[e^{|L|/t}] \leq 2\} \\ &= \frac{|L|}{\ln 2}, \end{aligned}$$

and $\|\cdot\|_{\psi_1}$ is a norm, we can conclude that the standardized statistic $\tilde{C}_{1,k} = \hat{C}_{1,k} - \mathbb{E}[\hat{C}_{1,k}]$ is also sub-exponential:

$$\begin{aligned} \|\tilde{C}_{1,k}\|_{\psi_1} &\leq \|\hat{C}_{1,k}\|_{\psi_1} + \|\mathbb{E}[\hat{C}_{1,k}]\|_{\psi_1} \\ &\leq M_g + \frac{|\mathbb{E}[\hat{C}_{1,k}]|}{\ln 2} \\ &\stackrel{(a)}{=} M_g + \frac{C_1 + \theta}{\ln 2}, \end{aligned}$$

where (a) is obtained by Assumption B.

Among different groups, the data are mutually independent, then we know that $\{\tilde{C}_{1,k}\}_{k=1}^{n_1}$ are independent random variables with mean 0.

Therefore, we can apply Bernstein inequality (R. Vershynin (2018)):

$$\mathbb{P}\left(\left|\frac{1}{n_1} \sum_{k=1}^{n_1} \tilde{C}_{1,k}\right| > \tilde{\epsilon} + \theta\right) \leq 2 \exp\left[-c \min\left(\frac{(\tilde{\epsilon} + \theta)^2}{(M_g + C_1/\ln 2)^2}, \frac{\tilde{\epsilon} + \theta}{M_g + C_1/\ln 2}\right) n_1\right],$$

where $c > 0$ is an absolute constant.

Since we have $M_g + \frac{C_1 + \theta}{\ln 2} > \tilde{\epsilon} > \theta$, which implies $\frac{\tilde{\epsilon}}{M_g + C_1 / \ln 2} < 1$, we can transform the inequality above into

$$\begin{aligned} \mathbb{P}\left(\left|\hat{C}_1 - C_1\right| > \tilde{\epsilon}\right) &= \mathbb{P}\left(\left|\frac{1}{n_1} \sum_{k=1}^{n_1} \tilde{C}_{1,k}\right| > \tilde{\epsilon} - \theta\right) \\ &\leq 2 \exp\left[-c \frac{(\tilde{\epsilon} - \theta)^2}{(M_g + (C_1 + \theta) / \ln 2)^2} n_1\right] \\ &\stackrel{(a)}{\leq} \delta, \end{aligned}$$

where (a) is obtained by $n_1 \geq \frac{1}{c(\tilde{\epsilon} - \theta)^2} (M_g + \frac{C_1 + \theta}{\ln 2})^2 \ln(\frac{2}{\delta})$.

Second, we can apply Theorem 4.3 to the case when using $\hat{\pi}$ instead of π . Therefore, by the end of **Step 2** in the proof of Theorem 4.3 we will derive the following conclusion:

For any $\delta \in (0, \frac{1}{3})$, $\epsilon \geq \sqrt{\frac{\ln(\frac{\delta}{2})}{2n} \frac{2\hat{C}_1 M |\mathcal{D}|}{\mathbb{P}(S=k^*)}}$, if $n \geq \frac{8 \ln(\frac{|\mathcal{D}|}{\delta})}{\min_k \mathbb{P}(S=k)}$, then

$$\mathbb{P}(|\hat{\mathcal{R}}^{LDP}(f) - \mathcal{R}(f)| > \epsilon) \leq 2\delta.$$

Third, Assume the events

$$\begin{aligned} A_1 &= \left\{ \left| \hat{C}_1 - C_1 \right| \leq \tilde{\epsilon} \right\}, \\ A_2 &= \left\{ |\hat{\mathcal{R}}^{LDP}(f) - \mathcal{R}(f)| \leq \epsilon, \epsilon \geq \sqrt{\frac{\ln(\frac{\delta}{2})}{2n} \frac{2\hat{C}_1 M |\mathcal{D}|}{\mathbb{P}(S=k^*)}}, n \geq \frac{8 \ln(\frac{|\mathcal{D}|}{\delta})}{\min_k \mathbb{P}(S=k)} \right\}, \\ A_3 &= \left\{ |\hat{\mathcal{R}}^{LDP}(f) - \mathcal{R}(f)| \leq \epsilon, \epsilon \geq \sqrt{\frac{\ln(\frac{\delta}{2})}{2n} \frac{2(C_1 + \tilde{\epsilon}) M |\mathcal{D}|}{\mathbb{P}(S=k^*)}}, n \geq \frac{8 \ln(\frac{|\mathcal{D}|}{\delta})}{\min_k \mathbb{P}(S=k)} \right\}, \end{aligned}$$

then we have $A_1 \cap A_2 \subseteq A_3$.

From the **First** part and **Second** part of the proof, we have $\mathbb{P}(A_1^C) \leq \delta$, $\mathbb{P}(A_2^C) \leq 2\delta$, then

$$\mathbb{P}(A_3) \geq \mathbb{P}(A_1 \cap A_2) \geq 1 - \mathbb{P}(A_1^C) - \mathbb{P}(A_2^C) \geq 1 - 3\delta,$$

which is equivalent to the following statement: For any $\delta \in (0, \frac{1}{3})$, $\epsilon \geq \sqrt{\frac{\ln(\frac{\delta}{2})}{2n} \frac{2(C_1 + \tilde{\epsilon}) M |\mathcal{D}|}{\mathbb{P}(S=k^*)}}$, if $n \geq \frac{8 \ln(\frac{|\mathcal{D}|}{\delta})}{\min_k \mathbb{P}(S=k)}$, then

$$\mathbb{P}(|\hat{\mathcal{R}}^{LDP}(f) - \mathcal{R}(f)| > \epsilon) \leq 3\delta.$$

Finally, similar to **Step 3** in the proof of Theorem 4.3 recall that one can easily show

$$\hat{\mathcal{R}}^{LDP}(f) - \mathcal{R}(f^*) \leq 2 \sup_{f \in \mathcal{F}} |\mathcal{R}^{LDP}(f) - \hat{\mathcal{R}}^{LDP}(f)|,$$

but we have already established similar results for one single hypothesis in **Step 2**. Therefore, what remains is to extend the previous result that bounds the generalization error between any single hypothesis and the optimal hypothesis in the entire hypothesis class. And this can be done easily by introducing the VC-dimension of the hypothesis \mathcal{F} . Denote the VC-dimension of our hypothesis class \mathcal{F} as $VC(\mathcal{F})$, then with some constant K and for any $\delta \in (0, \frac{1}{3})$, if $n \geq \frac{8 \ln(\frac{|\mathcal{D}|}{\delta})}{\min_k \mathbb{P}(S=k)}$, then with probability $1 - 3\delta$ we have:

$$\hat{\mathcal{R}}^{LDP}(f) \leq \mathcal{R}(f^*) + K \sqrt{\frac{VC(\mathcal{F}) + \ln(\frac{\delta}{2})}{2n} \frac{2(C_1 + \tilde{\epsilon}) M |\mathcal{D}|}{\mathbb{P}(S=k^*)}}.$$

This completes the proof. \square

D DEFERRED EXPERIMENT RESULTS

D.1 AUTO INSURANCE

D.1.1 DATA

The Auto Insurance data set contains 8150 observations, 17 features, and 1 binary response. In our experiment, we choose $D = \text{sex}$ to be the sensitive attribute taking values "Male" and "Female". privatized sensitive attribute S is generated under different privacy levels using a set of ϵ 's by Definition 4.1 D was used to set the performance benchmark and is masked under any other settings.

D.1.2 EXPERIMENTS SETUP

We conduct experiments 1) when the noise rate $\pi, \bar{\pi}$ are known and 2) when the noise rates are unknown. To investigate how a transformation T may affect the performance of our method, we consider a transformation $T(X) = \tilde{X}$ obtained via supervised learning (as shown in Example 4.1). Other transformations, such as grouping, and discretization are very commonly seen in insurance pricing.

For both scenarios, we let the hypothesis class \mathcal{F} be the class of linear models for all settings as this is a more practical setup in an insurance pricing setting due to requirements on transparency. We ran our algorithm on three pre-defined π 's, namely 0.8, 0.7, and 0.6 to demonstrate how noise rate may affect the performance of our method for both scenarios. We also created various subsets of the original data to demonstrate how sample size may affect the performance of our method for both scenarios. Additionally, under scenario 2, we compared performance using three different n_1 values, namely 1, 2, and 4 to verify our results in Theorem 4.5 To obtain the discrimination-free price h^* , we choose $P^*(d)$ to be the empirical marginal distribution of D .

In all figures in the remaining section, the curves represent the empirical test loss of the best-estimate price defined in 3.1 Blue curve (Best-Estimate) is obtained using a naive logistic regression. While the orange curve (MPTP) and the rest (MPTP-LDP) are also obtained using a naive logistic regression, group-specific score functions (defined in Eq. 1) were used instead.

D.1.3 SCENARIO 1 (KNOWN NOISE RATE)

Since the main issue is to estimate $\mu(X, D)$ when the true sensitive attribute is not accessible, we will focus on presenting our results in the estimation of $\mu(X, D)$. However, results for $h^*(X)$ under scenario 1 can also be found in Appendix F We first investigate the effect of noise rate on loss approximation when the sample size is fixed:

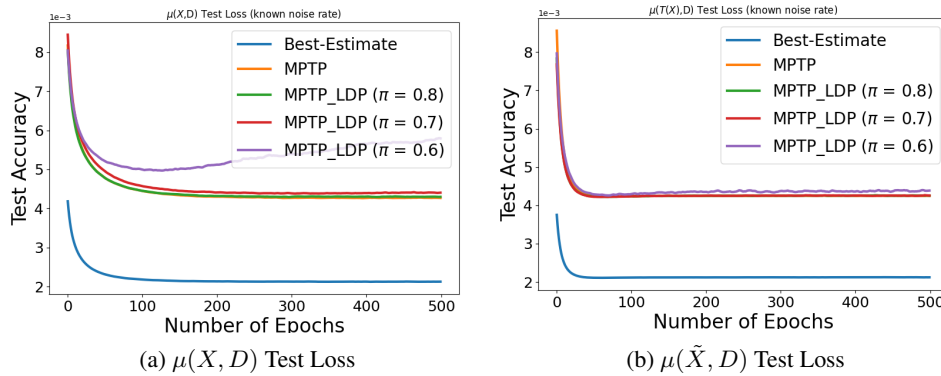


Figure 5: Test Loss for Scenario 1 (fixed sample size)

As expected, we observe very similar patterns compared to the healthcare insurance dataset.

Next, we study the effect of sample size on loss approximation for fixed noise rates ($\pi = 0.7$). To compare the performance on loss approximation on different sample sizes, we randomly create

a subset of the full data set (1338 observations) that contains only half of its observations (669 observations) and then run the same experiment over the full data set (green curve) and the subset (red curve) to obtain the following result:

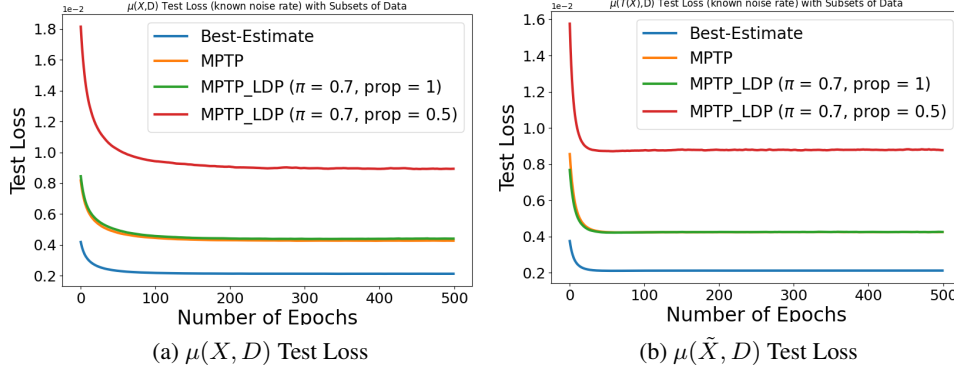


Figure 6: Test Loss for Scenario 1 (fixed noise rate: $\pi = 0.7$)

Again, we observe very similar patterns compared to the healthcare insurance dataset.

D.1.4 SCENARIO 2 (UNKNOWN NOISE RATE)

Similar to the procedure in scenario 1, with the key difference being that π is replaced by an estimate $\hat{\pi}$ obtained following Lemma 4.4 and Theorem 4.5. To demonstrate the consistency with our theoretical results in Theorem 4.5, on top of the comparison under fixed sample size or true noise rate, we also present the result with different n_1 values. As mentioned, results for $h^*(X)$ under scenario 2 can also be found in Appendix F. The same method is used to estimate π compared to the healthcare insurance dataset.

We first present the empirical results for the effect of noise rate on loss approximation when the sample size is fixed:

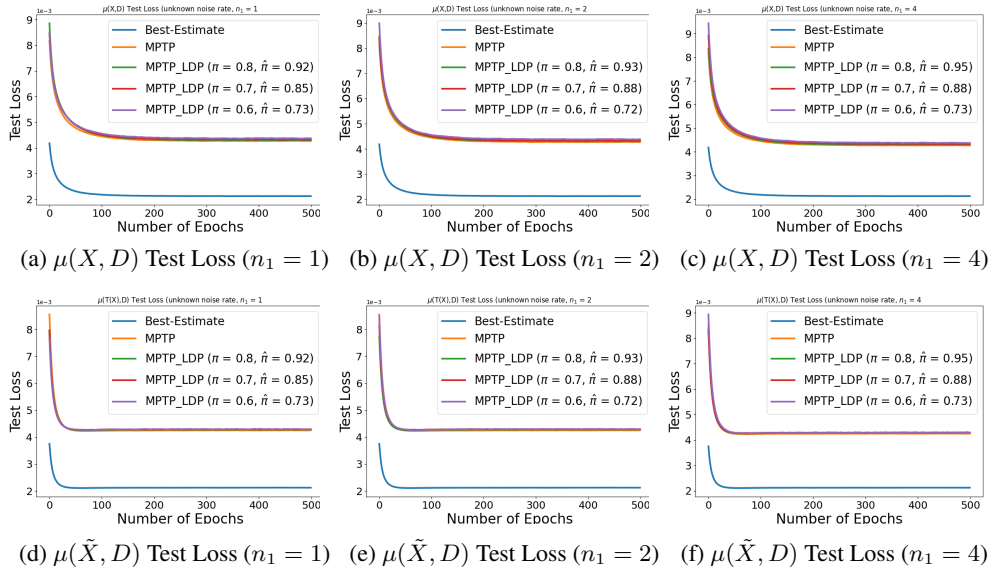
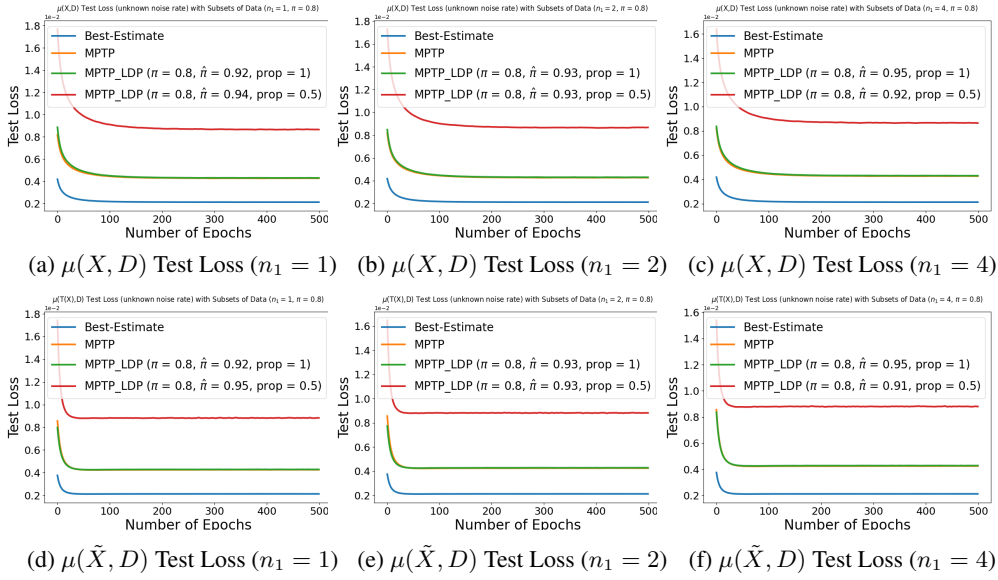


Figure 7: Test Loss for Scenario 2 (fixed sample size)

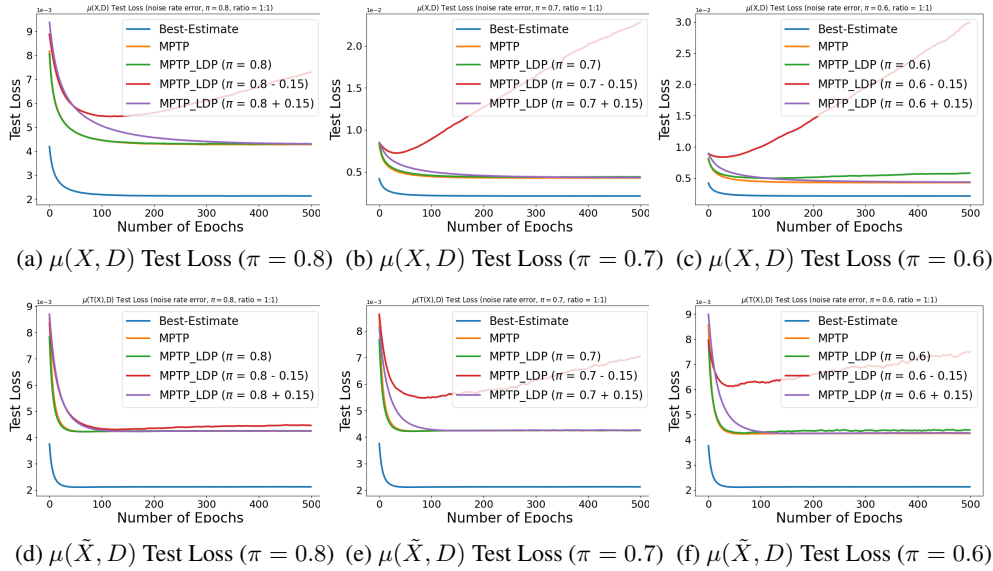
We will now show our findings for the effect of sample size on loss approximation for a fixed noise rate ($\pi = 0.8$) in the following:

Figure 8: Test Loss for Scenario 2 (fixed noise rate: $\pi = 0.8$)

We noticed that in this data set, although the noise rate is overestimated with all n_1 's, there is no issue in terms of convergence behavior. This provides further evidence that underestimation can be much more destructive to the algorithm than overestimation.

D.1.5 EMPIRICAL STUDY ON THE IMPACT OF NOISE RATE ESTIMATION ERROR

we present our findings on the effect of estimation error for π on the empirical performance of our algorithm. Particularly, we investigate the effect of underestimation and overestimation for π under both balanced and imbalanced distribution for privatized sensitive attributes S by introducing pre-defined estimation errors. We sampled subsets of the full data set to obtain data with imbalanced distribution for privatized sensitive attributes. For conciseness, we present our results for the balanced case below and defer results for the imbalanced case in Appendix E

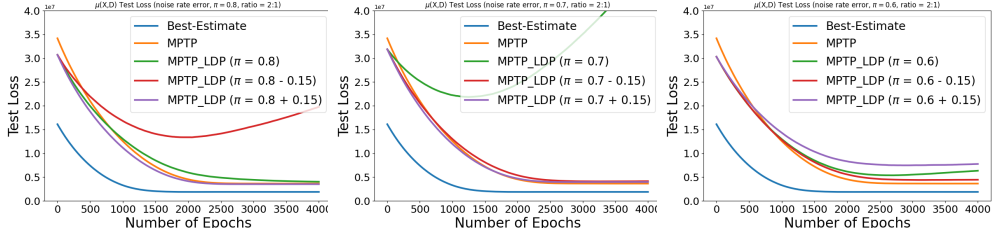
Figure 9: Test Loss for Noise Rate Estimation Error (error = $\pm 15\%$)

We observe further empirical evidence that underestimation can cause issues in terms of convergence behavior.

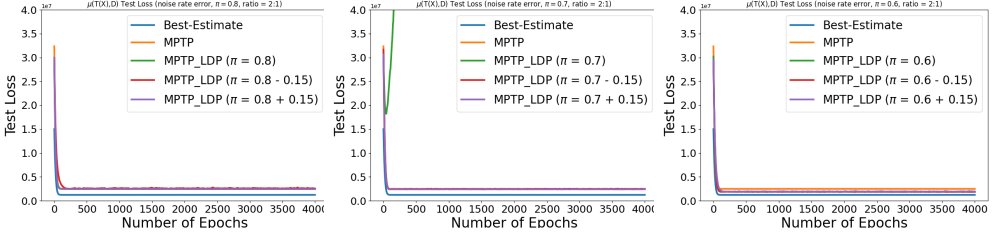
E DEFERRED INVESTIGATION OF NOISE RATE ESTIMATION ERROR

E.1 HEALTH INSURANCE

For completeness, we present the effect of noise rate estimation error on loss approximation when the privatized sensitive attributes S have unbalanced distribution.

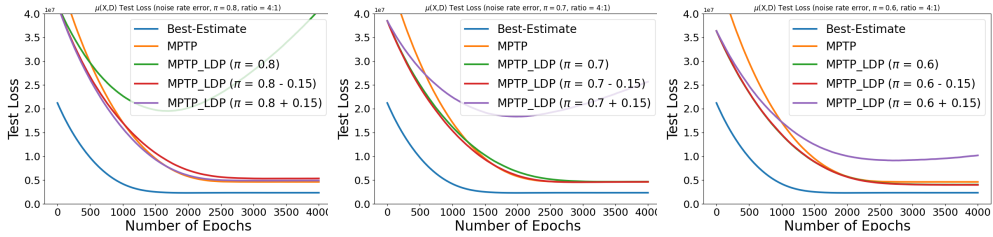


(a) $\mu(X, D)$ Test Loss ($\pi = 0.8$) (b) $\mu(X, D)$ Test Loss ($\pi = 0.7$) (c) $\mu(X, D)$ Test Loss ($\pi = 0.6$)

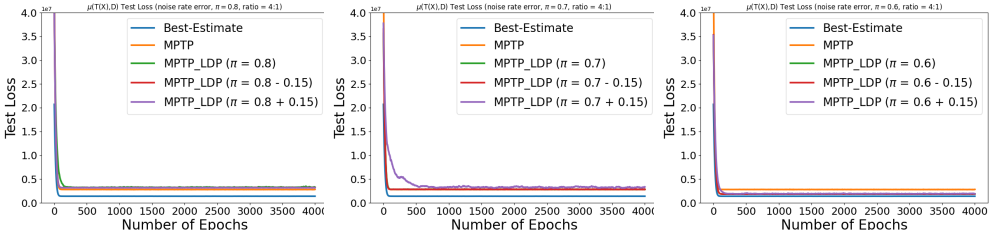


(d) $\mu(\tilde{X}, D)$ Test Loss ($\pi = 0.8$) (e) $\mu(\tilde{X}, D)$ Test Loss ($\pi = 0.7$) (f) $\mu(\tilde{X}, D)$ Test Loss ($\pi = 0.6$)

Figure 10: Test Loss for Noise Rate Estimation Error (ratio = 2:1, error = $\pm 15\%$)



(a) $\mu(X, D)$ Test Loss ($\pi = 0.8$) (b) $\mu(X, D)$ Test Loss ($\pi = 0.7$) (c) $\mu(X, D)$ Test Loss ($\pi = 0.6$)



(d) $\mu(\tilde{X}, D)$ Test Loss ($\pi = 0.8$) (e) $\mu(\tilde{X}, D)$ Test Loss ($\pi = 0.7$) (f) $\mu(\tilde{X}, D)$ Test Loss ($\pi = 0.6$)

Figure 11: Test Loss for Noise Rate Estimation Error (ratio = 4:1, error = $\pm 15\%$)

E.2 AUTO INSURANCE

For completeness, we present the effect of noise rate estimation error on loss approximation when the privatized sensitive attributes S have unbalanced distribution.

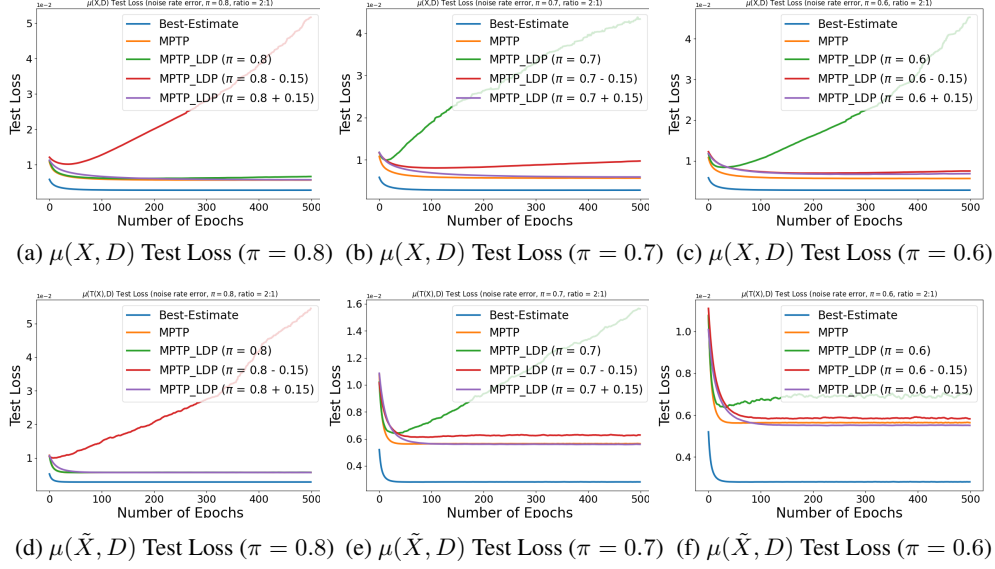


Figure 12: Test Loss for Noise Rate Estimation Error (ratio = 2:1, error = $\pm 15\%$)

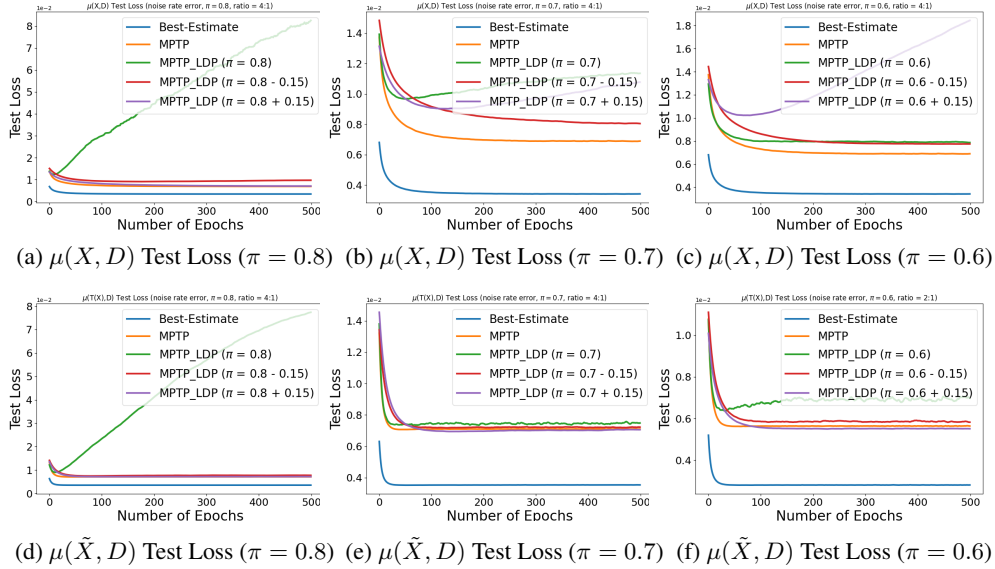


Figure 13: Test Loss for Noise Rate Estimation Error (ratio = 4:1, error = $\pm 15\%$)

F DEFERRED FIGURES

F.1 HEALTH INSURANCE

For completeness, we present the loss approximation performance under both scenarios of $h^*(X)$ computed with $\mathbb{P}^*(d)$ recovered using the empirical distribution of privatized sensitive attribute S .

F.1.1 SCENARIO 1 (KNOWN NOISE RATE)

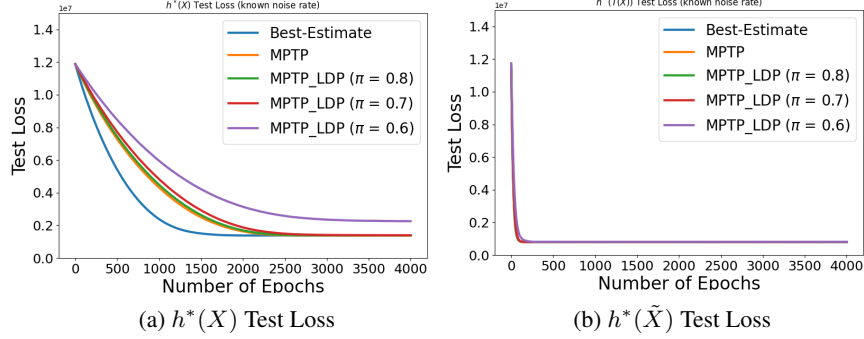


Figure 14: Test Loss for Scenario 1 (fixed sample size)

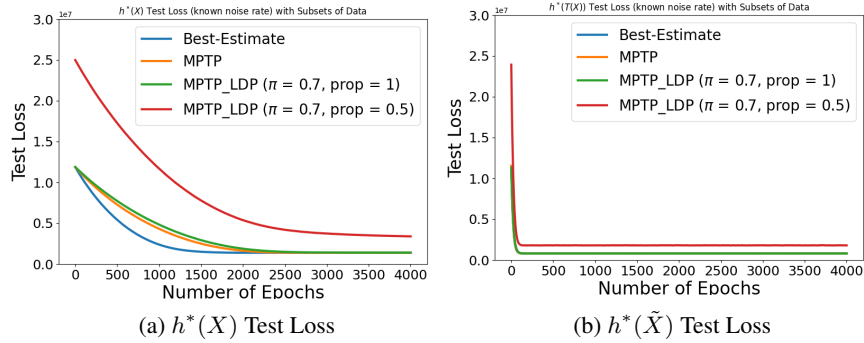


Figure 15: Test Loss for Scenario 1 (fixed noise rate: $\pi = 0.7$)

F.1.2 SCENARIO 2 (UNKNOWN NOISE RATE)

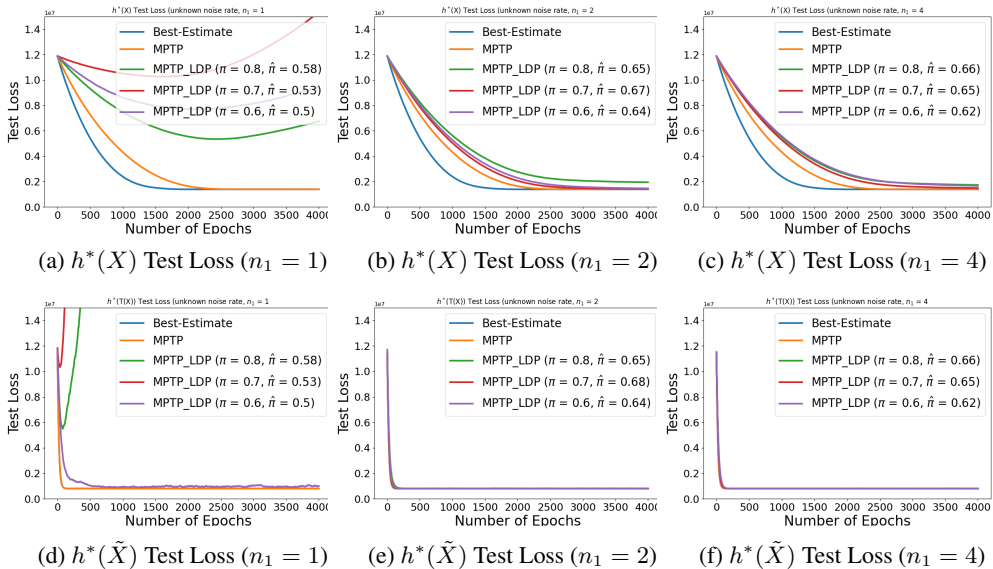
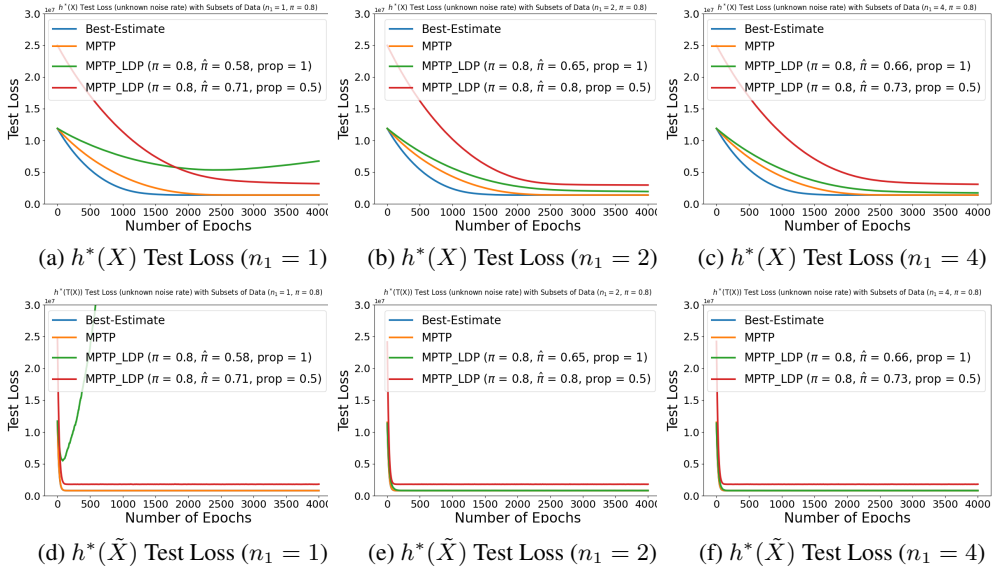


Figure 16: Test Loss for Scenario 2 (fixed sample size)

Figure 17: Test Loss for Scenario 2 (fixed noise rate: $\pi = 0.8$)

F.2 AUTO INSURANCE

For completeness, we present the loss approximation performance under both scenarios of $h^*(X)$ computed with $\mathbb{P}^*(d)$ recovered using the empirical distribution of privatized sensitive attribute S .

F.2.1 SCENARIO 1 (KNOWN NOISE RATE)

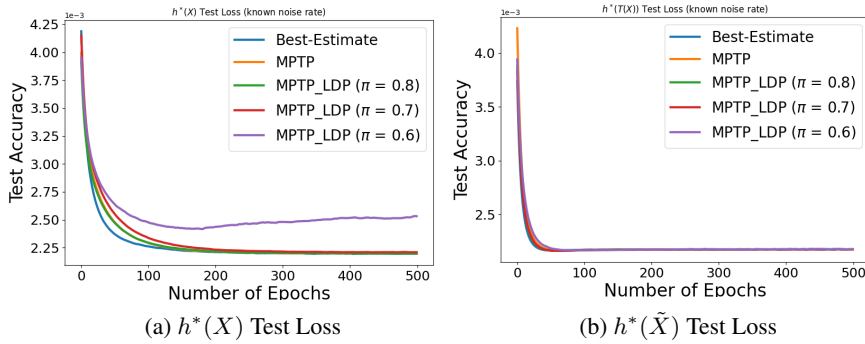
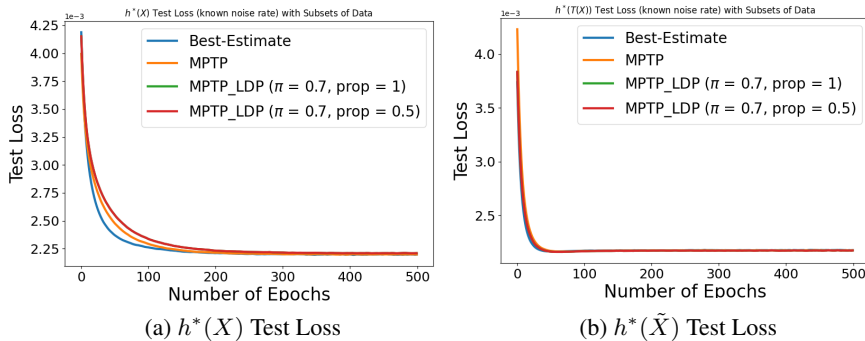


Figure 18: Test Loss for Scenario 1 (fixed sample size)

Figure 19: Test Loss for Scenario 1 (fixed noise rate: $\pi = 0.7$)

F.2.2 SCENARIO 2 (UNKNOWN NOISE RATE)

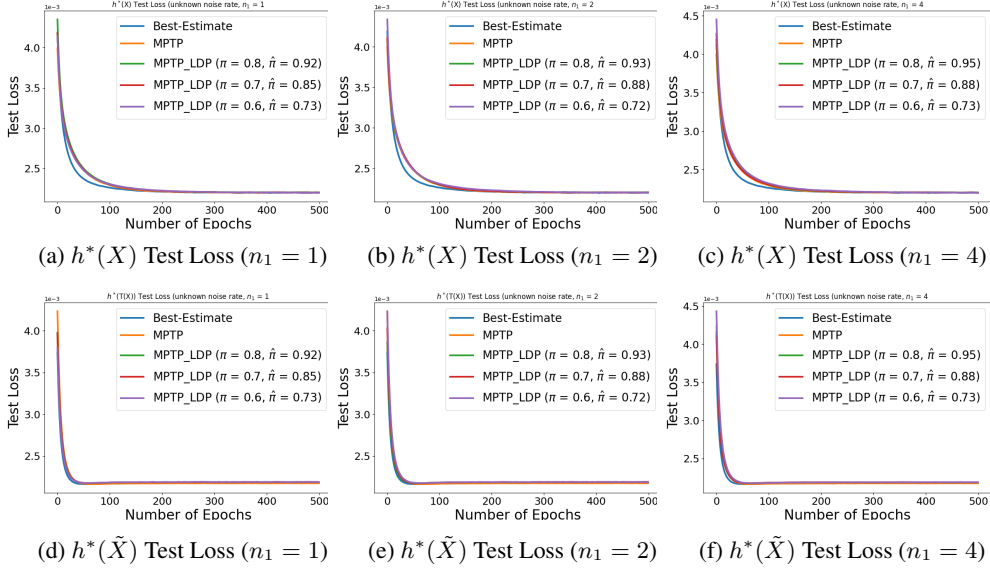
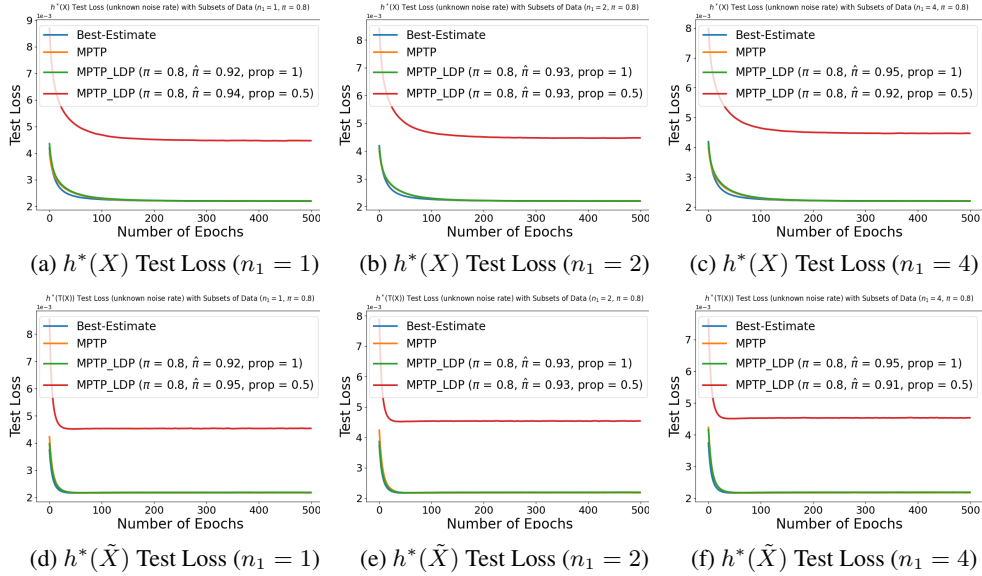


Figure 20: Test Loss for Scenario 2 (fixed sample size)

Figure 21: Test Loss for Scenario 2 (fixed noise rate: $\pi = 0.8$)