A SUMMARY OF UPDATES

We highlighted all revisions in blue.

A.1 UPDATE IN SECTION 1

We rearticulated some portion of the introduction section to make our work easier to follow since we believe adding more actuarial science related background to the introduction will make the audience understand better the motivation and the focus of the work. Specifically, we give definition for actuarial fair (see 1), interpretation of noise, and the generalization of the proposed method on non-sensitive attributes (see 1) (as Reviewer BixT, m8tn suggested).

A.2 UPDATE IN SECTION 2

We further reviewed existing work on algorithms that train discrimination-free models in the actuarial science literature (see 2.2) and pointed out the value and novelty of our work (as Reviewer m8tn, BixT suggested).

A.3 UPDATE IN SECTION 4.1

We slightly modified the example (Example 4.1) we presented to show how the choice of T and $\mathcal{F}$ relates to model transparency and performance so that both regression and classification tasks are included (as Reviewer rqKi suggested).

A.4 UPDATE IN SECTION 4.2

We relate the motivation of using LDP to the interpretation of noise to clarify that LDP is only used under two scenarios. 1) in data collection of vendors as an incentive for consumers to provide information about their sensitive attributes. 2) in data transmission for security purposes, note that 2) is not needed if the information the insurer collected is already privatized (as Reviewer m8tn suggested).

A.5 UPDATE IN SECTION 4.3

First, we added a more technical discussion (see 4.3) on assumptions A and B (especially on the relaxation of assumption B) but deferred the presentation to Appendix C and Appendix G.4 due to the page limit (as Reviewer BixT, rqKi suggested).

Second, We added a high-level description of the impact of the estimation error of π on the behavior of Risk-LDP (Eq. 7) (as Reviewer rqKi suggested).

A.6 UPDATE IN SECTION 5

First, we added another empirical experiment on regression task using an insurance data set (see 5.1) to show that our method is not limited to logistic loss but is compatible with other losses as well. The results we obtained are also in support of the theoretical guarantees we derived (as Reviewer rqKi suggested).

Second, we added an empirical study on the effect of the estimation error of noise rate on both evenly and unevenly distributed scenarios. The results and observations are presented in Appendix E due to page limit (as Reviewer rqKi suggested).

B DEFERRED ALGORITHMS

B.1 MPTP-D

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

Insurer Input: data: $\{X_i, Y_i\}_{i=1}^n$, hypothesis class: $\mathcal{H}$ (if obtain T via supervised learning)
Insurer Output: $\{T(X_i)\}_{i=1}^n$
TTP Input: data: $\{T(X_i), Y_i, D_i\}$, hypothesis class: $\mathcal{F}$, risk function: $\mathcal{R}(f_1, \ldots, f_{|\mathcal{D}|})$ (Eq. (1))
  repeat
    train $f_1, \ldots, f_{|\mathcal{D}|}$ by minimizing Eq. (1)
  until Convergence
  compute $h^*(T(X))$ using Eq. (2)
  return $f_1^*, \ldots, f_{|\mathcal{D}|}^*, h^*(X)$
TTP Output: $f_1^*, \ldots, f_{|\mathcal{D}|}^*, h^*(T(X))$

B.2 MPTP-S

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

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 $\tilde{T}$ via supervised learning)
Insurer Output: $\{T(X_i)\}_{i=1}^n, \{\tilde{T}(X_i)\}_{i=1}^n$
TTP Input: data: $\{T(X_i), Y_i, S_i\}_{i=1}^n, \{\tilde{T}(X_i), S_i\}_{i=1}^n$, hypothesis class $\mathcal{G}$, risk function: $\forall k \in [n_1]$, $\mathcal{R}(g_k) = \sum_{j=1}^m L(g_k(\tilde{T}_{k,j}, S_{k,j}))$ (see Lemma 4.4), hypothesis class: $\mathcal{F}$, risk function: $\mathcal{R}(f_1, \ldots, f_{|\mathcal{D}|})$ (Eq. (7)),
  if Scenario 2 ($\pi, \bar{\pi}$ unknown) **then**
    compute $\hat{\pi}_k, \hat{\bar{\pi}}_k, k \in [n_1]$ (by applying Lemma 4.4)
    compute $\hat{C}_1$ using $\hat{\pi}_k, \hat{\bar{\pi}}_k, k \in [n_1]$ (by C_1 estimation procedure 4.3)
    compute $\hat{\pi}, \hat{\bar{\pi}}$ using $\hat{C}_1$
    compute $\hat{\mathbf{\Pi}}^{-1}$ using $\hat{\pi}, \hat{\bar{\pi}}$
  else
    compute $\hat{\mathbf{\Pi}}^{-1}$ using $\pi, \bar{\pi}$
  end if
  repeat
    train $f_1, \ldots, f_{|\mathcal{D}|}$ by minimizing Eq. (7)
  until convergence
  compute $h^*(T(X))$ using Eq. (2)
  return $f_1^*, \ldots, f_{|\mathcal{D}|}^*, h^*(X)$
TTP Output: $f_1^*, \ldots, f_{|\mathcal{D}|}^*, h^*(T(X))$

C Deferred Discussion on Assumptions

C.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:

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

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

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

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.

C.2 Restrictions on Assumption B

For Assumption B, the condition is equivalent to $\mathbb{E}[\frac{1}{\hat{\pi}-1/|\mathcal{D}|}]$. Therefore, the closer π and $\frac{1}{|\mathcal{D}|}$ is the more accuracy of $\hat{\pi}$ is needed to suffice this assumption.

C.3 Relaxation of Assumption B

We do not acquire $\hat{C}_{1,k}$ to be strictly unbiased estimators of C_1– some perturbations can be allowed with some small modification to Theorem 4.5 but the general result still holds. Please see Section 4.3 and Appendix G.4 for a detailed discussion.

D DEFERRED EXPERIMENT RESULTS

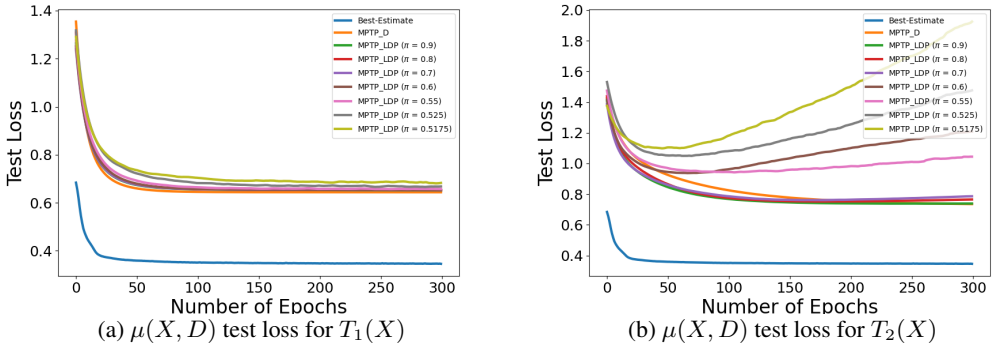
D.1 DATA

The Adult dataset contains 48842 observations, 14 features, and 1 target variable (income). In our experiment, we delete observations with missing values which results in a subset of 45222 observations. Further, we delete "fnlwgt", "education_num" where the former has no clear description and the latter is a duplicate of "education". We choose $D =$ sex to be our sensitive attribute which takes values of "male" and "female". The privatized sensitive attribute S is generated under different privacy levels using a set of ϵ's by Definition 4.1. D was only used to set the benchmark for performance and is masked under any other settings.

We conduct experiments 1) when the noise rate $\pi, \bar{\pi}$ are known (scenario 1) and 2) when the noise rates are unknown (scenario 2). For both scenarios, while we limited the hypothesis class $\mathcal{F}$ to the class of linear models, the insurer obtains two transformations T_1, T_2 for the main task and one transformation $\tilde{T}$ for noise rate estimation, where T_1 is obtained via supervised learning (same as example 4.1), $T_2, \tilde{T}$ are simply the identity. The reason that we choose such T_1, T_2 is to showcase the relationship between the complexity of T, model transparency, and performance on unseen data under the same $\mathcal{F}$. Further, under scenario 2, we set $n_1 = 1, 2, 4$ and conduct experiments for each n_1 respectively. With $\mathcal{F}$ being the class of linear models under both scenarios, TTP is essentially fitting a logistic regression w.r.t. $T_1(X)$ and $T_2(X)$ to obtain $\mu(T_1(X), D)$, and $\mu(T_2(X), D)$ respectively. For the calculation of $h^*(T(X))$, we choose the empirical marginal of D (estimated using S).

D.2 RESULTS

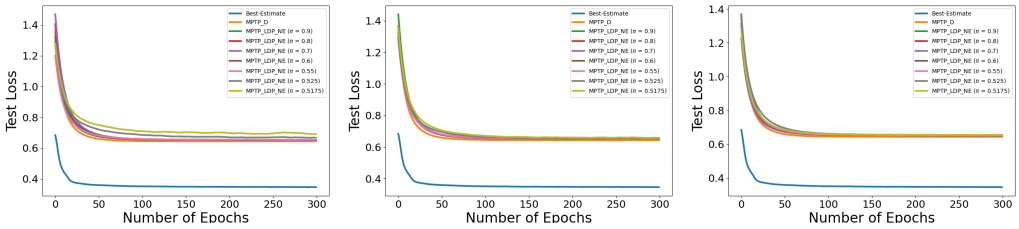
For each noise level, we generated S using 5 different seeds, hence each figure below (Figure 4, 5, 6) shows the mean values across all 5 different seeds. For both scenarios, we run experiments over 7 different privacy levels for $\pi = (0.9, 0.8, 0.7, 0.6, 0.55, 0.525, 0.5175)$. As the focus is to estimate $\mu(X, D)$, for conciseness, plots for test loss of $h^*(X)$ are deferred to Appendix F.



(a) $\mu(X, D)$ test loss for $T_1(X)$ (b) $\mu(X, D)$ test loss for $T_2(X)$

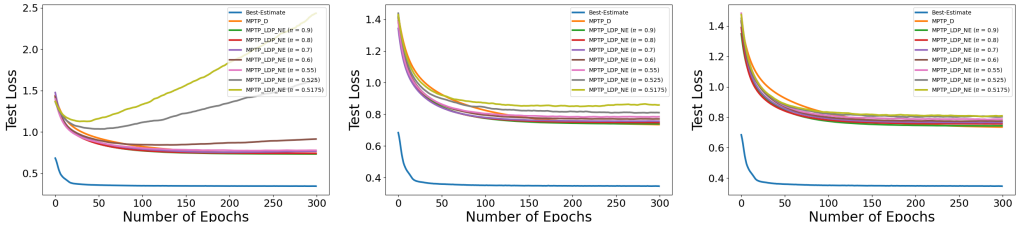
Figure 4: Test Loss for Scenario 1

From Figure 4, we observe that the T_1 is more robust against noise compared to T_2, and T_1 converges faster and has a better out-of-sample performance. Notice that as $\pi \to \frac{1}{|\mathcal{D}|}$, it requires a larger sample size to achieve the same loss approximation. Hence, for a fixed sample size, the larger the noise, the worse Eq. (7) approximates Eq. (1) which is in support of the result we obtain from Theorem 4.3. Although in terms of both accuracy and loss, there is a gap between Eq. (1), the trade-off comes from the ease of implementation (use of group-specific models) and transparency w.r.t. $T(X)$ in that we have limited $\mathcal{F}$ to be the class of linear models. Next, we present the test loss (see Figure 5, 6) for $\mu(X, D)$ estimation using T_1, T_2 under scenario 2 with $n_1 = 1, 2, 4$ respectively.



(a) $\mu(X, D)$ test loss with $n_1 = 1$ (b) $\mu(X, D)$ test loss with $n_1 = 2$ (c) $\mu(X, D)$ test loss with $n_1 = 4$

Figure 5: $\mu(X, D)$ test loss with $T_1(X)$ for scenario 2 with $n_1 = 1, 2, 4$



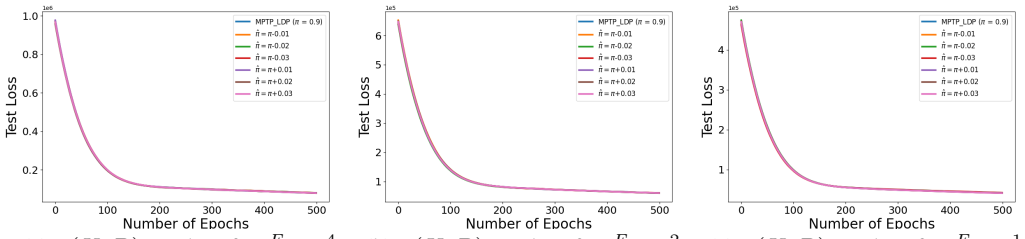
(a) $\mu(X, D)$ test loss with $n_1 = 1$ (b) $\mu(X, D)$ test loss with $n_1 = 2$ (c) $\mu(X, D)$ test loss with $n_1 = 4$

Figure 6: $\mu(X, D)$ test loss with $T_2(X)$ for scenario 2 with $n_1 = 1, 2, 4$

From Figure 5, 6, the loss behavior w.r.t. T_1, T_2 is similar to that under scenario 1 in general. However, as n_1 increases, we observe a better approximation of Risk-LDP (Eq. (7)) to Eq. (1) (more obvious under T_2). As n_1 increases, a smaller $\tilde{\epsilon}$ is achievable, hence resulting in a tighter bound as Theorem 4.5 suggests. Therefore, the experiment results under both scenarios align with our theoretical results.

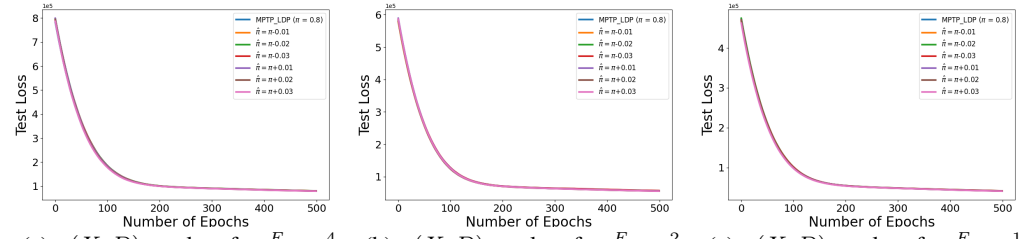
E DEFERRED INVESTIGATION OF NOISE RATE ESTIMATION ERROR

We manually set the error of estimation to be $\{\pm 0.01, \pm 0.02, \pm 0.03\}$, and our estimated noise rate is manually adjusted for each privacy level. Further, we created three subsets of the original data where the ratio between "Female" and "Male" observation is $\frac{4}{1}$, $\frac{2}{1}$ and $\frac{1}{1}$ to study the impact of the estimation error of Risk-LDP (Eq. 7) when the privatized sensitive attributes are unevenly distributed. We conduct experiments using T_1, T_2 with the manually adjusted erroneous noise rate on each subset respectively and compare the performance. We first present the results using T_1 below.



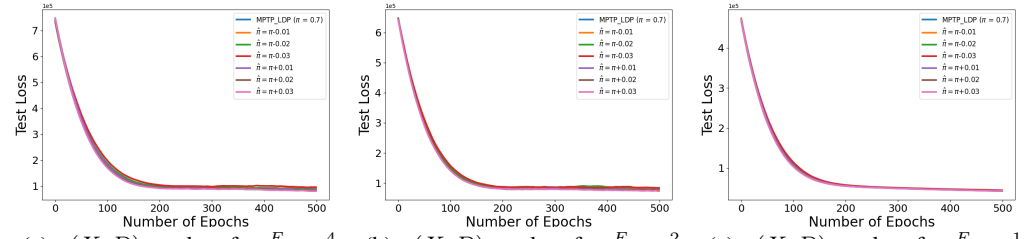
(a) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{4}{1}$ (b) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{2}{1}$ (c) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{1}{1}$

Figure 7: $\mu(X, D)$ Test Loss with $T_1(X)$ for Erroneous $\hat{\pi}$ when $\pi = 0.9$



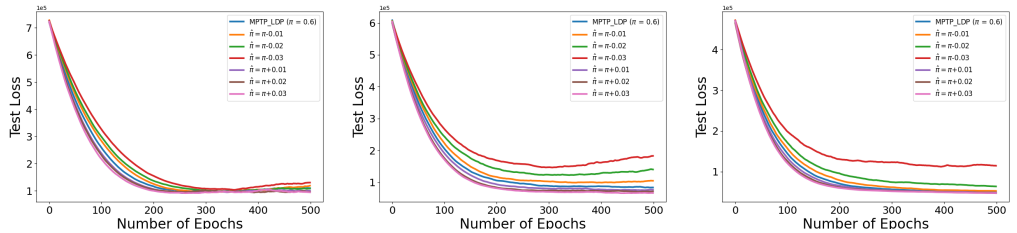
(a) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{4}{1}$ (b) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{2}{1}$ (c) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{1}{1}$

Figure 8: $\mu(X, D)$ Test Loss with $T_1(X)$ for Erroneous $\hat{\pi}$ when $\pi = 0.8$



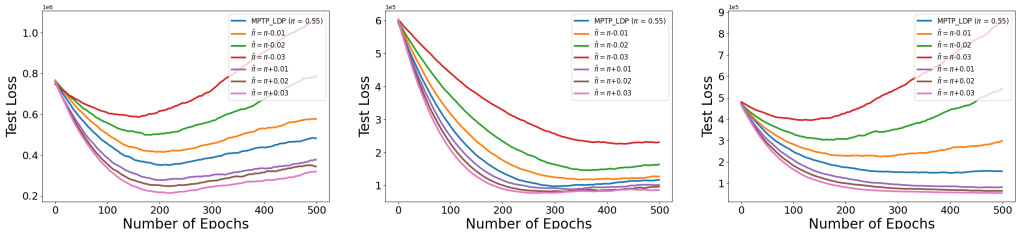
(a) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{4}{1}$ (b) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{2}{1}$ (c) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{1}{1}$

Figure 9: $\mu(X, D)$ Test Loss with $T_1(X)$ for Erroneous $\hat{\pi}$ when $\pi = 0.7$



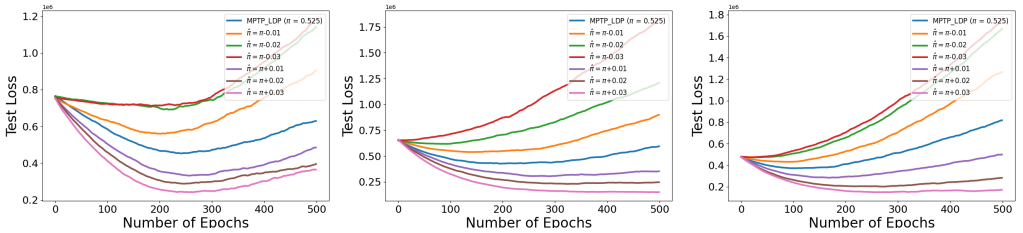
(a) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{4}{1}$ (b) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{2}{1}$ (c) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{1}{1}$

Figure 10: $\mu(X, D)$ Test Loss with $T_1(X)$ for Erroneous $\hat{\pi}$ when $\pi = 0.6$



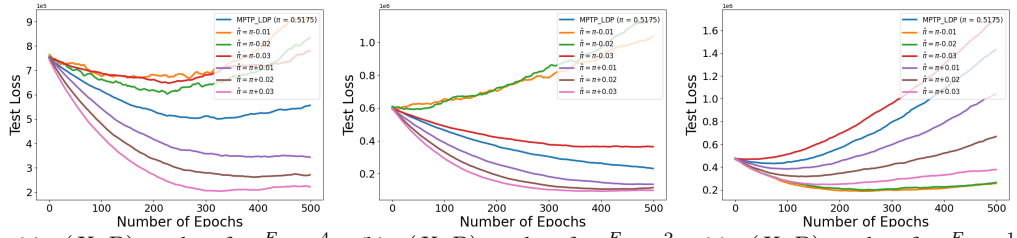
(a) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{4}{1}$ (b) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{2}{1}$ (c) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{1}{1}$

Figure 11: $\mu(X, D)$ Test Loss with $T_1(X)$ for Erroneous $\hat{\pi}$ when $\pi = 0.55$



(a) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{4}{1}$ (b) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{2}{1}$ (c) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{1}{1}$

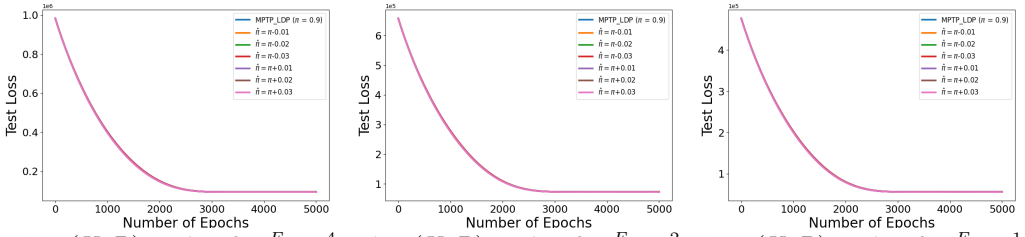
Figure 12: $\mu(X, D)$ Test Loss with $T_1(X)$ for Erroneous $\hat{\pi}$ when $\pi = 0.525$



(a) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{4}{1}$ (b) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{2}{1}$ (c) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{1}{1}$

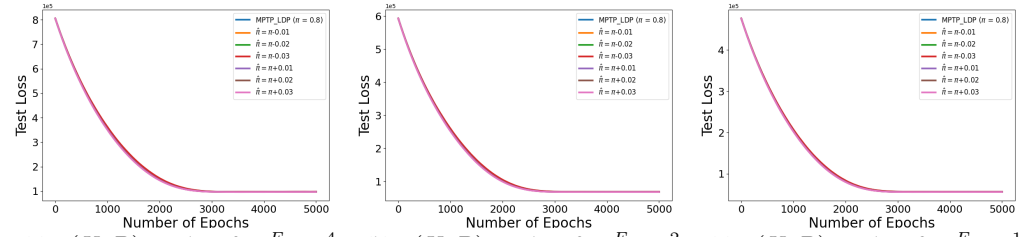
Figure 13: $\mu(X, D)$ Test Loss with $T_1(X)$ for Erroneous $\hat{\pi}$ when $\pi = 0.5175$

Next, we present the results using $T_2(X)$



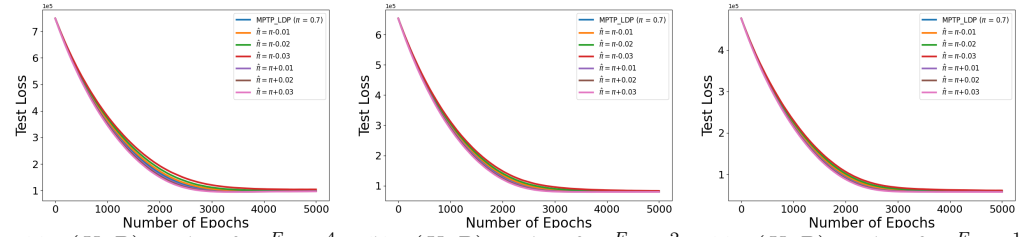
(a) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{4}{1}$ (b) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{2}{1}$ (c) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{1}{1}$

Figure 14: $\mu(X, D)$ Test Loss with $T_2(X)$ for Erroneous $\hat{\pi}$ when $\pi = 0.9$



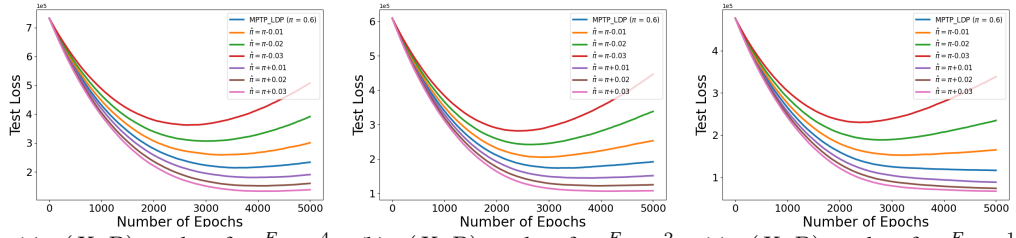
(a) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{4}{1}$ (b) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{2}{1}$ (c) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{1}{1}$

Figure 15: $\mu(X, D)$ Test Loss with $T_2(X)$ for Erroneous $\hat{\pi}$ when $\pi = 0.8$



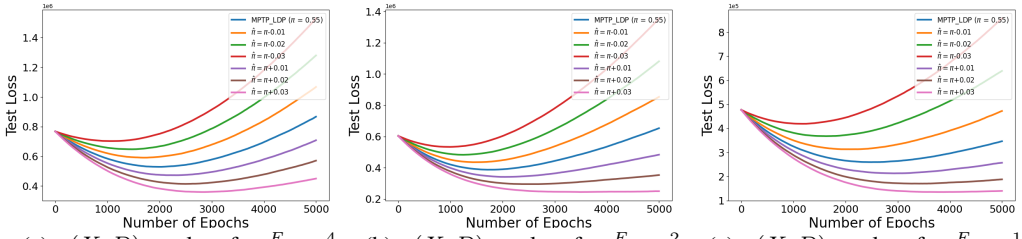
(a) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{4}{1}$ (b) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{2}{1}$ (c) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{1}{1}$

Figure 16: $\mu(X, D)$ Test Loss with $T_2(X)$ for Erroneous $\hat{\pi}$ when $\pi = 0.7$



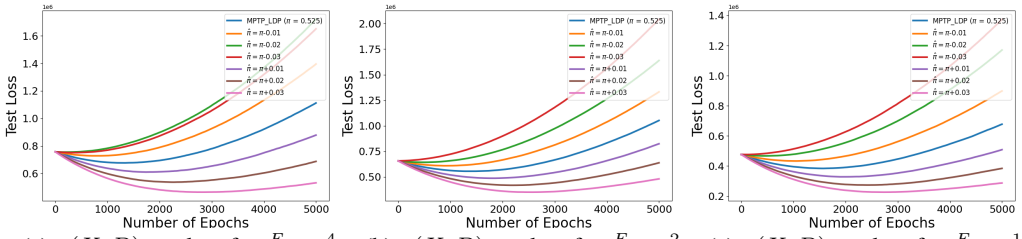
(a) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{4}{1}$ (b) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{2}{1}$ (c) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{1}{1}$

Figure 17: $\mu(X, D)$ Test Loss with $T_2(X)$ for Erroneous $\hat{\pi}$ when $\pi = 0.6$



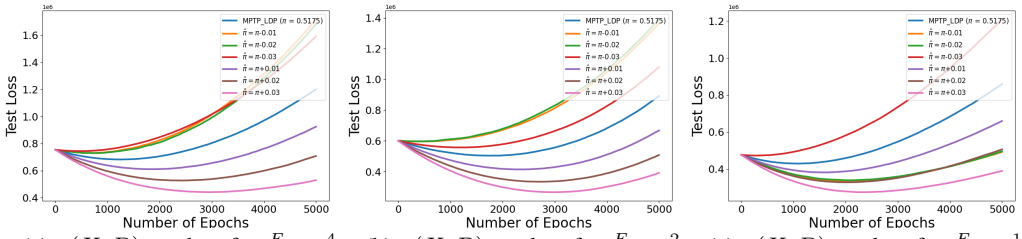
(a) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{4}{1}$ (b) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{2}{1}$ (c) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{1}{1}$

Figure 18: $\mu(X, D)$ Test Loss with $T_2(X)$ for Erroneous $\hat{\pi}$ when $\pi = 0.55$



(a) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{4}{1}$ (b) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{2}{1}$ (c) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{1}{1}$

Figure 19: $\mu(X, D)$ Test Loss with $T_2(X)$ for Erroneous $\hat{\pi}$ when $\pi = 0.525$



(a) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{4}{1}$ (b) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{2}{1}$ (c) $\mu(X, D)$ test loss for $\frac{F}{M} = \frac{1}{1}$

Figure 20: $\mu(X, D)$ Test Loss with $T_2(X)$ for Erroneous $\hat{\pi}$ when $\pi = 0.5175$

Observations: From the above figures we can see that T_1 converges much faster and is also more robust against estimation error on the noise rate than T_2 regardless of the distribution of sensitive attributes in general, combining the similar observation under scenario 1, we tend to conclude that the convergence rate regardless of the noise rate is known or unknown is considered closely related to the transformation chosen, further the robustness of Risk-LDP (Eq. 7) against noise rate estimation error is also impacted by the choice of transformation and this impact becomes more obvious as π gets closer to $\frac{1}{|\mathcal{D}|}$.

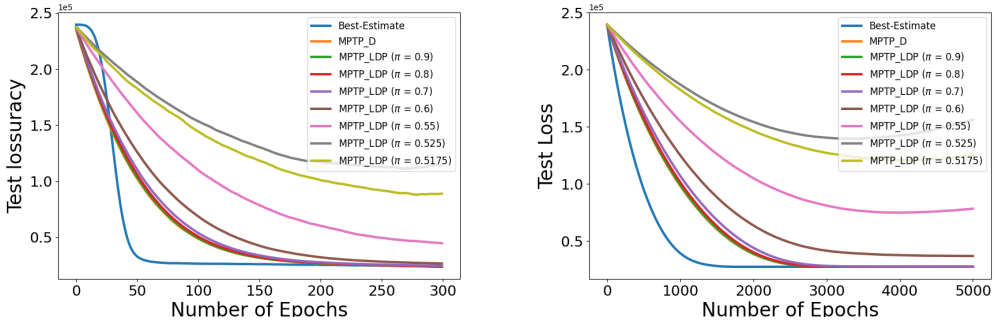
We also observe that when π is far away from $\frac{1}{|\mathcal{D}|}$ (in this case 0.5), regardless of the transformation chosen and the distribution of sensitive attribute, Risk-LDP (Eq. 7) is very robust against the estimation error (even when the error is large regardless overestimation or underestimation). However, as π becomes very close to $\frac{1}{|\mathcal{D}|}$ (Figure 11, 12, 13), we can see that underestimation of π is much more destructive than overestimation especially when the underestimation error is large.

As one should expect that different transformations should yield different behaviors of Risk-LDP. Fixing the transformation, we also noted that the distribution of the sensitive attributes does not have too much impact on the convergence behavior of Risk-LDP (Eq. 7). However, as the distribution becomes more imbalanced, Risk-LDP tends to give a higher loss than that of the less imbalanced scenario.

F DEFERRED FIGURES

F.1 INSURANCE

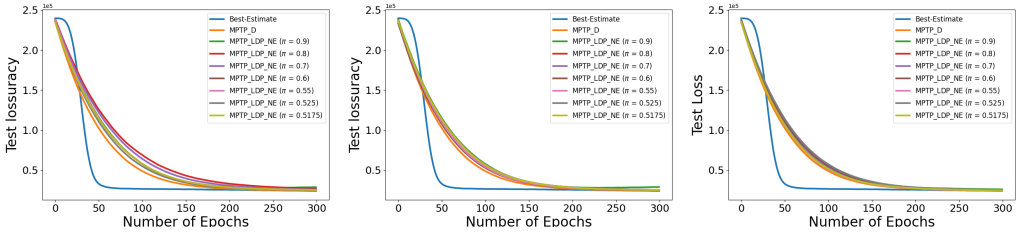
We now present the test loss for the estimation of $h^*(X)$ using T_1, T_2 under scenario 1 respectively:



(a) $h^*(X)$ test loss for $T_1(X)$ (b) $h^*(X)$ test loss for $T_2(X)$

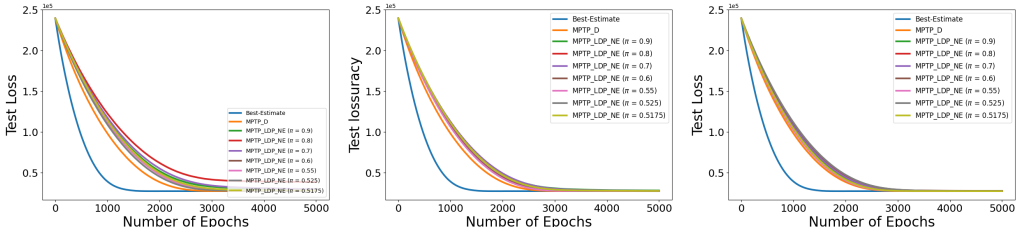
Figure 21: $h^*(X)$ test loss for scenario 1

Below is the test loss for $h^*(X)$ using T_1, T_2 with estimated π with $n_1 = 1, 2, 4$ under scenario 2 respectively:



(a) $h^*(X)$ test loss with $n_1 = 1$ (b) $h^*(X)$ test loss with $n_1 = 2$ (c) $h^*(X)$ test loss with $n_1 = 4$

Figure 22: $h^*(X)$ test loss with $T_1(X)$ for scenario 2 with $n_1 = 1, 2, 4$

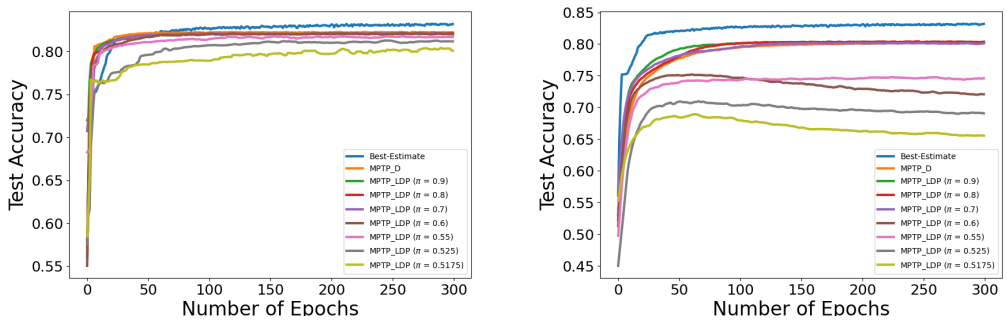


(a) $h^*(X)$ test acc with $n_1 = 1$ (b) $h^*(X)$ test acc with $n_1 = 2$ (c) $h^*(X)$ test acc with $n_1 = 4$

Figure 23: $h^*(X)$ test loss with $T_2(X)$ for scenario 2 with $n_1 = 1, 2, 4$

F.2 ADULT

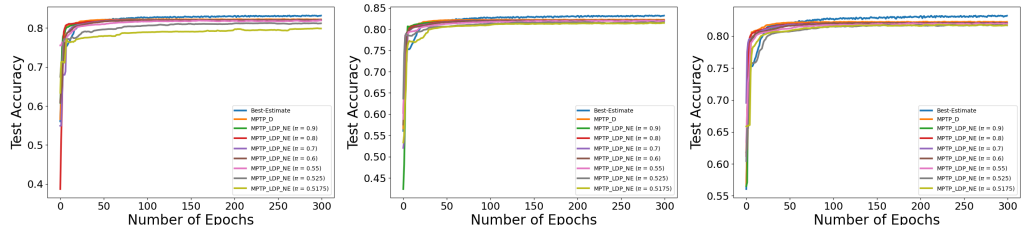
We first present test accuracy for the estimation of $\mu(X, D)$ and $h^*(X)$ using T_1, T_2 under scenario 1 respectively:



(a) $\mu(X, D)$ test acc for $T_1(X)$ (b) $\mu(X, D)$ test acc for $T_2(X)$

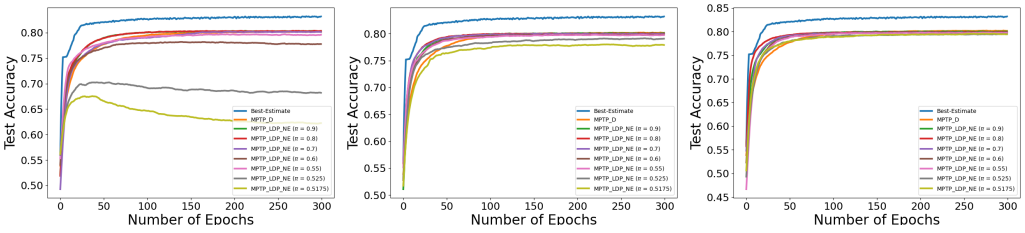
Figure 24: $\mu(X, D)$ test accuracy for scenario 1

Below is the test accuracy for the estimation of $\mu(X, D)$ and $h^*(X)$ using T_1, T_2 with $n_1 = 1, 2, 4$ under scenario 2 respectively:



(a) $\mu(X, D)$ test acc with $n_1 = 1$ (b) $\mu(X, D)$ test acc with $n_1 = 2$ (c) $\mu(X, D)$ test acc with $n_1 = 4$

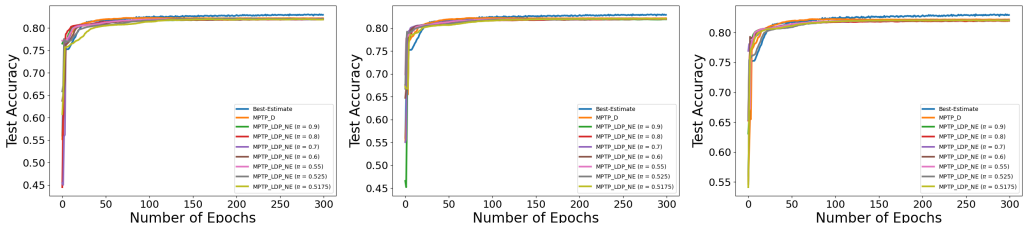
Figure 25: $\mu(X, D)$ test accuracy with $T_1(X)$ for scenario 2 with $n_1 = 1, 2, 4$



(a) $\mu(X, D)$ test acc with $n_1 = 1$ (b) $\mu(X, D)$ test acc with $n_1 = 2$ (c) $\mu(X, D)$ test acc with $n_1 = 4$

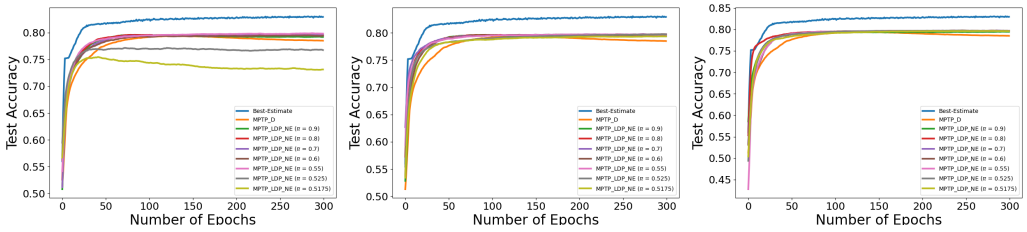
Figure 26: $\mu(X, D)$ test accuracy with $T_2(X)$ for scenario 2 with $n_1 = 1, 2, 4$

Next, we present the test accuracy for the estimation of $h^*(X)$ using T_1, T_2 with $n_1 = 1, 2, 4$ under scenario 2 respectively:



(a) $h^*(X)$ test acc with $n_1 = 1$ (b) $h^*(X)$ test acc with $n_1 = 2$ (c) $h^*(X)$ test acc with $n_1 = 4$

Figure 27: $h^*(X)$ test accuracy with $T_1(X)$ for scenario 2 with $n_1 = 1, 2, 4$



(a) $h^*(X)$ test acc with $n_1 = 1$ (b) $h^*(X)$ test acc with $n_1 = 2$ (c) $h^*(X)$ test acc with $n_1 = 4$

Figure 28: $h^*(X)$ test accuracy with $T_2(X)$ for scenario 2 with $n_1 = 1, 2, 4$

G Deferred Proofs

G.1 Lemma 4.2 Proof

Lemma 4.2 Given the privacy parameter ϵ, minimizing the following risk (Risk-LDP) Eq. (7) 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:

$$\mathcal{R}^{LDP}(f_1, \ldots, f_k) = \sum_{k=1}^{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} \mathbf{\Pi}_{kj}^{-1} \mathbb{E}_{Y,T(X)|S=j} \Big[L\big(Y, f_k(T(X))\big) \Big], \tag{8}$$

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 $(T(X), Y, \hat{Y})$, then consider the following probability:

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

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

$$\begin{aligned}
&\mathbb{P}(Y, T(X) \mid S = d) \\
&= \sum_{d' \in D} \mathbb{P}(Y, T(X) \mid S = d, D = d') \mathbb{P}(D = d' \mid S = d) \\
&= \sum_{d' \in D} \mathbb{P}(Y, T(X) \mid D = d') \mathbb{P}(D = d' \mid S = d) \\
&= \sum_{d' \in D} \mathbb{P}(Y, T(X) \mid D = d') \frac{\mathbb{P}(S = d \mid D = d') \mathbb{P}(D = d')}{\sum_{d'' \in D} \mathbb{P}(S = d \mid D = d'') \mathbb{P}(D = d'')} \\
&= P(Y, T(X) \mid D = d) \frac{\pi \mathbb{P}(D = d)}{\pi \mathbb{P}(D = d) + \sum_{d'' \setminus d} \bar{\pi} \mathbb{P}(D = d'')} + \sum_{d' \setminus d} P(Y, T(X) \mid D = d') \frac{\bar{\pi} \mathbb{P}(D = d')}{\pi \mathbb{P}(D = d) + \sum_{d'' \setminus d} \bar{\pi} \mathbb{P}(D = d'')}.
\end{aligned}$$

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:

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

denote as $\boldsymbol{s_1} = \mathbf{\Pi} \boldsymbol{d_1}$, where $\boldsymbol{s_1} = \mathbb{P}(Y, T(X) \mid S)$, $\boldsymbol{d_1} = \mathbb{P}(Y, T(X) \mid D)$.

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

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

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

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

where $\mathbf{\Pi}^{-1}_{k\cdot}$ 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 $\mathbb{P}(Y, X \mid D)$. 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 $\boldsymbol{T}$ be an $|\mathcal{D}| \times |\mathcal{D}|$ matrix with the following entries:

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

then we have the following system of linear equations:

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

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

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

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

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

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

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

At the population level, we have recovered that:

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

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

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

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

Therefore, we conclude that it is equivalent to minimizing:

$$(f_{1^*}, \ldots, f_{k^*}) \leftarrow \arg\min_{f_1,\ldots,f_k} \sum_{k=1}^{|\mathcal{D}|} \sum_{j=1}^{|\mathcal{D}|} \mathbf{\Pi}_{kj}^{-1} \mathbb{E}_{Y,T(X)|S=j}\Big[L\big(Y, f_k(T(X))\big)\Big]$$

This completes the proof. □

G.2 LEMMA 4.4 PROOF

Lemma 4.4 Under ϵ-LDP setting, with $\pi \in (\frac{1}{|\mathcal{D}|}, 1]$, $\bar{\pi} \in [0, \frac{1}{|\mathcal{D}|})$, assuming that there exists an anchor points $\tilde{T}(X)^*$ s.t. $\mathbb{P}(D = j^* | \tilde{T}(X)^*) = 1$ for some $j^* \in [|\mathcal{D}|]$, then $\pi = \mathbb{P}(S = j^* | \tilde{T}(X)^*)$. Empirically, denote the n-dimension vector $\boldsymbol{\eta}_s(\tilde{T}(X)^*) = \big(\hat{\mathbb{P}}(S = j^* | \tilde{T}(X_1)), \ldots, \hat{\mathbb{P}}(S = j^* | \tilde{T}(X_n))\big)$, then $\hat{\pi} = \|\boldsymbol{\eta}_s(\tilde{T}(X)^*)\|_\infty$ and $\{\hat{P}(S = j^* | \tilde{T}(X_i))\}_{i=1}^n$ can be obtained by specifying a hypothesis class $\mathcal{G}$ and minimize the following empirical risk:

$$\hat{\mathcal{R}}(k) = \sum_{i=1}^{n} L\big(k(\tilde{T}(X_i)), S_i\big).$$

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

$$\begin{aligned}
\mathbb{P}(S = j^* | \tilde{T}(X)^*) &= \sum_{k=1}^{|\mathcal{D}|} \mathbb{P}(S = j^* | D = k) \cdot \mathbb{P}(D = k | \tilde{T}(X)^*) \\
&\overset{(a)}{=} \sum_{k=1}^{|\mathcal{D}|} \mathbb{P}(S = j^* | D = k) \cdot \mathbf{1}\{j^* = k\} \\
&= \pi,
\end{aligned}$$

(a) is by followed by the definition of anchor point

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

Then one can easily see that $\mathbb{P}(S=j^*|\tilde{T}(X_i))$ attains its maximum when $\mathbb{P}(D=j^*|\tilde{T}(X_i))=1$, 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 $\mathbb{P}(S=j^*|\tilde{T}(X_i))$ is actually a weighted sum of π and $\bar{\pi}$, where the weights are simply $\{\mathbb{P}(D=k|\tilde{T}(X_i))\}_{k=1}^{|\mathcal{D}|}$. But we also know that $\pi>\bar{\pi}$. Hence, for empirical estimation, denote the n-dimension probability vector $\boldsymbol{\eta}_s(\tilde{T}(X)^*)=\Big(\hat{\mathbb{P}}(S=j^*|\tilde{T}(X_1)),\ldots,\hat{\mathbb{P}}(S=j^*|\tilde{T}(X_n))\Big)$

$$\hat{\pi}=\|\boldsymbol{\eta}_s(\tilde{T}(X)^*)\|_\infty.$$

Where $\{\hat{P}(S=j^*|\tilde{T}(X_i))\}_{i=1}^n$ can be obtained by specifying a hypothesis class $\mathcal{K}$ and minimize the following empirical risk:

$$\hat{\mathcal{R}}(k)=\sum_{i=1}^n L\big(k(\tilde{T}(X_i)),S_i\big).$$

This completes the proof □

G.3 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 $VC(\mathcal{F})$, then under a given loss function $L:Y\times Y\to\mathbb{R}_+$, and for $f=\{f_k\}_{k=1}^{|\mathcal{D}|}$ where $f_k\in\mathcal{F},\forall k\in[|\mathcal{D}|]$ with $f_k:T(\mathcal{X})\to\mathbb{R}_+$ s.t. $\sup_{X\in\mathcal{X}}|f_k(T(X))|\leq M\in\mathbb{R}_+,\forall k\in[|\mathcal{D}|]$ derived from Lemma 4.2, consequently, $L(f_k(T(X),Y))\leq\phi(M)\in\mathbb{R}_+,\forall k\in[|\mathcal{D}|],X\in\mathcal{X},Y\in\mathcal{Y}$, where ϕ is some function of M, denote $k^*\leftarrow\arg\max_k|\hat{\mathcal{R}}^{LDP}(f_k)-\mathcal{R}^{LDP}(f_k)|$, if $n\geq\frac{8\ln(\frac{|\mathcal{D}|}{\delta})}{\min_k\mathbb{P}(S=k)}$ then with probability $1-2\delta$:

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

Proof. For better presentation, denote $X=T(X)$ and $\tilde{X}=\tilde{T}(X)$ in the proof.

Step 1: simplify the objective

Denote $\mathcal{R}(f_k)$ 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

$$\begin{aligned}
&\mathbb{P}(|\hat{\mathcal{R}}^{LDP}(f)-\mathcal{R}(f)|>\epsilon)\\
=&\mathbb{P}(|\hat{\mathcal{R}}^{LDP}(f)+\mathcal{R}^{LDP}(f)-\mathcal{R}^{LDP}(f)-\mathcal{R}(f)|>\epsilon)\\
\leq&\mathbb{P}(|\hat{\mathcal{R}}^{LDP}(f)-\mathcal{R}^{LDP}(f)|+|\mathcal{R}^{LDP}(f)-\mathcal{R}(f)|>\epsilon)\\
\overset{(a)}{=}&\mathbb{P}(|\hat{\mathcal{R}}^{LDP}(f)-\mathcal{R}^{LDP}(f)|\geq\epsilon)\\
=&\mathbb{P}\Big(|\sum_{k=1}^{|\mathcal{D}|}\hat{\mathcal{R}}^{LDP}(f_k)-\sum_{k=1}^{D}\mathcal{R}^{LDP}(f_k)|>\epsilon\Big)\\
\leq&\mathbb{P}\Big(\sum_{k=1}^{D}|\hat{\mathcal{R}}^{LDP}(f_k)-\mathcal{R}^{LDP}(f_k)|>\epsilon\Big)\\
\overset{(b)}{\leq}&\mathbb{P}\Big(\max_k|\hat{\mathcal{R}}^{LDP}(f_k)-\mathcal{R}^{LDP}(f_k)|>\frac{\epsilon}{|\mathcal{D}|}\Big)\\
\overset{(c)}{=}&\mathbb{P}\Big(\Big|\sum_{j=1}^{|\mathcal{D}|}\hat{\mathbf{\Pi}}_{k^*j}^{-1}\frac{1}{n_j}\sum_{i:S_i=j}L\big(Y_i,f_{k^*}(T(X_i))\big)-\mathbf{\Pi}_{k^*\cdot}^{-1}\mathbb{E}_{Y,X|S}\Big[L\big(Y,f_{k^*}(X)\big)\Big]\Big|>\frac{\epsilon}{|\mathcal{D}|}\Big),
\end{aligned}$$

where (a) is obtained from the population equivalence of two losses from Lemma 4.2.

(b) is followed by for two events A, B, if A implies B then $P(A) < P(B)$, also denote that $k^* \leftarrow \arg\max_k |\hat{\mathcal{R}}^{LDP}(f_k) - \mathcal{R}^{LDP}(f_k)|$.

(c) is obtained by expanding the expression for $\hat{\mathcal{R}}^{LDP}(f_{k^*})$ and $\mathcal{R}^{LDP}(f_{k^*})$ respectively.

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

Denote $n^N_{yxs} = \sum_i \mathbf{1}(y_i = y, x_i = x, s_i = s)$, $\boldsymbol{Q}_{yxs} = \mathbb{P}(Y = y, X = x, S = j)$, and define the random variable $N_{yxs} = \{i \mid y_i = y, x_i = x, s_i = s\}$. We can deduce $n^N_s = \sum_{x\in X, y\in Y} \mathbf{1}(y_i = y, x_i = x, s_i = s)$. Then, we have $\mathbb{E}[\hat{\mathcal{R}}^{LDP}(f_{k^*}) \mid N_{YXS}] = \mathcal{R}^{LDP}(f_{k^*})$, where N_{YXS} denotes all possible N_{yxs}. Using similar approach of Lemma 2 in Mozannar et al. (2020), we can write:

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

where (a) follows by conditioning over all $2^n|X|^n|\mathcal{D}|^n$ possible configurations of $N_{yxs} \subset [n]$. (b) is obtained by splitting the configurations where $\forall x, y, N_{yxs} : n^N_s \geq \frac{n\sum_{x\in X, y\in Y} \boldsymbol{Q}_{yxs}}{2}$ and the complement of the event and upper bound the complement of the event by the probability that $\exists s$ s.t. $n^N_s < \frac{n\sum_{x\in X, y\in Y} \boldsymbol{Q}_{yxs}}{2}$. (c) is obtained by the union bound and we know $n^N_s \sim \text{Binomial}\big(n, \sum_{x\in X, y\in Y} \boldsymbol{Q}_{yxs}\big)$ and apply the Chernoff bound on n^N_{yxj}.

Now, we will apply the McDiarmid Inequality McDiarmid (1989). Let $X^n = (X_1, \ldots, X_n) \in X^n$ be n independent random variables and let $g : X^n \to \mathbb{R}$, if there exists constants $c_1, \ldots, c_n$ s.t.

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

then $\forall \epsilon > 0$:

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

Since by conditioning on N_{YXS}, then for $\hat{\mathcal{R}}^{LDP}(f_{k^*})$, everything else is now deterministic except for f_{k^*}, in other words, by conditioning on N_{YXS}, 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 $f_{k^*}(X_i)$, we try to bound how much f_{k^*} can change.

Recall from Lemma 4.2, we computed the entries of $\mathbf{\Pi}^{-1}$ takes the following form:

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

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 $\mathbf{\Pi}^{-1}$ in terms of $\boldsymbol{P}_s$ where $\boldsymbol{P}_s = \mathbb{P}(S)$:

$$\begin{cases} \mathbf{\Pi}^{-1}_{i,i} = C_1 \frac{\pi \boldsymbol{T}^{-1}_{i\cdot}\boldsymbol{P}_s + \sum_{l\setminus i} \bar{\pi}\boldsymbol{T}^{-1}_{l\cdot}\boldsymbol{P}_s}{\boldsymbol{T}^{-1}_{i\cdot}\boldsymbol{P}_s}, \text{for } i \in D \\ \mathbf{\Pi}^{-1}_{i,j} = C_2 \frac{\bar{\pi} \boldsymbol{T}^{-1}_{i\cdot}\boldsymbol{P}_s + \sum_{l\setminus i} \pi\boldsymbol{T}^{-1}_{l\cdot}\boldsymbol{P}_s}{\boldsymbol{T}^{-1}_{i\cdot}\boldsymbol{P}_s}, \text{for } i,j \in D \text{ s.t.}, i \neq j \end{cases},$$

we also computed $\boldsymbol{T}^{-1}$ as:

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

then we have

$$\begin{aligned}
&\sup_{N,N'} |\hat{\mathcal{R}}^{LDP}(f_{k^*})^N - \hat{\mathcal{R}}^{LDP}(f_{k^*})^{N'}| \\
=&\Big| C_1 \frac{\pi \boldsymbol{T}^{-1}_{k^*\cdot}\boldsymbol{P}^N_s + \sum_{l\setminus k} \bar{\pi}\boldsymbol{T}^{-1}_{l\cdot}\boldsymbol{P}^N_s}{\boldsymbol{T}^{-1}_{k^*\cdot}\boldsymbol{P}^N_s} \hat{\mathcal{R}}^{LDP}(f_{k^*})^N + \sum_{j\setminus k} C_2 \frac{\bar{\pi} \boldsymbol{T}^{-1}_{k^*\cdot}\boldsymbol{P}^N_s + \sum_{l\setminus k} \pi\boldsymbol{T}^{-1}_{l\cdot}\boldsymbol{P}^N_s}{\boldsymbol{T}^{-1}_{k^*\cdot}\boldsymbol{P}^N_s} \hat{\mathcal{R}}^{LDP}(f_{k^*})^N \\
&- C_1 \frac{\pi \boldsymbol{T}^{-1}_{k^*\cdot}\boldsymbol{P}^N_s + \sum_{l\setminus k} \bar{\pi}\boldsymbol{T}^{-1}_{l\cdot}\boldsymbol{P}^N_s}{\boldsymbol{T}^{-1}_{k^*\cdot}\boldsymbol{P}^N_s} \hat{\mathcal{R}}^{LDP}(f_{k^*})^{N'} + \sum_{j\setminus k} C_2 \frac{\bar{\pi} \boldsymbol{T}^{-1}_{k^*\cdot}\boldsymbol{P}^N_s + \sum_{l\setminus k} \pi\boldsymbol{T}^{-1}_{l\cdot}\boldsymbol{P}^N_s}{\boldsymbol{T}^{-1}_{k^*\cdot}\boldsymbol{P}^N_s} \hat{\mathcal{R}}^{LDP}(f_{k^*})^{N'} \Big| \\
=&\Big| C_1 \frac{\pi \boldsymbol{T}^{-1}_{k^*\cdot}\boldsymbol{P}^N_s + \sum_{l\setminus k} \bar{\pi}\boldsymbol{T}^{-1}_{l\cdot}\boldsymbol{P}^N_s}{\boldsymbol{T}^{-1}_{k^*\cdot}\boldsymbol{P}^N_s} \Big(\frac{\sum_{i\in N, x\in X, y\in Y, S=k} L\big(y_i, f_{k^*}(x_i)\big)}{n_{k^*}} - \frac{\sum_{i\in N', x\in X, y\in Y, S=k} L\big(y_i, f_{k^*}(x_i)\big)}{n_{k^*}} \Big) \\
&+ \sum_{j\setminus k} C_2 \frac{\bar{\pi} \boldsymbol{T}^{-1}_{k^*\cdot}\boldsymbol{P}^N_s + \sum_{l\setminus k} \pi\boldsymbol{T}^{-1}_{l\cdot}\boldsymbol{P}^N_s}{\boldsymbol{T}^{-1}_{k^*\cdot}\boldsymbol{P}^N_s} \Big(\frac{\sum_{i\in N, x\in X, y\in Y, S=j} L\big(y_i, f_{k^*}(x_i)\big)}{n_{k^*}} - \frac{\sum_{i\in N', x\in X, y\in Y, S=j} L\big(y_i, f_{k^*}(x_i)\big)}{n_{k^*}} \Big) \Big| \\
\overset{(a)}{\leq}&\Big| C_1 \frac{\pi \max_{m\in[|\mathcal{D}|]} \boldsymbol{T}^{-1}_{m\cdot}\boldsymbol{P}^N_s + (|\mathcal{D}|-1)\bar{\pi} \max_{m\in[|\mathcal{D}|]} \boldsymbol{T}^{-1}_{m\cdot}\boldsymbol{P}^N_s}{\boldsymbol{T}^{-1}_{k^*\cdot}\boldsymbol{P}^N_s} \cdot \frac{\phi(M)}{n_{k^*}} \Big| \\
=&\Big| C_1 \big(\pi + \bar{\pi}(|\mathcal{D}|-1)\big) \cdot \frac{\phi(M)}{n_{k^*}} \Big| \\
\overset{(b)}{=}&\Big| \frac{C_1\phi(M)}{n_{k^*}} \Big|,
\end{aligned}$$

where (a) is obtained by $C_2 \leq 0, \forall \pi \in (\frac{1}{|\mathcal{D}|}, 1]$. (b) is followed by the fact that $\pi + \bar{\pi}(|\mathcal{D}|-1) = 1$.

Now, we are ready to apply the McDiarmid Inequality:

$$\begin{aligned}
&\sum_{\forall x,y,N_{yxs}: n^N_s \geq \frac{n\sum_{x\in X, y\in Y} \boldsymbol{Q}_{yxs}}{2}} \mathbb{P}\Big(\hat{\mathcal{R}}^{LDP}(f_{k^*})^N - \mathcal{R}^{LDP}(f_{k^*}) > \frac{\epsilon}{|\mathcal{D}|} \Big| N_{YXS} \Big) \cdot \mathbb{P}(N_{YXS}) \\
\leq &\sum_{\forall x,y,N_{yxs}: n^N_s \geq \frac{n\sum_{x\in X, y\in Y} \boldsymbol{Q}_{yxs}}{2}} 2\exp\left\{ - \frac{\frac{2\epsilon^2}{|\mathcal{D}|^2}}{n\cdot\big(\frac{C_1\phi(M)}{n_{k^*}}\big)^2} \right\} \cdot \mathbb{P}(N_{YXS}) \\
\overset{(a)}{\leq} &2\exp\left\{ - 2n\epsilon^2 \Big(\frac{\mathbb{P}(S=k)}{2C_1\phi(M)|\mathcal{D}|} \Big)^2 \right\},
\end{aligned}$$

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

Now, we have:

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

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

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

Step 3: Obtain the final result

Recall that one can easily show

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

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:

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

This completes the proof. □

G.4 THEOREM 4.5 PROOF

Theorem 4.5 For any $\delta \in (0, \frac{1}{3})$, $C_1 = \frac{\pi+|\mathcal{D}|-2}{|\mathcal{D}|\pi-1} > 0$, $\hat{C}_1 = \frac{1}{n_1}\sum_{k=1}^{n_1} \hat{C}_{1,k}$, where $\hat{C}_{1,k}$ is defined in Lemma 4.4, 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 (4.3),B (4.3), and Lemma 4.4 hold, given a loss function $L : Y \times Y \to \mathbb{R}_+$, $M_g + \frac{C_1+\theta}{\ln 2} > \tilde{\epsilon} > \theta$, and for $f = \{f_k\}_{k=1}^{|\mathcal{D}|}$ where $f_k \in \mathcal{F}, \forall k \in [|\mathcal{D}|]$ with $f_k : T(\mathcal{X}) \to \mathbb{R}_+$ s.t. $\sup_{X\in\mathcal{X}} |f_k(T(X))| \leq M \in \mathbb{R}_+, \forall k \in [|\mathcal{D}|]$ derived from Lemma 4.2, consequently, $L(f_k(T(X), Y)) \leq \phi(M) \in \mathbb{R}_+, \forall k \in [|\mathcal{D}|], X \in \mathcal{X}, Y \in \mathcal{Y}$, where ϕ is some function of M, denote $k^* \leftarrow \arg\max_k |\hat{\mathcal{R}}^{LDP}(f_k) - \mathcal{R}^{LDP}(f_k)|$, if $n \geq \frac{8\ln(\frac{|\mathcal{D}|}{\delta})}{\min_k \mathbb{P}(S=k)}$, $n_1 \geq \frac{1}{c(\tilde{\epsilon}-\theta)^2}(M_g + \frac{C_1+\theta}{\ln 2})^2 \ln(\frac{2}{\delta})$ where c is an absolute constant, then with probability $1 - 3\delta$:

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

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: Given the observed data $\{T(X_i), S_i\}_{i=1}^n$, we evenly divide them into n_1 groups, with $m = \frac{n}{n_1}$ samples each.

Step 2: Estimating within groups: for any $k \in [n_1]$, within every group $\{T(X_{k,j}), S_{k,j}\}_{j=1}^m$, we can derive an m-dimension vector $\boldsymbol{\eta}_{s,k}(\tilde{T}(X_{k,\cdot})^*) = \big(\hat{\mathbb{P}}_k(S = j^*|\tilde{T}(X_{k,1})), \ldots, \hat{\mathbb{P}}_k(S = j^*|\tilde{T}(X_{k,m}))\big)$ and $\hat{\pi}_k = \|\boldsymbol{\eta}_{s,k}(T(X)^*)\|_\infty$, which is defined in Lemma 4.4. Then, applying a straight-forward plug in $\hat{C}_{1,k} = \frac{\hat{\pi}_k+|\mathcal{D}|-2}{|\mathcal{D}|\hat{\pi}_k-1}$.

Step 3:Averaging: Finally, our estimator for C_1, denoted by $\hat{C}_1 = \frac{1}{n_1}\sum_{k=1}^{n_1} \hat{C}_{1,k}$, can be derived by averaging $\hat{C}_{1,k}, k \in [n_1]$.

Next, we state two assumptions that we used to derive the generalization error bound for Risk-LDP (Eq. (7)) when the noise rate is estimated from the data.

Assumption A: (Sub-exponentiality) For all $k \in [n_1]$, define $\hat{g}_k(\tilde{T}(X)) = \hat{\mathbb{P}}_k(S = j^*|\tilde{T}(X))$ There exists a constant $M_g > 0$, such that $\|\hat{C}_{1,k}\|_{\psi_1} = \|\min_{i\in[m]} \frac{\hat{g}_k(\tilde{T}(X_{k,i}))+|\mathcal{D}|-2}{|\mathcal{D}|\hat{g}_k(\tilde{T}(X_{k,i}))-1}\|_{\psi_1} \leq M_g$ 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 $\left|\mathbb{E}[\hat{C}_{1,k}] - C_1\right| < \theta$ for all $k \in [m]$, 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

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

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:

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

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

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

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

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

Third, Assume the events

$$A_1 = \left\{ \left| \hat{C}_1 - C_1 \right| \leq \tilde{\epsilon} \right\},$$

$$A_2 = \left\{ |\hat{\mathcal{R}}^{LDP}(f) - \mathcal{R}(f)| \leq \epsilon, \epsilon \geq \sqrt{\frac{\ln\left(\frac{\delta}{2}\right)}{2n}} \frac{2\hat{C}_1 \phi(M)|\mathcal{D}|}{\mathbb{P}(S = k^*)}, n \geq \frac{8 \ln\left(\frac{|\mathcal{D}|}{\delta}\right)}{\min_k \mathbb{P}(S = k)} \right\},$$

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

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 $\delta \in (0, \frac{1}{3}), \epsilon \geq \sqrt{\frac{\ln\left(\frac{\delta}{2}\right)}{2n}} \frac{(2(C_1 + \tilde{\epsilon})\phi(M)|\mathcal{D}|}{\mathbb{P}(S = k^*)}$, if $n \geq \frac{8 \ln\left(\frac{|\mathcal{D}|}{\delta}\right)}{\min_k \mathbb{P}(S = k)}$, then

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

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

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

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 $n \geq \frac{8 \ln\left(\frac{|\mathcal{D}|}{\delta}\right)}{\min_k \mathbb{P}(S = k)}$, then with probability $1 - 3\delta$ we have:

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

This completes the proof. □