Optimal Accuracy-Privacy Trade-off of Inference as Service

Yulu Jin and Lifeng Lai

***Abstract*—In this paper, we propose a general framework to provide a desirable trade-off between inference accuracy and privacy protection in the inference as service scenario (IAS). Instead of sending data directly to the server, the user will preprocess the data through a privacy-preserving mapping, which will increase privacy protection but reduce inference accuracy. To properly address the trade-off between privacy protection and inference accuracy, we formulate an optimization problem to find the privacy-preserving mapping. Even though the problem is non-convex in general, we characterize nice structures of the problem and develop an iterative algorithm to find the desired privacy-preserving mapping, with convergence analysis provided under certain assumptions. From numerical examples, we observe that the proposed method has better performance than gradient ascent method in the convergence speed, solution quality and algorithm stability.**

I. Introduction

The Internet of Things (IoT) is an emerging communication paradigm that aims at connecting different kinds of devices to the Internet [2]–[4]. Within the past decade, the number of IoT devices being introduced in the market has increased dramatically due to its low cost and convenience [5]. Sensors of IoT devices could generate contexts at a high velocity and the inference with the contexts becomes an essential component for IoT applications [6]. However, building inference systems is costly due to the overhead of maintaining contexts repositories, running inference algorithms and learning from the inference results for further applications of inference tasks. One of the emerging solutions to this problem is so-called inference-as-a-service (IAS) [7], [8]. In IAS, the devices will send data to a server in the cloud, who will make inference using sophisticated algorithms. However, the IAS paradigm brings privacy issues, as the devices will send their data to the cloud without knowing where these data is stored or what future purposes these data might serve. There are some interesting works that attempt to address this issue using Homomorphic Encryption (HE) technique [9]–[11]. Unfortunately, the complexity of HE-based solution is very high, and its privacy relies on the (unproved) assumption that certain mathematical problems are difficult to solve.

The goal of our paper is to address the fundamental trade-off between inference accuracy and privacy protection from information theory perspective. Instead of sending data directly to the server, the user will preprocess the data through a privacy-preserving mapping. This privacy-preserving mapping has two opposing effects. On one hand, it will prevent the server from observing the data directly and hence enhance the privacy protection. On the other hand, this might reduce the inference accuracy. To properly address the trade-off between these two competing goals, we formulate an optimization problem to find the privacy-preserving mapping. As the inference accuracy is directly related to the mutual information between parameters of interest and post-mapping data, we use mutual information to measure the inference accuracy. However, determining the privacy measure is tricky, as there are many existing information leakage measures [12], each of which is useful for certain specific scenarios. Hence, in our problem formulation, instead of using a specific privacy leakage measure, we propose a general framework that is applicable for different privacy metrics. The proposed framework is defined by a continuous function f with certain properties. Different choices of f lead to different privacy measures. For example, if f is chosen to be $-\log$ function, the proposed privacy leakage metric is the same as mutual information, a widely used information leakage measure. Moreover, we introduce a parameter β to represent the relative weight between these two measures. Thus, the trade-off problem between privacy and accuracy can be solved through a maximization problem where the objective function is composed of a weighted sum of accuracy and privacy terms.

To solve the maximization problem, if we optimize over the space of the privacy-preserving mapping directly, the formulated problem is a complicated non-concave problem with multiple constraints. Through various transformations and variable augmentations, we transform the optimization problem into a form that has three dominating arguments with certain nice concavity properties. In particular, if any two arguments are fixed, the problem is concave in the remaining argument. We then exploit this structure and design an algorithm with two nested loops to solve the optimization problem for general f by iterating between those three dominating arguments until reaching convergence. For the outer loop, we solve the optimization on the first dominating argument, for which we have a closed-form update formula. For the inner loop, using certain concavity properties of the objective function on the other two dominating arguments, we apply the Alternating Direction Method of Multipliers (ADMM) to solve the non-convex problem efficiently. Compared with solving the optimization problem using gradient ascent in the space of the privacy-preserving mapping directly, the proposed

Y. Jin and L. Lai are with the Department of Electrical and Computer Engineering, University of California, Davis, CA. Email: {yuljin,lflai}@ucdavis.edu. The work of Y. Jin and L. Lai was supported by the National Science Foundation under Grants CCF-1717943, ECCS-1711468, CNS-1824553 and CCF-1908258. This paper has been presented in part in the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing [1].

method does not need parameter tuning, converges much faster and finds solutions that have much better qualities. To further illustrate the proposed framework and algorithm, we also provide several examples by specializing f to particular function choices and provide numerical results.

Moreover, we provide the convergence analysis of the proposed method. Since there are two nested loops in the proposed algorithm, we first prove the convergence of the inner loop, which is the convergence proof of the ADMM procedure. Although there are many existing convergence proofs for typical ADMM, most of them focus on separable problems only. In our case, the considered optimization problem has non-separable structure. Inspired by recent research works about convergence analysis of ADMM with non-separable objective functions [13]–[15], we provide two proofs with different assumptions on f. Based on the convergence proof of ADMM, we further prove that the function value is non-decreasing between two iterations of the outer-loop. Then with a guarantee that the objective function is upper-bounded, the proposed algorithm is shown to converge.

There exist many other privacy-preserving techniques that are based on perturbations of data, which provide privacy guarantees at the expense of a loss of accuracy [16]–[20]. k-anonymity is proposed by Samarati and Sweeney [16], which requires that each record is indistinguishable from at least k–1 other records within the dataset. Differential privacy works by adding a pre-determined amount of randomness into a computation performed on a data set [17]. For example, a local randomization approach is proposed in [20] to solve the privacy concern in distributed machine learning whose privacy-preserving property is measured by local differential privacy, and ADMM is used as a parallel computing approach. These concepts and techniques are very useful for the privacy protection of data analysis through a dataset or database, which is different from the setup considered in this paper. Moreover, various minimax formulations and algorithms have also been proposed to defend against inference attacks in different scenarios [21]–[23]. Bertran et al. [21] proposed an optimization problem where the terms in the objective function were defined in terms of mutual information, showed the performance bound for the optimization problem and learned the sanitization transform in a data-driven fashion using an adversarial approach with Deep Neural Networks (DNNs). Under their formulation, they analyzed a trade-off between utility loss and attribute obfuscation under the constraint of the attribute obfuscation $I(A;Z) \leq k$. Feutry et al. [22] measured the utility and privacy by expected risks, formulated the utility-privacy trade-off as a min-diff-max optimization problem and proposed a learning-based and task-dependent approach to solving this problem, while only deterministic mechanisms are considered. To address this issue, a privacy-preserving adversarial network was proposed in [23] by employing adversarially-trained neural networks to implement randomized mechanisms and to perform a variational approximation of mutual information privacy. Different from them, we propose a more general framework of privacy protection and avoid the reliance on DNNs to derive the privacy-preserving mapping.

This journal paper is an extension of conference paper [1]. Compared with [1], the algorithms in this paper are significantly improved with theoretical convergence guarantee. In particular, while the algorithm in our conference paper [1] converges in various numerical examples, there is no convergence proof. It is in fact difficult to provide theoretical convergence guarantee for the algorithm in [1]. In this journal paper, by designing an improved iterative algorithm with two nested loops involving ADMM procedure, we can provide convergence guarantee both theoretically and numerically.

The remainder of the paper is organized as follows. In Section II, we introduce the problem formulation. In Section III, we present the proposed algorithm and provide the convergence analysis. In Section IV, we present numerical results. Finally, we offer concluding remarks and future work directions in Section V.

II. Problem formulation

Consider an inference problem, in which one would like to infer the parameter $S \in \mathcal{S}$ of data $Y \in \mathcal{Y}$, in which $\mathcal{Y}$ has a finite alphabet. In the inference as service scenario, one would send Y to the server who will determine the parameter S using its sophisticated models and powerful computing capabilities. However, directly sending data Y to the server brings the privacy issue, as now the server knows Y perfectly. To reduce the privacy leakage, instead of sending Y directly, one can employ a privacy-preserving mapping to transform data Y to $U \in \mathcal{U}$ and send U to the server. Here, $\mathcal{U}$ also has a finite alphabet and is allowed to be different from $\mathcal{Y}$. Without loss of generality, we will employ a randomized privacy-preserving mapping and use $p(u|y)$ to denote the probability that data $Y = y$ will be mapped to $U = u$ and the whole mapping is denoted as $P_{U|Y}$. Furthermore, we use P_S to denote the prior distribution of S and $P_{Y|S}$ to denote the conditional distribution Y given S, while the lower-case letter p is used to denote the component-wise probability (e.g., $p(s), p(y), p(y|s)$ will be used in the sequel).

To measure the inference accuracy, note that the distributional difference between P_S and $P_{S|U}$ characterizes the information about S contained in U. Since the inference at the server side is solely based on U, such information determines the inference accuracy. As $I(S;U)$ is the averaged Kullback–Leibler (KL) divergence between P_S and $P_{S|U}$, we use it to measure the inference accuracy. We would like to make $I(S;U)$ as large as possible, which means that we would like to retain as much information about the parameter of interest S in U as possible so that the server can make a more accurate inference.

To measure the privacy leakage, instead of choosing one particular privacy metric, we intend to investigate a general form $\mathbb{E}_{Y,U}[d(y,u)]$ that is applicable for different privacy metrics. Here, $d(y,u) = f(\frac{p(y)}{p(y|u)})$ and f is a continuous function defined on $(0, +\infty)$. We note that $\mathbb{E}_{Y,U}[d(y,u)] = \mathbb{E}_{Y,U}[f(\frac{p(y)}{p(y|u)})]$ measures the distributional distance between

P_Y and $P_{Y|U}$, where P_Y is the prior distribution of Y and $P_{Y|U}$ is the posterior distribution of Y after observing U. Hence, the smaller the distance, the less information U can provide about Y and the better the privacy protection. Note that $\frac{p(y)}{p(y|u)} = \frac{p(u)}{p(u|y)}$. Hence we will also use $\frac{p(u)}{p(u|y)}$ as the argument to f in the sequel. Since $p(u|y)$ shows in the denominator, we assume that $\epsilon \leq p(u|y) \leq 1, \forall y, u$, where $\epsilon > 0$.

To balance the inference accuracy and privacy protection, we propose to find the privacy-preserving mapping $P_{U|Y}$ by solving the following optimization problem

$$\max_{P_{U|Y}} \quad \mathcal{F}[P_{U|Y}] \triangleq I(S;U) - \beta \mathbb{E}_{Y,U}\left[f\left(\frac{p(y)}{p(y|u)}\right)\right], \quad (1)$$

$$\text{s.t.} \quad p(u|y) \geq \epsilon, \forall y, u,$$

$$\sum_u p(u|y) = 1, \forall y. \quad (2)$$

Here, $\beta \in (0, \infty)$ is a trade-off parameter that indicates the relative importance of maximizing $I(S;U)$ (i.e., maximizing inference accuracy) and minimizing the distance $\mathbb{E}_{Y,U}[d(y,u)]$ between P_Y and $P_{Y|U}$ (i.e., maximizing the privacy).

Another possible problem formulation is to maximize the inference accuracy under the constraint that the privacy leakage is less than certain threshold δ:

$$\max_{P_{U|Y}} \quad I(S;U) \quad (3)$$

$$\text{s.t.} \quad \mathbb{E}_{Y,U}\left[f\left(\frac{p(y)}{p(y|u)}\right)\right] \leq \delta,$$

$$p(u|y) \geq \epsilon, \forall y, u,$$

$$\sum_u p(u|y) = 1, \forall y.$$

However, directly solving such constrained optimization problems is very challenging. A typical way to solve this kind of problems is to form the Lagrangian of the maximization problem, whose objective is written as the weighted sum of the original objective and the constraints. Hence, our problem formulation can be viewed as the Lagrangian of the problem formulation (3). The trade-off parameter β can be treated as the Lagrangian multiplier. Different value of β corresponds to different privacy constraint δ in (3), whose value depends on different applications. In particular, using the proposed algorithms, solutions can be computed for a broad range of β. We can then obtain the Pareto optimal curve for accuracy and privacy leakage, where each point corresponds to one sub-problem solved to maximize the inference performance subject to a certain upper bound of privacy leakage. Then the user can select an operating point from the Pareto optimal curve depending on the user's preference and the constraint imposed by the applications.

For the privacy measure function f, we assume that

(a) $f(\cdot)$ is strictly convex;

(b) $f(\cdot)$ is twice-differentiable;

(c) $f^{'}(t)$ is l_f- Lipschitz continuous of t.

Here we provide some comments about these assumptions. (a) guarantees certain convexity of the problem. In particular, under (a), the sub-problems are shown to be convex, which ensures the feasibility and simplification of the proposed method. (b) and (c) are needed to ensure the convergence of the proposed method. These assumptions are fairly weak. As will be discussed in Section IV, most of the widely used distance measures satisfy these assumptions.

The proposed framework in (1) is very general. Different choices of f will lead to different privacy measures. For example, if we choose f to be $-\log(\cdot)$, then we have

$$\mathbb{E}_{Y,U}[d(y,u)] = -\sum_{y,u} p(y)p(u|y)\log\left(\frac{p(u)}{p(u|y)}\right)$$

$$= \sum_y p(y) D_{KL}[P_{U|y} \parallel P_U] = I[U;Y],$$

in which $D_{KL}(\cdot \parallel \cdot)$ is the KL divergence. As the result, choosing f to be the $-\log$ function means we will use mutual information between U and Y to measure the information leakage, a very common choice in information theory study. More examples will be provided in Section IV.

III. Algorithms and Convergence Proof

In this section, we discuss how to solve the optimization problem defined in (1) for general f. One natural approach to solving (1) is to apply the gradient ascent (GA) algorithm. However, GA faces several challenges such as proper step size, computation complexity, convergence speed and the quality of the optimal point found etc. To overcome these challenges, we propose a new algorithm that transforms the maximization problem over single argument to an alternative maximization problem over multiple arguments and then employ ideas from ADMM to solve the transformed problem.

A. Algorithm

We first have the following lemma that are useful for transforming the objective function.

Lemma 1:

$$I(S;U) = I(S;Y) - \sum_{u,y} p(y)p(u|y) D_{KL}[P_{S|y} \parallel P_{S|u}].$$

Proof: Please refer to Appendix A. ∎

By Lemma 1, the objective function defined in (1) can be written as

$$\mathcal{F}[P_{U|Y}, P_U, P_{S|U}] = I(S;Y) - \beta \mathbb{E}_{Y,U}[d(y,u)] - \sum_{u,y} p(y)p(u|y) D_{KL}[P_{S|y} \parallel P_{S|u}].$$

Note that $I(S;Y)$, $p(y)$ and $p(s|y)$ are fixed, hence the cost function can be viewed as a function of three arguments $P_{U|Y}$, P_U and $P_{S|U}$. For consistency, we require the following equations to be satisfied simultaneously

$$p(u) = \sum_y p(u|y)p(y), \forall u, \quad (4)$$

$$p(s|u) = \frac{\sum_y p(u|y)p(s,y)}{p(u)}, \forall u, \forall s. \quad (5)$$

By (5), we further require that $p(u) > 0, \forall u$. As the result, we can reformulate (1) as the following alternative optimization problem

$$\begin{aligned}\max_{P_{S|U},P_U,P_{U|Y}} \quad & \mathcal{F}[P_{U|Y}, P_U, P_{S|U}]. \qquad (6)\\ \text{s.t.} \quad & p(u|y) \geq \epsilon, \forall y, \forall u, \ \sum_u p(u|y) = 1, \forall y,\\ & p(u) > 0, \forall u, \ \sum_u p(u) = 1,\\ & p(u) = \sum_y p(u|y)p(y), \forall u,\\ & p(s|u) \geq 0, \forall u, \forall s, \ \sum_s p(s|u) = 1, \forall u,\\ & p(s|u) = \frac{\sum_y p(u|y)p(s,y)}{p(u)}, \forall u, \forall s.\end{aligned}$$

The following lemma illustrates the nice property of the alternative formulation (6): the alternative optimization problem is convex in each argument given the other two arguments.

Lemma 2: Suppose that $f(\cdot)$ is a strictly convex function. Then for given $P_U, P_{S|U}$, $\mathcal{F}[P_{U|Y}, P_U, P_{S|U}]$ is concave in each $P_{U|y_i}, \forall y_i \in \mathcal{Y}$. Similarly, for given $P_{U|Y}, P_{S|U}$, $\mathcal{F}[P_{U|Y}, P_U, P_{S|U}]$ is concave in P_U. For given $P_{U|Y}, P_U$, $\mathcal{F}[P_{U|Y}, P_U, P_{S|U}]$ is concave in $P_{S|U}$.

Proof: Please refer to Appendix B. ∎

Using this lemma, a natural approach to maximizing the objective function in (6) is to alternately iterate between $P_{U|Y}$, P_U and $P_{S|U}$ until reaching convergence. In particular, we propose an iterative algorithm with two blocks to obtain a solution to (6): update of $P_{S|U}$ and update of $P_{U|Y}, P_U$. Firstly, for a given P_U and $P_{U|Y}$, we update $P_{S|U}$ by solving the maximization on $P_{S|U}$ and derive an analytical result as a function of P_U and $P_{U|Y}$. Secondly, for the derived $P_{S|U}$, we update P_U and $P_{U|Y}$ by using the ADMM scheme to solve the maximization on P_U and $P_{U|Y}$. In this paper, we show that the proposed algorithm will converge. We would like to note that, however, as the problem in (6) is non-convex in the product space of $\{P_{U|Y}, P_U, P_{S|U}\}$, the derived limit point is not expected to be the global optimal solution of (6). In the following, we provide details for each iteration. The convergence proof of the proposed algorithm will be presented in Section III-B.

1) Updating $P_{S|U}$: For the $P_{S|U}$ subproblem, the maximization problem is

$$\begin{aligned}\max_{P_{S|U}} \quad & \mathcal{F}[P_{S|U}|P_{U|Y}, P_U],\\ \text{s.t.} \quad & p(s|u) \geq 0, \forall u, \forall s, \qquad (7)\\ & \sum_s p(s|u) = 1, \forall u, \qquad (8)\\ & p(s|u) = \frac{\sum_y p(u|y)p(s,y)}{p(u)}, \forall u, \forall s. \qquad (9)\end{aligned}$$

Lemma 3: The solution to the $P_{S|U}$ subproblem is

$$p(s|u) = \frac{\sum_y p(u|y)p(s,y)}{p(u)}. \qquad (10)$$

Proof: Please refer to Appendix C. ∎

2) Updating $P_{U|Y}$ and P_U: Now, for a given $P_{S|U}$, we discuss how to update $P_{U|Y}$ and P_U by solving

$$\max_{P_{U|Y}\in\mathcal{P}_{U|Y}, P_U\in\mathcal{P}_U} \ \mathcal{F}[P_{U|Y}, P_U|P_{S|U}], \qquad (11)$$

$$\text{s.t. } \delta(u) = p(u) - \sum_y p(u|y)p(y) = 0, \forall u, \qquad (12)$$

where

$$\mathcal{P}_{U|Y} = \{P_{U|Y} : p(u|y) \geq \epsilon, \sum_u p(u|y) = 1\}, \qquad (13)$$

$$\mathcal{P}_U = \{P_U : p(u) > 0, \sum_u p(u) = 1\}, \qquad (14)$$

and (12) corresponds to the consistency requirement (4).

Moreover, note that each row in the matrix $P_{U|Y}$ is independent and we further show that the objective function in (11) can be written as the sum of $|\mathcal{Y}|$ terms, each of which depends only on one row of $P_{U|Y}$.

$$\begin{aligned}\mathcal{F}[P_{U|Y}, P_U|P_{S|U}] &= -\beta \sum_{i=1}^{|\mathcal{Y}|} \left[p(y_i) \sum_u p(u|y_i) d\left(\frac{p(u)}{p(u|y_i)}\right)\right]\\ &\quad - \sum_{i=1}^{|\mathcal{Y}|} \left[p(y_i) \sum_u p(u|y_i) D_{KL}[P_{S|y_i} \parallel P_{S|u}]\right] + I(S;Y)\\ &= \sum_{i=1}^{|\mathcal{Y}|} \mathcal{F}_i'\left[P_{U|Y}, P_U|P_{S|U}\right] + I(S;Y), \qquad (15)\end{aligned}$$

where

$$\begin{aligned}\mathcal{F}_i'\left[P_{U|Y}, P_U|P_{S|U}\right] &= p(y_i)\left[-\beta \sum_u p(u|y_i) f\left(\frac{p(u)}{p(u|y_i)}\right)\right.\\ &\quad \left. - \sum_u p(u|y_i) D_{KL}[P_{S|y_i} \parallel P_{S|u}]\right]. \qquad (16)\end{aligned}$$

Thus, the optimization on $P_{U|Y}$ can be divided into $|\mathcal{Y}|$-problems, each of which corresponds to one row in $P_{U|Y}$.

As the result, although (11) is a non-convex problem in $(P_{U|Y}, P_U)$ jointly, it is a convex problem of one argument given the others, as shown in Lemma 2. This motivates us to apply the ADMM approach to solve the problem.

The augmented Lagrangian for the above problem is

$$\begin{aligned}&\mathcal{L}[P_{U|Y}, P_U, P_{S|U}; \Lambda]\\ &= \mathcal{F}[P_{U|Y}, P_U|P_{S|U}] + \sum_u \lambda(u)\delta(u) - \frac{\rho}{2}\sum_u \delta^2(u), \qquad (17)\end{aligned}$$

where Λ is a vector of size $|\mathcal{U}|$ and each component is denoted as $\lambda(u)$. Since $P_{S|U}$ is given, we will omit it from the expression of $\mathcal{L}$.

In the ADMM approach, there are updates of $P_{U|Y}$, P_U and Λ respectively. Exploiting the structure in (15), we can solve (11) using the following iterative procedure

$$P_{U|y_i}^{t+1} = \arg \max_{P_{U|y_i} \in \mathcal{P}_{U|y_i}} \mathcal{L}[P_{U|y_i}, P_{U|Y^{(i-)}}^{t+1}, P_{U|Y^{(i+)}}^{t}, P_U^t; \Lambda^t], \quad i = 1, 2, \cdots, |\mathcal{Y}|, \tag{18}$$

$$P_U^{t+1} = \arg \max_{P_U \in \mathcal{P}_U} \mathcal{L}[P_{U|Y}^{t+1}, P_U; \Lambda^t], \tag{19}$$

$$\Lambda^{t+1} = \Lambda^t - \rho(P_U^{t+1} - (P_{U|Y}^{t+1})^T P_Y), \tag{20}$$

or $$\lambda^{t+1}(u) = \lambda^t(u) - \rho[p^{t+1}(u) - \sum_y p^{t+1}(u|y)p(y)] = \lambda^t(u) - \rho\delta^{t+1}(u),$$

where $\mathcal{P}_{U|y_i} = \{P_{U|y_i} : p(u|y) \geq \epsilon, \sum_u p(u|y_i) = 1\}$, $P_{U|Y^{(i-)}}$ denotes all rows before the i-th row in the matrix $P_{U|Y}$ and $P_{U|Y^{(i+)}}$ denotes all rows after the i-th row. Note that here we use Gauss–Seidel ADMM where the local variables are updated sequentially in the Gauss–Seidel order and current conditional distributions ($P_{U|Y^{i-}}^{t+1}$ and $P_{U|Y^{i+}}^{t}$) are used to obtain $P_{U|y_i}^{t+1}$. Another update approach is to use $P_{U|Y^{i-}}^{t}$ to update $P_{U|y_i}$ in the $(t+1)$-th iteration. It has been shown that for multi-block problems, Gauss–Seidel ADMM often performs numerically better in practice than the directly extended ADMM [24]–[28], as the updated information $P_{U|Y^{i-}}^{t+1}$ is immediately utilized.

For $P_{U|y_i}$, the optimization problem is

$$\begin{aligned} \max_{P_{U|y_i}} \quad & \mathcal{L}[P_{U|y_i}, P_{U|Y^{(i-)}}^{t+1}, P_{U|Y^{(i+)}}^{t}, P_U^t; \Lambda^t], \\ \text{s.t.} \quad & p(u|y_i) \geq \epsilon, \forall u, \sum_u p(u|y_i) = 1. \end{aligned} \tag{21}$$

We have the following lemma regarding the objective function in (21). The proof follows similar steps to the proof of Lemma 2.

Lemma 4: The objective function in (21) is a strictly concave function.

Proof: Please refer to Appendix D. ∎

Hence, each sub-problem is a convex optimization problem with $|\mathcal{U}|$ inequality constraints and one equality constraint. In practice, under a specified $f(\cdot)$, the sub-problem can be solved numerically.

The sub-problem with respect to P_U is

$$\begin{aligned} \max_{P_U} \quad & \mathcal{L}[P_{U|Y}^{t+1}, P_U; \Lambda^t], \\ \text{s.t.} \quad & p(u) > 0, \forall u, \ \sum_u p(u) = 1. \end{aligned} \tag{22}$$

Following similar steps of Lemma 2, we can prove the following lemma.

Lemma 5: The objective function in (22) is a strictly concave function.

Proof: Please refer to Appendix D. ∎

Although there is a constraint, $P_U \in \mathcal{P}_U$, in this sub-problem, we can ignore it first and in the convergence proof, we will show that for the limit point, the constraint is naturally satisfied. We represent the solution to the unconstrained problem as $P_U^{t+1} = \arg\max_{P_U} \mathcal{L}[P_{U|Y}^{t+1}, P_U; \Lambda^t]$.

After solving two sub-problems on $P_{U|Y}$ and P_U respectively, we update the value of Λ.

In summary, we employ two nested loops to find the privacy-preserving mapping. In the outer loop, there are two update steps: update of $P_{S|U}$ and update of $(P_{U|Y}, P_U)$, where the update of $(P_{U|Y}, P_U)$ is performed by ADMM (which will be referred to as the inner loop). In the inner loop, we update $P_{U|Y}$ and P_U by going through the process of (18), (19), (20). We will use (j) to denote the j-th outer iteration and use $(j), t$ to denote the arguments at the t-th inner iteration of the j-th outer iteration. The algorithm is summarized in Algorithm 1. To quantify the matrix differences, we use the Frobenius norm [29], where for an $m \times n$ matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^m \sum_{j=1}^n |a_{i,j}|^2}$. To quantify the vector differences, we use the ℓ_2 norm, where for vector $\boldsymbol{b} = (b_1, b_2, \cdots, b_n)$, $\|\boldsymbol{b}\|_2^2 = \sum_{i=1}^n b_i^2$. For the thresholds, η is chosen to be a small value such that the function value is converged and η_p is chosen to be a small value such that $\mathcal{L}[P_{U|Y}^{t+1}, P_U^{t+1}; \Lambda^t] \geq \mathcal{L}[P_{U|Y}^{t+1}, P_U^t; \Lambda^t] \geq \mathcal{L}[P_{U|Y}^t, P_U^t, \Lambda^t]$ is true.

Algorithm 1 Design the privacy-preserving mapping

Input:
Prior distribution P_S and conditional distribution $P_{Y|S}$.
Trade-off parameter β.
Converge parameter η, η_p, η_d.
Output:
A mapping $P_{U|Y}$ from $Y \in \mathcal{Y}$ to $U \in \mathcal{U}$.
Initialization:
Randomly initiate $P_{U|Y}$ and calculate $P_U, P_{S|U}$ by (4) and (5).

$j = 1$.
while $\left\| P_{S|U}^{(j)} - P_{S|U}^{(j-1)} \right\|_F > \eta$ **do**
  $P_U^{(j),1} = P_U^{(j-1)}$.
  $P_{U|Y}^{(j),1} = P_{U|Y}^{(j-1)}$.
  $t = 1$.
  while $t = 1$ or $\left\| P_U^{(j),t} - P_U^{(j),t-1} \right\|_2^2 > \eta_p$ **do**
    Update $P_{U|y_i}$ by solving (21).
    Update P_U by solving (22).
    Update Λ by (20).
    $t = t + 1$.
  Update $P_{S|U}^{(i)}$ by (10).
  $j = j + 1$.
return $P_{U|Y}$

B. Convergence Analysis

In this section, we provide the convergence proof for Algorithm 1. To prove the convergence of the proposed iterative algorithm, we need to verify that the value of the functional $\mathcal{F}$ does not decrease while iterating, and that this functional is bounded from above.

The following lemma shows that $\mathcal{F}$ is upper-bounded.

Lemma 6: For a continuous function $f(\cdot)$, $\mathcal{F}[P_{U|Y}, P_U, P_{S|U}]$ is bounded from above.

Proof: Please refer to Appendix E. ∎

Then we prove that the value of $\mathcal{F}$ is non-decreasing between two iterations of the outer loop. There are two steps in the outer loop, updating $P_{S|U}$ by (10), and updating $(P_{U|Y}, P_U)$ by applying ADMM. For the update of $P_{S|U}$, since the optimization with respect to $P_{S|U}$ is a convex optimization problem and has a closed-form solution as the update function, the objective function $\mathcal{F}$ is non-decreasing in this step. To show that the value of $\mathcal{F}$ is non-decreasing for the limit point found by ADMM, it is necessary to prove that the proposed ADMM procedure converges subsequently. Otherwise, the consistency requirement between P_U and $P_{U|Y}$ may not be satisfied. In particular, in the following we prove that any sequence generated by the proposed ADMM procedure is bounded and has a limit point that is also the stationary point of (11), and the value of $\mathcal{F}$ is upper-bounded and non-decreasing between iterations of ADMM.

We note that the convergence proof of the proposed ADMM procedure for our problem setup is non-trivial, as the considered objective function has more than 2 local variables and is non-separable with respect to these local variables. Directly using multi-block ADMM may be non-convergent, even if the functions are separable with respect to these blocks of variables [30], and numerous research efforts have been devoted to analyzing the convergence of multi-block ADMM under certain assumptions [26], [27], [31]. In contrast to the separable case, studies on the convergence properties of n-block ADMM with non-separable objective, even for $n = 2$, are limited [15], [32], and the convergence is not guaranteed and has to be handled differently.

To make the presentation clear, in the following, we consider the case $|\mathcal{Y}| = 2$ and the proof can be easily generalized to the case when $\mathcal{Y}$ has a finite alphabet. For $|\mathcal{Y}| = 2$, the optimization problem in (1) can be further represented as

$$\begin{aligned}
\max \quad & -\Bigg[p(y_1)\sum_u p\left(u|y_1\right) D_{KL}[P_{S|y_1} \parallel P_{S|u}] \\
& +p(y_2)\sum_u p\left(u|y_2\right) D_{KL}[P_{S|y_2} \parallel P_{S|u}]\Bigg] \\
& -\beta\sum_u \left[p\left(u|y_1\right)p(y_1) f\left(\frac{p(u)}{p(u|y_1)}\right)\right. \\
& \left. +p(u|y_2)p(y_2) f\left(\frac{p(u)}{p(u|y_2)}\right)\right], \\
\text{s. t} \quad & p(u|y_i) \geq \epsilon, \forall u, \sum_u p(u|y_i) = 1, i = 1, 2, \\
& p(u) > 0, \forall u, \sum_u p(u) = 1, \\
& -p(u|y_1)p(y_1) - p(u|y_2)p(y_2) + p(u) = 0, \forall u,
\end{aligned}$$

in which the last constraint can also be written in the vector form, $-p(y_1)P_{U|y_1} - p(y_2)P_{U|y_2} + P_U = \mathbf{0}$.

For presentation convenience, we denote

$$h_i(P_{U|y_i}) = -p(y_i)\sum_u p(u|y_i) D_{KL}[P_{S|y_i} \parallel P_{S|u}], \quad i = 1, 2,$$

$$\begin{aligned}
g(P_{U|y_1}, P_{U|y_2}, P_U) = -\beta\sum_u \Bigg[& p(u|y_1)p(y_1) f\left(\frac{p(u)}{p(u|y_1)}\right) \\
& +p(u|y_2)p(y_2) f\left(\frac{p(u)}{p(u|y_2)}\right)\Bigg].
\end{aligned}$$

Thus, the objective function is

$$h_1(P_{U|y_1}) + h_2(P_{U|y_2}) + g(P_{U|y_1}, P_{U|y_2}, P_U),$$

and the augmented Lagrangian is

$$\begin{aligned}
& \mathcal{L}[P_{U|Y}, P_U, P_{S|U}; \Lambda] \\
= \; & \mathcal{F}[P_{U|Y}, P_U|P_{S|U}] + \sum_u \lambda(u)\delta(u) - \frac{\rho}{2}\sum_u \delta(u)^2 \\
= \; & h_1(P_{U|y_1}) + h_2(P_{U|y_2}) + g(P_{U|y_1}, P_{U|y_2}, P_U) \\
& + \sum_u \lambda(u)\delta(u) - \frac{\rho}{2}\sum_u \delta(u)^2.
\end{aligned}$$

For the update of the dual variable Λ, we have the following lemma which characterizes the relationship between the dual variable Λ and the primal variables.

Lemma 7: Suppose that $f(\cdot)$ is twice-differentiable and $f^{'}(t)$ is l_f- Lipschitz continuous of t. We have

$$\begin{aligned}
\left\|\Lambda^{t+1} - \Lambda^t\right\|_2^2 \leq & l_\Lambda \left(\left\|P_{U|y_1}^{t+1} - P_{U|y_1}^t\right\|_2^2 + \left\|P_{U|y_2}^{t+1}\right.\right. \\
& \left.\left. - P_{U|y_2}^t\right\|_2^2 + \left\|P_U^{t+1} - P_U^t\right\|_2^2\right),
\end{aligned} \tag{23}$$

with $l_\Lambda = \frac{16\beta^2 l_f^2}{\epsilon^4}$.

Proof: Please refer to Appendix F. ∎

For the ascent of $\mathcal{L}$ between two iterations, we have the following lemma.

Lemma 8: Suppose that $f(\cdot)$ is twice-differentiable and $f^{'}(t)$ is l_f- Lipschitz continuous of t. We have

$$\begin{aligned}
& \mathcal{L}\left[P_{U|Y}^{t+1}, P_U^{t+1}; \Lambda^{t+1}\right] - \mathcal{L}\left[P_{U|Y}^t, P_U^t; \Lambda^t\right] \\
\geq \; & \left[\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}\right]\|P_{U|y_1}^{t+1} - P_{U|y_1}^t\|_2^2 \\
& + \left[\frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}\right]\left\|P_{U|y_2}^{t+1} - P_{U|y_2}^t\right\|_2^2 \\
& + \left(\frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\right)\left\|P_U^{t+1} - P_U^t\right\|_2^2,
\end{aligned} \tag{24}$$

where $l_{y_1} = l_{y_2} = \frac{\beta l_f}{\epsilon^3}, l_u = \frac{\beta l_f}{\epsilon}$.

Proof: Please refer to Appendix G. ∎

With these supporting results, we now analyze the convergence of the proposed ADMM procedure. We first show that $\mathcal{L}$ is monotonic and upper-bounded, and the sequence $\{P_{U|Y}, P_U, \Lambda\}^t$ generated by ADMM is bounded.

Proposition 1: Suppose that $f(\cdot)$ is twice-differentiable and $f^{'}(t)$ is l_f-Lipschitz continuous of t. We have that

1) if $\min\{\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\} \geq 0$, $\mathcal{L}[P_{U|Y}^{t+1}, P_U^{t+1}; \Lambda^{t+1}] \geq \mathcal{L}[P_{U|Y}^t, P_U^t, \Lambda^t]$;
2) $\forall t \in \mathbb{N}$, $\mathcal{L}[P_{U|Y}^t, P_U^t; \Lambda^t]$ is upper-bounded;
3) $\{P_{U|Y}, P_U, \Lambda\}^t$ is bounded.

Proof: Please refer to Appendix H. ∎

We then show the asymptotic regularity of the sequence $\{P_{U|Y}, P_U, \Lambda\}^t$.

Proposition 2: Suppose that $f(\cdot)$ is twice-differentiable and $f^{'}(t)$ is l_f- Lipschitz continuous of t. When ρ is sufficiently large such that $\min\{\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\} \geq 0$, as $t \to \infty$, we have

1) $\left\| P_{U|y_1}^{t+1} - P_{U|y_1}^t \right\|_2^2 \to 0$,
2) $\left\| P_{U|y_2}^{t+1} - P_{U|y_2}^t \right\|_2^2 \to 0$,
3) $\left\| P_U^{t+1} - P_U^t \right\|_2^2 \to 0$.
4) $\left\| \Lambda^{t+1} - \Lambda^t \right\|_2^2 \to 0$,
5) $P_U^{t+1} - p(y_1) P_{U|y_1}^{t+1} - p(y_2) P_{U|y_2}^{t+1} \to 0$.

Proof: Please refer to Appendix I. ∎

Proposition 3: The sequence $\{P_{U|Y}, P_U, \Lambda\}^t$ has a limit point $(\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda})$, which is also a stationary point of (11).

Proof: Please refer to Appendix J. ∎

We now summarize the convergence results in the following theorem.

Theorem 1: Suppose that $f(\cdot)$ is twice-differentiable and $f^{'}(t)$ is l_f-Lipschitz continuous of t. Choose ρ such that $\min\{\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\} \geq 0$. The proposed ADMM procedure could converge subsequently, that is, staring from any $(P_{U|Y}^0, P_U^0, \Lambda^0)$, it generates a sequence that is bounded, has a limit point $(\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda})$, and the limit point is a stationary point of (11).

Proof: Please refer to Appendix K. ∎

Therefore, for the limit point $(\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda})$, the value of $\mathcal{F}$ is non-decreasing after the ADMM procedure. Then $\mathcal{F}$ is also non-decreasing between two iterations of the outer loop, which indicates that the proposed algorithm will converge.

For the case $|Y| = k$, there will be $(k+1)$ terms on the right hand side of (23) and (24). Then Propositions 1, 2, 3 and Theorem 1 still hold in a similar manner and the convergence analysis also applies.

C. Stronger Convergence for f with More Assumptions

In Section III-B, for the convergence analysis of ADMM, the value of ρ should be chosen large enough such that $\min\{\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\} \geq 0$. Thus, the feasible set of ρ will depend on the choice of ϵ. In this subsection, we propose another ADMM procedure with Bregman distance and make stronger assumptions on f to provide a convergence analysis with weaker constraints on ρ.

First we introduce the definition of Bregman distance. Let $\phi : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable and strictly convex function. Denote $\nabla\phi(y)$ as the gradient of ϕ on y. Then the Bregman distance induced by ϕ is defined as

$$\Delta_\phi(x, y) = \phi(x) - \phi(y) - \langle \nabla\phi(y), x - y \rangle, \tag{25}$$

where ϕ is called the kernel function or distance-generating function. From the property of Bregman distance, we have that $\Delta_\phi(x, y)$ is convex in x for fixed y [33]. The Bregman distance plays an important role in iterative algorithms. In particular, Bregman divergences are used to replace the quadratic penalty term in the standard ADMM (see $\delta^2(u)$ in (17)). Then we can choose a suitable Bregman divergence so that the subproblems can be solved more efficiently [33].

To solve the optimization problem in (11), for notation simplicity, we denote $x_1 : P_{U|y_1}$, $x_2 : P_{U|y_2}$, and $v : P_U$.

Recall the definition of $h_1(\cdot), h_2(\cdot), g(\cdot)$ in Section III-B. We propose an algorithm starting with $\left(x_1^0, x_2^0, v^0\right)$ and Λ^0. Suppose that $\varphi_1, \varphi_2, \phi$ are differentiable and strictly convex functions. Then with the given iteration point $w^k = \left(x_1^k, x_2^k, v^k, \Lambda^k\right)$, the new iteration point $w^{k+1} = \left(x_1^{k+1}, x_2^{k+1}, v^{k+1}, \Lambda^{k+1}\right)$ is given as:

$$\begin{aligned}
x_1^{k+1} &= \arg\max \Big\{ h_1(x_1) + \left(x_1 - x_1^k\right)^T \nabla_{x_1} g\left(x_1^k, x_2^k, v^k\right) \\
&\quad - \frac{\rho}{2} \left\| p(y_1)x_1 + p(y_2)x_2^k - v^k - \frac{\Lambda^k}{\rho} \right\|_2^2 - \Delta_{\varphi_1}\left(x_1, x_1^k\right) \Big\}, \\
x_2^{k+1} &= \arg\max \Big\{ h_2(x_2) + \left(x_2 - x_2^k\right)^T \nabla_{x_2} g\left(x_1^k, x_2^k, v^k\right) \\
&\quad - \frac{\rho}{2} \left\| p(y_1)x_1^{k+1} + p(y_2)x_2 - v^k - \frac{\Lambda^k}{\rho} \right\|_2^2 - \Delta_{\varphi_2}\left(x_2, x_2^k\right) \Big\}, \\
v^{k+1} &= \arg\max \Big\{ g\left(x_1^{k+1}, x_2^{k+1}, v\right) \\
&\quad - \frac{\rho}{2} \left\| p(y_1)x_1^{k+1} + p(y_2)x_2^{k+1} - v - \frac{\Lambda^k}{\rho} \right\|_2^2 - \Delta_\phi\left(v, v^k\right) \Big\}, \\
\Lambda^{k+1} &= \Lambda^k - \rho\left(p(y_1)x_1^{k+1} + p(y_2)x_2^{k+1} - v^{k+1}\right),
\end{aligned} \tag{26}$$

where $\Delta_{\varphi_1}\left(x_1, x_1^k\right), \Delta_{\varphi_2}\left(x_2, x_2^k\right)$, and $\Delta_\phi\left(v, v^k\right)$ are the Bregman distances associated with φ_1, φ_2, and ϕ respectively. Here, φ_1, φ_2, and ϕ should be properly chosen with respect to different $f(\cdot)$ adopted in the privacy measure.

To guarantee that the algorithm converges, we assume that

(i) ∇g is l_g-Lipschitz continuous;
(ii) $\nabla\varphi_1, \nabla\varphi_2, \nabla\phi$ are Lipshitz continuous with the modulus $l_{\varphi_1}, l_{\varphi_2}, l_\phi$, respectively;
(iii) $\varphi_1, \varphi_2, \phi$ are strongly convex with the modulus $\delta_{\varphi_1}, \delta_{\varphi_2}, \delta_\phi$, and $\delta_{\varphi_1}, \delta_{\phi_2} > l_g$.

Then we have

Lemma 9:

$$\begin{aligned}
&\left\| \Lambda^{k+1} - \Lambda^k \right\|_2^2 \\
&\leq 3l_g^2 \left(\left\| x_1^{k+1} - x_1^k \right\|_2^2 + \left\| x_2^{k+1} - x_2^k \right\|_2^2 \right) \\
&\quad + 3\left(l_g^2 + l_\phi^2\right) \left\| v^{k+1} - v^k \right\|_2^2 + 3l_\phi^2 \left\| v^k - v^{k-1} \right\|_2^2.
\end{aligned} \tag{27}$$

Proof: Please refer to Appendix L. ∎

By considering the updates of 3 primal variables, we have

Lemma 10:

$$\begin{aligned}
&\left(\mathcal{L}\left(w^{k+1}\right)-\frac{3l_\phi^2}{\rho}\left\|v^{k+1}-v^k\right\|_2^2\right)\\
&\quad-\left(\mathcal{L}\left(w^{k}\right)-\frac{3l_\phi^2}{\rho}\left\|v^{k}-v^{k-1}\right\|_2^2\right)\\
\geq&\left(\frac{\delta_{\varphi_1}-l_g}{2}-\frac{3l_g^2}{\rho}\right)\left\|x_1^{k+1}-x_1^k\right\|_2^2\\
&+\left(\frac{\delta_{\varphi_2}-l_g}{2}-\frac{3l_g^2}{\rho}\right)\left\|x_2^{k+1}-x_2^k\right\|_2^2\\
&+\left(\frac{\delta_{\phi}}{2}-\frac{3l_g^2+6l_\phi^2}{\rho}\right)\left\|v^{k+1}-v^k\right\|_2^2.
\end{aligned}$$

Proof: Please refer to Appendix M. ∎

Proposition 4: Under assumptions (i), (ii), (iii), we have

1) if $\rho \geq \max\{\frac{6l_g^2}{\delta_{\varphi_1}-l_g}, \frac{6l_g^2}{\delta_{\varphi_2}-l_g}, \frac{6l_g^2+12l_\phi^2}{\delta_\phi}\}$ (feasible under assumption (iii)), $\left(\mathcal{L}\left(w^{k+1}\right)-\frac{3l_\phi^2}{\rho}\left\|v^{k+1}-v^k\right\|_2^2\right)$ $-\left(\mathcal{L}\left(w^{k}\right)-\frac{3l_\phi^2}{\rho}\left\|v^{k}-v^{k-1}\right\|_2^2\right)\geq 0$;
2) $\forall k\in\mathbb{N}$, $\mathcal{L}[w^k]$ is upper-bounded;
3) $\{w\}^k$ is bounded.

Then following similar analysis in Section III-B, when ρ is chosen properly such that $\geq \max\{\frac{6l_g^2}{\delta_{\varphi_1}-l_g}, \frac{6l_g^2}{\delta_{\varphi_2}-l_g}, \frac{6l_g^2+12l_\phi^2}{\delta_\phi}\}$, we have $\left\|x_1^{k+1}-x_1^k\right\|_2^2 \to 0$, $\left\|x_2^{k+1}-x_2^k\right\|_2^2 \to 0$, and $\left\|v^{k+1}-v^k\right\|_2^2 \to 0$. By Lemma 9, we have $\left\|\Lambda^{k+1}-\Lambda^k\right\|_2^2 \to 0$. Moreover, the limit point of $\{w\}^k$ can also be shown to be the stationary point of (11). Thus, when replacing the ADMM procedure in Section III-A with this ADMM procedure with Bregman distance, Algorithm 1 converges in a similar manner.

IV. Examples and Numerical Results

In this section, we first give examples of different choices of f and then provide numerical results with specific f to show the performance of the proposed method.

A. Examples of f

We now provide examples of f, each of which leads to a well-known and widely used divergence measure.

In the first example, we consider $f(t)=-\log(t)$. As shown in Section II, if $f(t)=-\log(t)$, the privacy measure is then the mutual information. For the algorithm proposed in this chapter, we check whether all the assumptions are satisfied. Since $\epsilon \leq p(u|y) \leq 1$, we have $\epsilon \leq \frac{p(u)}{p(u|y)} \leq \frac{1}{\epsilon}$. Then we first have that $-\log(\cdot)$ is strictly convex on $\left[\epsilon, \frac{1}{\epsilon}\right]$. Secondly, we have that $f'(t)=-\frac{1}{t}$ is Lipschitz continuous since it is everywhere differentiable on $\left[\epsilon, \frac{1}{\epsilon}\right]$ and the absolute value of the derivative is bounded above by $\frac{1}{\epsilon^2}$.

In the second example, we consider the following strictly convex function $f(t) = t\log\frac{2t}{t+1}+\log\frac{2}{t+1}$. This choice leads to the Jensen-Shannon divergence [34]: $\mathbb{E}_{Y,U}[d(y,u)] = \sum_y p(y) JS[P_{U|y}, P_U]$, in which $JS[P_{U|y}, P_U] = D_{KL}\left[P_{U|y} \parallel \frac{P_{U|y}+P_U}{2}\right] + D_{KL}\left[P_U \parallel \frac{P_{U|y}+P_U}{2}\right]$

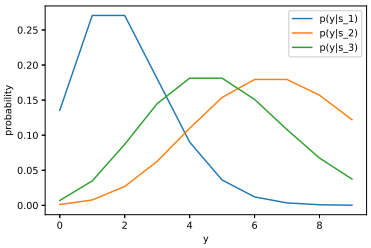

Fig. 1. Conditional distribution $p(y|s)$

To check the assumption (b), we have $f'(t) = \log\frac{2t}{t+1}$, $f''(t) = \frac{1}{t(t+1)} \leq \frac{1}{\epsilon(\epsilon+1)}$, and thus it is Lipschitz continuous.

In the third example, consider the strictly convex function $f(t) = (1-t)^2/(2t+2)$, which leads to the Le Cam divergence [35] as the privacy measure, $\mathbb{E}_{Y,U}[d(y,u)] = \sum_y p(y) LC[P_{U|y} \parallel P_U]$, in which

$$LC[P_{U|y} \parallel P_U] = \frac{1}{2}\sum_u \frac{[p(u)-p(u|y)]^2}{p(u|y)+p(u)}. \tag{28}$$

For this choice of f, again, assumptions (b) and (c) are satisfied.

In the fourth example, we consider the following function $f(t)=(1-\sqrt{t})^2$, which corresponds to the squared Hellinger distance [36]. It is easy to check that the assumptions are satisfied.

B. Numerical results

In this subsection, we provide numerical examples to show that our methods converge much faster than GA, and the solution found by our methods has much better quality than the one found by GA. Moreover, we explore how the weight parameter β and the alphabet size of $\mathcal{U}$ affects the privacy protection as well as the inference accuracy.

In the first example, we set the prior distribution $P_S = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$ and let $|\mathcal{Y}| = 10, |\mathcal{U}| = 12$. The conditional distributions $P_{Y|S}$ under each s are shown in Fig. 1. Under this setup, we will perform both Algorithm 1 and GA to find the transition mapping $P_{U|Y}$ that maximizes the functional defined in (1). Suppose that the trade-off parameter $\beta = 2$ and Jensen-Shannon divergence is used as the privacy metric. The initial mapping $P_{U|Y}$ is obtained by selecting random numbers conforming to uniform distribution and normalizing them.

For the convergence speed, we investigate the relationship between $\mathcal{F}$ and the outer iteration, which is illustrated in Fig 2. We notice that the function value is increasing and converges as the iterative process progresses. For comparison purposes, we also plot the corresponding figures for GA in Fig. 3 (with step size 0.0001) and Fig. 4 (with step size 0.00005). From these figures, we can see that Algorithm 1 converges within 20 iterations. On the other hand, for gradient ascent algorithm, even for a pretty small step size 0.0001, the function value

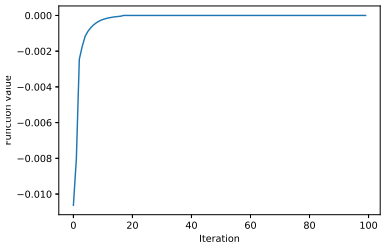

Fig. 2. Function value v.s. iteration (Algorithm 1)

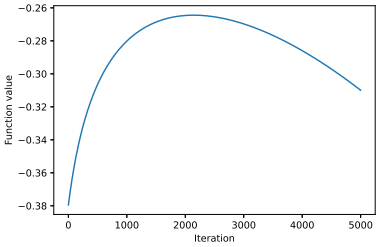

Fig. 3. Function value v.s. iteration (GA)

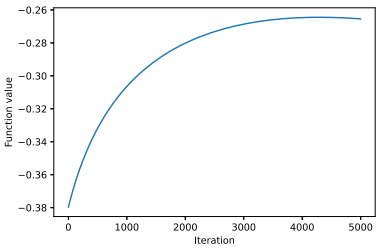

Fig. 4. Function value v.s. iteration (GA)

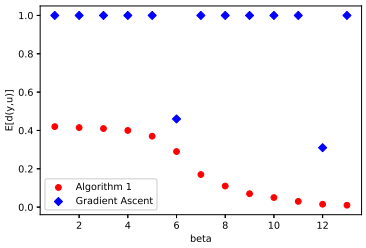

Fig. 5. β v.s. privacy protection (Algorithm 1 and GA)

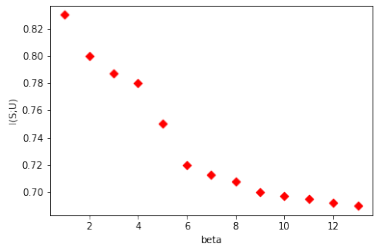

Fig. 6. β v.s. inference accuracy (Algorithm 1)

fails to keep increasing, which indicates that the step size is too large. Then for a smaller step size 0.00005, the function value converges as shown in Fig. 4. However, the value of the objective function found by GA is smaller than the value found by Algorithm 1.

For the relationship between β and the privacy protection, after random initialization, we run Algorithm 1 and GA until they terminate. The stopping criterion is either $||P_{U|Y}^{t+1} - P_{U|Y}^{t}||_F < 10^{-5}$ (convergence case) or a maximum number of iterations is reached (divergence case). We repeat this procedure 100 times for each β. Recall that the smaller the term $\mathbb{E}[d(y,u)]$, the better the privacy protection. In particular, we set $\mathbb{E}[d(y,u)]$ to be 1 for divergence cases since the maximum $\mathbb{E}[d(y,u)]$ under the converge scenario is smaller than 1. As shown in Fig. 5, we notice that $\mathbb{E}[d(y,u)]$ decreases as β increases for our proposed method while it is non-decreasing for GA. By setting the maximum number of iterations to be 3000, GA diverges under many choices of β. Even for the scenarios where GA converges, compared with Algorithm 1, the privacy protection obtained by GA is weaker. Therefore, the privacy-preserving mapping designed by GA could hardly guarantee the protection of privacy. In addition, we also explore the relationship between β and the information accuracy. As shown in Fig. 6, the inference accuracy measure $I(S;U)$ decreases as β increases, which indicates that the predictive ability becomes weaker. The reason is that as U leaks less information about Y when β increases, it also provides less information about the parameter of interest, which will reduce the predictive performance. However, Fig. 6 shows that the reduction of $I(S;U)$ is not very large, which implies that the model still has good predictive ability when there are stronger protections for privacy.

To explore other privacy measures, we now set f as $f(t) = (1-t)^2/(2t+2)$, which corresponds to the Le Cam divergence as discussed in Section IV-A. We again compare Algorithm 1 and GA. The results are shown in Table I. From the table, we can see that the maximum function value found by our method is greater than those found by GA.

Methods	**Convergent value**
Algorithm 1	-6.697e-14
Gradient ascent($\alpha = 0.05$)	-0.251
Gradient ascent($\alpha = 0.07$)	-0.245
Gradient ascent($\alpha = 0.1$)	-0.317
Gradient ascent($\alpha = 0.15$)	-0.235
Gradient ascent($\alpha = 0.2$)	Diverge

TABLE I
CONVERGENT VALUE OF ALGORITHM 1 AND GA

To compare different privacy measures, we set the trade-off parameter $\beta = 8$, which indicates that the privacy term is dominant in the objective function. As shown in Fig. 7, although the function values under JS-divergence and LC-divergence are different, the convergence speed and convergence curve

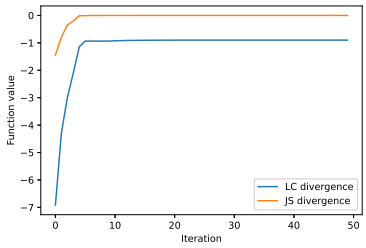

Fig. 7. Convergence process for JS and LC divergences (Algorithm 1)

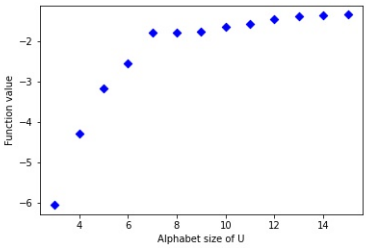

Fig. 8. Function value v.s. Alphabet size of $\mathcal{U}$ (Algorithm 1)

are almost the same, which shows that the proposed algorithm can converge in a similar manner under different metrics. However, the optimal privacy-preserving mapping $P_{U|Y}$ found by those two privacy measures are different. Therefore, in practical applications, an appropriate task-oriented privacy measure needs to be chosen.

Finally, we explore the relationship between $|\mathcal{U}|$ and the privacy protection. Note that in the proposed method, the alphabet sizes of $\mathcal{Y}$ and $\mathcal{U}$ are not necessarily equal. Thus, for $|\mathcal{Y}| = 10$, we explore how $|\mathcal{U}|$ affects the convergent function value. Here, we set $\beta = 8$ and use the LC-divergence to measure the privacy leakage. From Fig. 8, it is shown that although the function value is increasing as $|\mathcal{U}|$ increases, the alphabet size $|\mathcal{U}|$ has limited effects on the function value when $|\mathcal{U}| \geq 7$, which indicates that a large alphabet size of $\mathcal{U}$ is not necessary to derive a satisfactory privacy-preserving mapping. By setting $|\mathcal{Y}|$ to different values, we notice that when $\frac{|\mathcal{U}|}{|\mathcal{Y}|} \geq 0.8$, the convergent function value is relatively large.

V. Conclusion

We have proposed a general framework to design privacy-preserving mapping to achieve privacy-accuracy trade-off in the IAS scenarios. We have formulated optimization problems to find the desirable mapping. We have discussed the structure of the formulated problems and designed an iterative method to solve these complicated optimization problems. We have also proved the convergence of the proposed method under certain assumptions. Moreover, we have provided numerical results showing that this method has better performance than GA in the convergence speed, solution quality and algorithm stability.

In terms of future work, we will address the limitations of the currently work along the following lines. Firstly, we have several technical assumptions on the function f. In the future, we will try to weaken those assumptions. Secondly, for the proposed algorithm, we are only able to show the convergence, but we have not characterized the convergence rate, of the proposed algorithms. Moreover, the proposed method is only guaranteed to converge but not to the global optima. Thus, it is of interest to further modify the proposed method to find the global optimal solution and determine the corresponding convergence rate. Thirdly, we are also interested in comparing our proposed privacy protection scheme with other existing private mechanisms. Finally, in this work, we only consider the case when Y is discrete and generate the privacy-preserving mapping $P_{U|Y}$. In the future, we will consider the continuous case and find the optimal conditional pdf $f_{U|Y}$.

Appendix A
Proof of Lemma 1

$$
\begin{aligned}
& I(S;U) + \sum_{u,y} p(y)p(u|y)D_{KL}[P_{S|y} \parallel P_{S|u}] \\
= & \sum_{s,u,y} p(s,u,y) \log \frac{p(s|u)}{p(s)} \\
& + \sum_{s,u,y} p(y)p(u|y)p(s|y) \log \frac{p(s|y)}{p(s|u)} \\
\overset{(a)}{=} & \sum_{s,u,y} p(s,u,y) \left[\log \frac{p(s|u)}{p(s)} + \log \frac{p(s|y)}{p(s|u)} \right] \\
= & \sum_{s,y} p(s,y) \log \frac{p(s|y)}{p(s)} = I(S;Y),
\end{aligned}
$$

where (a) uses the fact that $S \to Y \to U$ is a Markov chain since given Y, S and U are independent.

Appendix B
Proof of Lemma 2

First, prove that $\mathcal{F}[P_{U|Y}]$ is concave with respect to $P_{S|U}$. By applying Lemma 1, (1) can be written in the following form,

$$
\begin{aligned}
\mathcal{F}[P_{U|Y}] = & \; I(S;Y) - \beta \mathbb{E}_{Y,U}[d(y,u)] \\
& - \sum_{u,y} p(y)p(u|y)D_{KL}[P_{S|y} \parallel P_{S|u}]. \qquad (29)
\end{aligned}
$$

Note that $I(S;Y)$ is a constant under our setup. Given $P_{U|Y}$ and P_U, $\mathbb{E}_{Y,U}[d(y,u)]$ is independent of $P_{S|U}$. Moreover, $P_{S|u}$ and $P_{S|u'}$ are two independent vectors. For given u and y, we have

$$
D_{KL}[P_{S|y} \parallel P_{S|u}] = \sum_{s} p(s|y) \log \frac{p(s|y)}{p(s|u)}. \qquad (30)
$$

Since $a \log(x)$ is concave in x, (30) is convex in $P_{S|u}$ and $\mathcal{F}[P_{U|Y}]$ is concave with respect to $P_{S|U}$.

Second, we prove that $\mathcal{F}[P_{U|Y}]$ is concave w.r.t P_U when f is strictly convex. Note that P_U only shows up in $\mathbb{E}_{Y,U}[d(y,u)]$

and since f is strictly convex, taking the sum doesn't change the concavity and $\mathcal{F}[P_{U|Y}]$ is also concave in P_U.

Third, we consider $P_{U|Y}$. There are $|\mathcal{Y}|$ conditional distributions in the mapping $P_{U|Y}$, where $P_{U|y}$ and $P_{U|y'}$ are independent when $y \neq y'$. Then we consider a particular row $P_{U|y}$ and prove the concavity. The Hessian matrix of $\mathcal{F}$ with respect to $P_{U|y}$ is

$$\boldsymbol{H}_{\mathcal{F}} = \begin{bmatrix} \frac{\partial^2 \mathcal{F}[p(u|y)]}{\partial p(u_1|y)^2} & \cdots & \frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_1|y)\partial p(u_{|\mathcal{U}|}|y)} \\ \cdots & \cdots & \cdots \\ \frac{\partial^2 \mathcal{F}[p(u|y)]}{\partial p(u_{|\mathcal{U}|}|y)\partial p(u_1|y)} & \cdots & \frac{\partial^2 \mathcal{F}[p(u|y)]}{\partial p(u_{|\mathcal{U}|}|y)^2} \end{bmatrix}.$$

Then we calculate each element in $\boldsymbol{H}_{\mathcal{F}}$. Assume that $i \neq j$. Taking derivative based on the form in (29), we have

$$\frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_i|y)^2} = -\beta \frac{\partial^2 \mathbb{E}_{Y,U}[d(y,u)]}{\partial p(u_i|y)^2},$$
$$\frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_i|y)\partial p(u_j|y)} = -\beta \frac{\partial^2 \mathbb{E}_{Y,U}[d(y,u)]}{\partial p(u_i|y)\partial p(u_j|y)},$$

in which

$$\frac{\partial^2 \mathbb{E}_{Y,U}[d(y,u)]}{\partial p(u_i|y)^2} = p(y)\left[f'(t)\frac{-p(u_i)}{p(u_i|y)^2} - f'(t)\frac{-p(u_i)}{p(u_i|y)^2} - tf''(t)\frac{-p(u_i)}{p(u_i|y)^2}\right] = p(y)f''(t)\frac{t^2}{p(u_i|y)} > 0,$$
$$\frac{\partial^2 \mathbb{E}_{Y,U}[d(y,u)]}{\partial p(u_i|y)\partial p(u_j|y)} \overset{(a)}{=} 0,$$

where $t = \frac{p(u_i)}{p(u_i|y)}$ and (a) is due to the fact that t is independent of $p(u_j|y)$ when $i \neq j$ and P_U is given. Then we have $\frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_i|y)^2} < 0$ and $\frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_i|y)\partial p(u_j|y)} = 0$. Thus, the Hessian matrix $\boldsymbol{H}_{\mathcal{F}}$ is a diagonal matrix with negative entries, which indicates that the objective function $\mathcal{F}$ is concave in $P_{U|y_i}$ and the lemma is proved.

Appendix C
Proof of Lemma 3

We first ignore (7), (9) and solve the optimization problem subject to (8) only. We will then check that the obtained solution satisfy constraints (7), (9).

For a $u \in \mathcal{U}$, the Lagrangian is

$$\mathcal{L}_{S|u} = \mathcal{F}[P_{S|U}|P_{U|Y}, P_U] + \alpha\left(\sum_s p(s|u) - 1\right),$$

where α is the Lagrangian multiplier with respect to constraint (8) . Since P_U and $P_{U|Y}$ are given, $\mathcal{L}_{S|u}$ is a convex function with respect to $P_{S|u}$. By taking the derivative, we have

$$\frac{\partial \mathcal{L}_{S|u}}{\partial p(s|u)} = \frac{\sum_y p(y)p(u|y)p(s|y)}{p(s|u)} + \alpha = 0,$$

which indicates

$$p(s|u) = \frac{\sum_y p(y)p(u|y)p(s|y)}{-\alpha}. \tag{31}$$

Since $\sum_s p(s|u) = 1$, we have

$$\sum_s p(s|u) = \sum_s \frac{\sum_y p(y)p(u|y)p(s|y)}{-\alpha} = 1$$
$$\implies \alpha = -\sum_s \sum_y p(y)p(u|y)p(s|y)$$
$$= -\sum_y p(y)p(u|y)\sum_s p(s|y)$$
$$= -\sum_y p(y)p(u|y) = -p(u).$$

Plugging the value of α into (31), we have

$$p(s|u) = \frac{\sum_y p(u|y)p(s,y)}{p(u)} \geq 0,$$

which guarantees the non-negativity condition in (7). It is also easy to check that this satisfies the constraint in (9) exactly, preserves the consistency of different arguments and thus is the solution to the $P_{S|U}$ subproblem.

Appendix D
Proof of Lemma 4 and 5

Note that for $i \neq j$,

$$\frac{\partial^2 \sum_{i=1}^m \lambda(u_i)\delta(u_i) - \frac{\rho}{2}\sum_{i=1}^m \delta(u_i)^2}{\partial p(u_i|y)^2} = -\rho p^2(y) \leq 0,$$
$$\frac{\partial^2 \sum_{i=1}^m \lambda(u_i)\delta(u_i) - \frac{\rho}{2}\sum_{i=1}^m \delta(u_i)^2}{\partial p(u_i|y)\partial p(u_j|y)} = 0.$$

In Lemma 2, we have shown that $\frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_i|y)^2} < 0$ and $\frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_i|y)\partial p(u_j|y)} = 0$. Hence, we have

$$\frac{\partial^2 \mathcal{L}[P_{U|Y}]}{\partial p(u_i|y)^2} = \frac{\partial^2 \mathcal{F}[P_{U|Y}]}{\partial p(u_i|y)^2} - \rho p^2(y) < 0,$$
$$\frac{\partial^2 \mathcal{L}[P_{U|Y}]}{\partial p(u_i|y)\partial p(u_j|y)} = 0,$$

and that the Hessian matrix $\boldsymbol{H}_{\mathcal{L}}$ is negative-definite. Moreover, the constraint $\sum_{i=1}^m p(u_i|y) = 1, \forall y \in \mathcal{Y}$ defines a convex set and thus the sub-problem on $P_{U|y_i}$ is a convex problem. Similarly, we also have $\frac{\partial^2 \mathcal{L}[P_U]}{\partial p(u_i)^2} < 0$ and $\frac{\partial^2 \mathcal{L}[P_U]}{\partial p(u_i)\partial p(u_j)} = 0$, which indicates that the Hessian matrix of $\mathcal{L}$ with respect to P_U is negative-definite. Combined with the fact that the constraint set is convex, the sub-problem on P_U is a convex optimization problem.

Appendix E
Proof of Lemma 6

First, note that $I(S;U) \leq H(S)$, which is bounded. Thus, $\mathcal{F}[P_{U|Y}]$ is upper bounded if $\mathbb{E}_{Y,U}[d(y,u)]$ is bounded from above. Let $t(y,u) = \frac{p(u)}{p(u|y)}$. We have that

$$\mathbb{E}_{Y,U}[d(y,u)] = \sum_{y,u} p(y)p(u|y)f(t(y,u))$$
$$= \sum_{y,u} p(y)p(u)\frac{f(t(y,u))}{t(y,u)},$$

where $p(y)p(u) \leq 1$. Since $\epsilon \leq p(u|y) \leq 1$, we have that $t(y,u) \in \left[\epsilon, \frac{1}{\epsilon}\right], \forall y, u$. Since f is continuous, it's natural to have $\frac{f(t(y,u))}{t(y,u)} < +\infty$. Then $\mathbb{E}_{Y,U}[d(y,u)]$ is bounded from above.

Appendix F
Proof of Lemma 7

By the optimality of P_U, we have

$$\begin{cases} 0 = \nabla_{P_U} g\left(P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^{t+1}\right) \\ \qquad - \rho\left(-p(y_1)P_{U|y_1}^{t+1} - p(y_2)P_{U|y_2}^{t+1} + P_U^{t+1}\right) + \Lambda^t, \\ \Lambda^{t+1} = \Lambda^t - \rho\left(-p(y_1)P_{U|y_1}^{t+1} - p(y_2)P_{U|y_2}^{t+1} + P_U^{t+1}\right), \end{cases}$$

which implies

$$0 = \nabla_{P_U} g\left(P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^{t+1}\right) + \Lambda^{t+1}. \tag{32}$$

Then we have

$$\begin{aligned} & \left\|\Lambda^{t+1} - \Lambda^t\right\|_2^2 \\ = & \left\|\nabla_{P_U} g\left(P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^{t+1}\right)\right. \\ & \left. -\nabla_{P_U} g\left(P_{U|y_1}^{t}, P_{U|y_2}^{t}, P_U^{t}\right)\right\|_2^2 \\ = & \sum_{u\in\mathcal{U}} \left(\frac{\partial g(P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^{t+1})}{\partial p^{t+1}(u)}\right. \\ & \left. -\frac{\partial g(P_{U|y_1}^{t}, P_{U|y_2}^{t}, P_U^{t})}{\partial p^{t}(u)}\right)^2. \end{aligned} \tag{33}$$

Given that $f^{'}(t)$ is l_f-Lipschitz continuous of t, we further have that for any given u,

$$\begin{aligned} & \left(\frac{\partial g(P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^{t+1})}{\partial p^{t+1}(u)} - \frac{\partial g(P_{U|y_1}^{t}, P_{U|y_2}^{t}, P_U^{t})}{\partial p^{t}(u)}\right)^2 \\ = & \left(\beta p(y_1)\left[f'\left(\frac{p^t(u)}{p^t(u|y_1)}\right) - f'\left(\frac{p^{t+1}(u)}{p^{t+1}(u|y_1)}\right)\right]\right. \\ & \left. +\beta p(y_2)\left[f'\left(\frac{p^t(u)}{p^t(u|y_2)}\right) - f'(\frac{p^{t+1}(u)}{p^{t+1}(u|y_2)})\right]\right)^2 \\ \leq & \beta^2 l_f^2 \left[p(y_1)\left|\frac{p^t(u)}{p^t(u|y_1)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_1)}\right|\right. \\ & \left. +p(y_2)\left|\frac{p^t(u)}{p^t(u|y_2)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_2)}\right|\right]^2 \\ \leq & 2\beta^2 l_f^2 \left[p(y_1)^2\left(\frac{p^t(u)}{p^t(u|y_1)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_1)}\right)^2\right. \\ & \left. +p(y_2)^2\left(\frac{p^t(u)}{p^t(u|y_2)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_2)}\right)^2\right], \end{aligned} \tag{34}$$

where $\frac{p^t(u)}{p^t(u|y_1)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_1)} = \frac{p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)}{p^t(u|y_1)p^{t+1}(u|y_1)}$. Using the assumption that $\frac{1}{p(u|y)} \leq \frac{1}{\epsilon} < \infty$, we have

$$\begin{aligned} & \left|\frac{p^t(u)}{p^t(u|y_1)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_1)}\right| \\ \leq & \left(\frac{1}{\epsilon}\right)^2 \left|p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)\right|. \end{aligned}$$

To further bound $\left|p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)\right|$, we have

$$\begin{aligned} & \left|\frac{p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)}{p^t(u) - p^{t+1}(u)}\right| \\ = & p^{t+1}(u|y_1) + p^{t+1}(u)\frac{|p^{t+1}(u|y_1) - p^t(u|y_1)|}{|p^t(u) - p^{t+1}(u)|} \\ \leq & 1 + \frac{|p^{t+1}(u|y_1) - p^t(u|y_1)|}{|p^t(u) - p^{t+1}(u)|}, \end{aligned}$$

and

$$\begin{aligned} & \left|\frac{p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)}{p^{t+1}(u|y_1) - p^t(u|y_1)}\right| \\ = & p^t(u) + p^t(u|y_1)\frac{|p^t(u) - p^{t+1}(u)|}{|p^{t+1}(u|y_1) - p^t(u|y_1)|} \\ \leq & 1 + \frac{|p^t(u) - p^{t+1}(u)|}{|p^{t+1}(u|y_1) - p^t(u|y_1)|}. \end{aligned}$$

Moreover, $\min\left\{\frac{|p^{t+1}(u|y_1) - p^t(u|y_1)|}{|p^t(u) - p^{t+1}(u)|}, \frac{|p^t(u) - p^{t+1}(u)|}{|p^{t+1}(u|y_1) - p^t(u|y_1)|}\right\} \leq 1$. Then we have

$$\begin{aligned} & \min\{\left|\frac{p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)}{p^t(u) - p^{t+1}(u)}\right|, \\ & \left|\frac{p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)}{p^{t+1}(u|y_1) - p^t(u|y_1)}\right|\} \leq 2, \end{aligned}$$

and thus

$$\begin{aligned} & |p^t(u)p^{t+1}(u|y_1) - p^{t+1}(u)p^t(u|y_1)| \\ \leq & 2|p^t(u) - p^{t+1}(u)| + 2|p^{t+1}(u|y_1) - p^t(u|y_1)|, \end{aligned}$$

and thus

$$\begin{aligned} & \left|\frac{p^t(u)}{p^t(u|y_1)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_1)}\right| \\ \leq & \frac{2}{\epsilon^2}\left[|p^t(u) - p^{t+1}(u)| + |p^{t+1}(u|y_1) - p^t(u|y_1)|\right]. \end{aligned} \tag{35}$$

Similarly, for $\left(\frac{p^t(u)}{p^t(u|y_2)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_2)}\right)$, we have

$$\begin{aligned} & \left|\frac{p^t(u)}{p^t(u|y_2)} - \frac{p^{t+1}(u)}{p^{t+1}(u|y_2)}\right| \\ \leq & \frac{2}{\epsilon^2}\left[|p^t(u) - p^{t+1}(u)| + |p^{t+1}(u|y_2) - p^t(u|y_2)|\right]. \end{aligned} \tag{36}$$

Plugging (35) and (36) into (34) and (33), we have

$$\begin{aligned} & \left\|\Lambda^{t+1} - \Lambda^t\right\|_2^2 \\ = & \sum_{u\in\mathcal{U}} \left(\frac{\partial g(P_{U|y_1}^{t+1}, P_{U|y_2}^{t+1}, P_U^{t+1})}{\partial p^{t+1}(u)}\right. \\ & \left. -\frac{\partial g(P_{U|y_1}^{t}, P_{U|y_2}^{t}, P_U^{t})}{\partial p^{t}(u)}\right)^2 \\ \leq & 2\beta^2 l_f^2 \left(\frac{2}{\epsilon^2}\right)^2 \sum_{u\in\mathcal{U}} \left\{p(y_1)^2[|p^t(u) - p^{t+1}(u)|\right. \\ & +|p^{t+1}(u|y_1) - p^t(u|y_1)|]^2 \\ & +p(y_2)^2[|p^t(u) - p^{t+1}(u)| \\ & \left. +|p^{t+1}(u|y_2) - p^t(u|y_2)|]^2\right\} \end{aligned}$$

$$
\begin{aligned}
&\leq 2\beta^2 l_f^2\left(\frac{2}{\epsilon^2}\right)^2\sum_{u\in\mathcal{U}}\big\{2p(y_1)^2[(p^t(u)-p^{t+1}(u))^2\\
&\quad+(p^{t+1}(u|y_1)-p^t(u|y_1))^2]\\
&\quad+2p(y_2)^2[(p^t(u)-p^{t+1}(u))^2\\
&\quad+(p^{t+1}(u|y_2)-p^t(u|y_2))^2]\big\}\\
&= 4\beta^2 l_f^2\left(\frac{2}{\epsilon^2}\right)^2\sum_{u\in\mathcal{U}}\big[p(y_1)^2(p^{t+1}(u|y_1)-p^t(u|y_1))^2\\
&\quad+p(y_2)^2(p^{t+1}(u|y_2)-p^t(u|y_2))^2\\
&\quad+(p(y_1)^2+p(y_2)^2)(p^t(u)-p^{t+1}(u))^2\big]\\
&\leq 4\beta^2 l_f^2\left(\frac{2}{\epsilon^2}\right)^2\sum_{u\in\mathcal{U}}\big[(p^{t+1}(u|y_1)-p^t(u|y_1))^2\\
&\quad+(p^{t+1}(u|y_2)-p^t(u|y_2))^2\\
&\quad+(p^t(u)-p^{t+1}(u))^2\big]\\
&= l_\Lambda\left(\left\|P^{t+1}_{U|y_1}-P^t_{U|y_1}\right\|_2^2+\left\|P^{t+1}_{U|y_2}-P^t_{U|y_2}\right\|_2^2\right.\\
&\quad\left.+\left\|P^{t+1}_U-P^t_U\right\|_2^2\right),
\end{aligned}
$$

where $l_\Lambda=4\beta^2l_f^2\left(\frac{2}{\epsilon^2}\right)^2=\frac{16\beta^2l_f^2}{\epsilon^4}$.

Appendix G
Proof of Lemma 8

Since $f'(t)$ is l_f-Lipschitz continuous and $t=\frac{p(u)}{p(u|y_i)}\in(0,\frac{1}{\epsilon})$, we have that g is differentiable and $\nabla_{P_{U|y_1}}g,\nabla_{P_{U|y_2}}g,\nabla_{P_U}g$ are Lipschitz continuous with constants l_{y_1}, l_{y_2}, l_u for $P_{U|y_1},P_{U|y_2},P_U$ respectively. In particular, we have $l_{y_1}=l_{y_2}=\frac{\beta l_f}{\epsilon^3}$ and $l_u=\frac{\beta l_f}{\epsilon}$. Then we have

$$
\begin{aligned}
&\mathcal{L}[P^{t+1}_{U|y_1},P^t_{U|y_2},P^t_U;\Lambda^t]-\mathcal{L}[P^t_{U|y_1},P^t_{U|y_2},P^t_U;\Lambda^t]\\
=\;&\mathcal{F}[P^{t+1}_{U|y_1},P^t_{U|y_2},P^t_U]-\mathcal{F}[P^t_{U|y_1},P^t_{U|y_2},P^t_U]\\
&+\langle\Lambda^t,p(y_1)(P^t_{U|y_1}-P^{t+1}_{U|y_1})\rangle\\
&+\frac{\rho}{2}\|P^t_U-p(y_1)P^t_{U|y_1}-p(y_2)P^t_{U|y_2}\|_2^2\\
&-\frac{\rho}{2}\|P^t_U-p(y_1)P^{t+1}_{U|y_1}-p(y_2)P^t_{U|y_2}\|_2^2\\
\overset{(a)}{=}\;&\mathcal{F}[P^{t+1}_{U|y_1},P^t_{U|y_2},P^t_U]-\mathcal{F}[P^t_{U|y_1},P^t_{U|y_2},P^t_U]\\
&+\langle\Lambda^t,p(y_1)(P^t_{U|y_1}-P^{t+1}_{U|y_1})\rangle\\
&+\langle\rho(P^t_U-p(y_1)P^{t+1}_{U|y_1}-p(y_2)P^t_{U|y_2}),\\
&\quad p(y_1)(P^{t+1}_{U|y_1}-P^t_{U|y_1})\rangle+\frac{\rho}{2}\|p(y_1)(P^t_{U|y_1}-P^{t+1}_{U|y_1})\|_2^2\\
=\;&\mathcal{F}[P^{t+1}_{U|y_1},P^t_{U|y_2},P^t_U]-\mathcal{F}[P^t_{U|y_1},P^t_{U|y_2},P^t_U]\\
&+\frac{\rho}{2}\|p(y_1)(P^t_{U|y_1}-P^{t+1}_{U|y_1})\|_2^2+\langle P^{t+1}_{U|y_1}-P^t_{U|y_1},\\
&-p(y_1)\Lambda^t+p(y_1)\rho(P^t_U-p(y_1)P^{t+1}_{U|y_1}-p(y_2)P^t_{U|y_2})\rangle\\
=\;&\mathcal{F}[P^{t+1}_{U|y_1},P^t_{U|y_2},P^t_U]-\mathcal{F}[P^t_{U|y_1},P^t_{U|y_2},P^t_U]\\
&+\frac{\rho}{2}\|p(y_1)(P^t_{U|y_1}-P^{t+1}_{U|y_1})\|_2^2-\langle P^{t+1}_{U|y_1}-P^t_{U|y_1},\\
&\nabla_{P_{U|y_1}}\mathcal{F}[P^{t+1}_{U|y_1},P^t_{U|y_2},P^t_U]\rangle\\
\overset{(b)}{\geq}\;&\left[\frac{\rho}{2}p(y_1)^2-\frac{l_{y_1}}{2}\right]\|P^{t+1}_{U|y_1}-P^t_{U|y_1}\|_2^2,
\end{aligned}
\tag{37}
$$

where (a) follows from the cosine rule and (b) follows from the fact that $\mathcal{F}=h_1+h_2+g$, $h_i(P_{U|y_i})$ is linear in $P_{U|y_i}$, and $\nabla_{P_{U|y_1}}g$ is l_{y_1}-Lipschitz continuous of $P_{U|y_1}$.

Similarly, for the update of $P_{U|y_2}$, we have

$$
\begin{aligned}
&\mathcal{L}[P^{t+1}_{U|y_1},P^{t+1}_{U|y_2},P^t_U;\Lambda^t]-\mathcal{L}[P^{t+1}_{U|y_1},P^t_{U|y_2},P^t_U;\Lambda^t]\\
\geq\;&\left[\frac{\rho}{2}p(y_2)^2-\frac{l_{y_1}}{2}\right]\|P^{t+1}_{U|y_1}-P^t_{U|y_1}\|_2^2.
\end{aligned}
\tag{38}
$$

For the update of P_U and Λ, we have

$$
\begin{aligned}
&\mathcal{L}[P^{t+1}_{U|y_1},P^{t+1}_{U|y_2},P^{t+1}_U;\Lambda^{t+1}]-\mathcal{L}[P^{t+1}_{U|y_1},P^{t+1}_{U|y_2},P^t_U;\Lambda^t]\\
=\;&g(P^{t+1}_{U|Y},P^{t+1}_U)-g(P^{t+1}_{U|Y},P^t_U)+\langle\Lambda^{t+1},P^{t+1}_U-P^t_U\rangle\\
&+\frac{\rho}{2}\|P^{t+1}_U-P^t_U\|_2^2-\frac{1}{\rho}\|\Lambda^{t+1}-\Lambda^t\|_2^2\\
\geq\;&\frac{\rho-l_u}{2}\|P^{t+1}_U-P^t_U\|_2^2-\frac{1}{\rho}\|\Lambda^{t+1}-\Lambda^t\|_2^2.
\end{aligned}
\tag{39}
$$

Combining (37), (38) and (39), we have

$$
\begin{aligned}
&\mathcal{L}[P^{t+1}_{U|Y},P^{t+1}_U;\Lambda^{t+1}]-\mathcal{L}[P^t_{U|Y},P^t_U;\Lambda^t]\\
\geq\;&\left[\frac{\rho}{2}p(y_1)^2-\frac{l_{y_1}}{2}\right]\|P^{t+1}_{U|y_1}-P^t_{U|y_1}\|_2^2\\
&+\left[\frac{\rho}{2}p(y_2)^2-\frac{l_{y_2}}{2}\right]\left\|P^{t+1}_{U|y_2}-P^t_{U|y_2}\right\|_2^2\\
&+\frac{\rho-l_u}{2}\left\|P^{t+1}_U-P^t_U\right\|_2^2-\frac{1}{\rho}\left\|\Lambda^{t+1}-\Lambda^t\right\|_2^2\\
\overset{(c)}{\geq}\;&\left[\frac{\rho}{2}p(y_1)^2-\frac{l_{y_1}}{2}-\frac{l_\Lambda}{\rho}\right]\|P^{t+1}_{U|y_1}-P^t_{U|y_1}\|_2^2\\
&+\left[\frac{\rho}{2}p(y_2)^2-\frac{l_{y_2}}{2}-\frac{l_\Lambda}{\rho}\right]\left\|P^{t+1}_{U|y_2}-P^t_{U|y_2}\right\|_2^2\\
&+\left(\frac{\rho-l_u}{2}-\frac{l_\Lambda}{\rho}\right)\left\|P^{t+1}_U-P^t_U\right\|_2^2,
\end{aligned}
$$

where (c) follows from Lemma 7.

Appendix H
Proof of Proposition 1

1) If $\min\{\frac{\rho}{2}p(y_1)^2-\frac{l_{y_1}}{2}-\frac{l_\Lambda}{\rho},\frac{\rho}{2}p(y_2)^2-\frac{l_{y_2}}{2}-\frac{l_\Lambda}{\rho},\frac{\rho-l_u}{2}-\frac{l_\Lambda}{\rho}\}\geq 0$, according to Lemma 8, we have

$$
\mathcal{L}[P^{t+1}_{U|Y},P^{t+1}_U;\Lambda^{t+1}]-\mathcal{L}[P^t_{U|Y},P^t_U,\Lambda^t]\geq 0.
$$

2) $\forall t\in\mathbb{N}$, $\mathcal{L}[P^t_{U|Y},P^t_U,P^t_{S|U};\Lambda^t]$ is upper-bounded. Assume that there exists P'_U, such that $P'_U-(P^t_{U|Y})^TP_Y=\mathbf{0}$. Then we have

$$
\begin{aligned}
&\mathcal{L}[P^t_{U|Y},P^t_U;\Lambda^t]\\
=\;&h_1(P^t_{U|y_1})+h_2(P^t_{U|y_2})+g(P^t_{U|y_1},P^t_{U|y_2},P^t_U)\\
&+\sum_u\lambda^t(u)\delta^t(u)-\frac{\rho}{2}\sum_u\delta^t(u)^2\\
=\;&h_1(P^t_{U|y_1})+h_2(P^t_{U|y_2})+g(P^t_{U|y_1},P^t_{U|y_2},P^t_U)\\
&+(\Lambda^t)^T[P^t_U-(P^t_{U|Y})^TP_Y]\\
&-\frac{\rho}{2}[P^t_U-(P^t_{U|Y})^TP_Y]^T[P^t_U-(P^t_{U|Y})^TP_Y]
\end{aligned}
$$

$$
\begin{aligned}
&\leq h_1(P^t_{U|y_1}) + h_2(P^t_{U|y_2}) + g(P^t_{U|y_1}, P^t_{U|y_2}, P^t_U) \\
&\quad +(\Lambda^t)^T[P^t_U - (P^t_{U|Y})^T P_Y] \\
&= h_1(P^t_{U|y_1}) + h_2(P^t_{U|y_2}) + g(P^t_{U|y_1}, P^t_{U|y_2}, P^t_U) \\
&\quad +\langle \Lambda^t, P^t_U - P'_U\rangle \\
&\overset{(a)}{=} h_1(P^t_{U|y_1}) + h_2(P^t_{U|y_2}) + g(P^t_{U|y_1}, P^t_{U|y_2}, P^t_U) \\
&\quad -\langle \nabla_{P_U} g\left(P^t_{U|y_1}, P^t_{U|y_2}, P^t_U\right), P^t_U - P'_U\rangle \\
&\overset{(b)}{\leq} h_1(P^t_{U|y_1}) + h_2(P^t_{U|y_2}) + g(P^t_{U|y_1}, P^t_{U|y_2}, P'_U) \\
&\quad +\frac{l_u}{2}\|P^t_U - P'_U\|_2^2 \\
&< \infty,
\end{aligned}
$$

where (a) follows from (32) and (b) is true as $\nabla_{P_U} g$ is l_u-Lipschitz continuous.

3) $\{P^t_{U|Y}, P^t_U, \Lambda^t\}$ is bounded.

Since $\forall t \in \mathbb{N}$, $P^t_{U|y_1}, P^t_{U|y_2}$ are PMFs, $\{P_{U|Y}\}^t$ is bounded. Similarly, $\{P_U\}^t$ is also bounded. For Λ^t, Lemma 7 can be generalized to the case where the iteration difference is k and we have

$$
\begin{aligned}
\left\|\Lambda^{t+k} - \Lambda^t\right\|_2^2 \leq\ & l_\Lambda\left(\left\|P^{t+k}_{U|y_1} - P^t_{U|y_1}\right\|_2^2\right. \\
&\left. + \left\|P^{t+k}_{U|y_2} - P^t_{U|y_2}\right\|_2^2 + \left\|P^{t+k}_U - P^t_U\right\|_2^2\right), \forall k \in \mathbb{N}^+.
\end{aligned}
$$

Thus, since $\{P_{U|Y}\}^t$ and $\{P_U\}^t$ are bounded, $\{\Lambda\}^t$ is also bounded.

Appendix I
Proof of Proposition 2

When ρ is sufficiently large, e.g. $\rho = \frac{7\beta l_f}{\epsilon^3 \min\{p(y_1)^2, p(y_2)^2\}}$, we will have $\min\{\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}, \frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\} \geq 0$. In this case, since $\mathcal{L}[P_{U|Y}, P_U; \Lambda]$ is non-decreasing between iterations and upper-bounded, there exists t_0, such that

$$
\begin{aligned}
\infty >\ & \sum_{t=t_0}^{\infty}\left|\mathcal{L}\left[P^t_{U|y_1}, P^t_{U|y_2}, P^t_U; \Lambda^t\right]\right. \\
&\qquad \left. -\mathcal{L}\left[P^{t+1}_{U|y_1}, P^{t+1}_{U|y_2}, P^{t+1}_U; \Lambda^{t+1}\right]\right| \\
\overset{(b)}{\geq}\ & \left[\frac{\rho}{2}p(y_1)^2 - \frac{l_{y_1}}{2} - \frac{l_\Lambda}{\rho}\right]\sum_{t=t_0}^{\infty}\|P^{t+1}_{U|y_1} - P^t_{U|y_1}\|_2^2 \\
&+\left[\frac{\rho}{2}p(y_2)^2 - \frac{l_{y_2}}{2} - \frac{l_\Lambda}{\rho}\right]\sum_{t=t_0}^{\infty}\left\|P^{t+1}_{U|y_2} - P^t_{U|y_2}\right\|_2^2 \\
&+\left(\frac{\rho - l_u}{2} - \frac{l_\Lambda}{\rho}\right)\sum_{t=t_0}^{\infty}\left\|P^{t+1}_U - P^t_U\right\|_2^2,
\end{aligned}
$$

where (b) is from Lemma 8. Then as $t \to \infty$, we have $\left\|P^{t+1}_{U|y_1} - P^t_{U|y_1}\right\|_2 \to 0$, $\left\|P^{t+1}_{U|y_2} - P^t_{U|y_2}\right\|_2 \to 0$, and $\left\|P^{t+1}_U - P^t_U\right\|_2 \to 0$. By Lemma 7, we have $\left\|\Lambda^{t+1} - \Lambda^t\right\|_2 \to 0$, which implies

$$
P^{t+1}_U - p(y_1) P^{t+1}_{U|y_1} - p(y_2) P^{t+1}_{U|y_2} \to 0.
$$

Appendix J
Proof of Proposition 3

Since $\{P^t_{U|Y}, P^t_U, \Lambda^t\}$ is bounded, there exists a subsequence $\{P^{t_s}_{U|Y}, P^{t_s}_U, \Lambda^{t_s}\}$ that converges to the limit point $(\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda})$, i.e. $\lim_{s\to\infty}(P^{t_s}_{U|Y}, P^{t_s}_U, \Lambda^{t_s}) = (\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda})$. For the limit point $(\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda})$, we will show that it is the stationary point of (11).

By the optimality of $P_{U|y_1}$, $P_{U|y_2}$ and P_U, we have

$$
\begin{aligned}
\mathbf{0} \in\ & \partial_{P_{U|y_1}}\mathcal{F}[P^{t_s+1}_{U|y_1}, P^{t_s}_{U|y_2}] - p(y_1)\Lambda^{t_s} \\
&+\rho p(y_1)[P^{t_s}_U - p(y_1)P^{t_s+1}_{U|y_1} - p(y_1)P^t_{U|y_2}], \\
\mathbf{0} \in\ & \partial_{P_{U|y_2}}\mathcal{F}[P^{t_s+1}_{U|Y}] - p(y_2)\Lambda^{t_s} \\
&+\rho p(y_2)[P^{t_s}_U - (P^{t_s+1}_{U|Y})^T P_Y], \\
\mathbf{0} \in\ & \partial_{P_U} g\left(P^{t_s+1}_{U|y_1}, P^{t_s+1}_{U|y_2}, P^{t_s+1}_U\right) + \Lambda^{t_s} \\
&-\rho\left(P^{t_s+1}_U - p(y_1)P^{t_s+1}_{U|y_1} - p(y_2)P^{t_s+1}_{U|y_2}\right).
\end{aligned}
$$

Taking the limit along the subsequence and using Proposition 2, we have

$$
\begin{aligned}
0 \in\ & \partial_{P_{U|y_1}}\mathcal{F}[\hat{P}_{U|y_1}] - p(y_1)\hat{\Lambda} \\
0 \in\ & \partial_{P_{U|y_2}}\mathcal{F}[\hat{P}_{U|y_2}] - p(y_2)\hat{\Lambda} \\
0 \in\ & \partial_{P_U}\mathcal{F}[\hat{P}_U] + \hat{\Lambda},
\end{aligned}
$$

which indicates that the stationary condition is satisfied at the limit point $(\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda})$.

Now we check all constraints in (11) are also satisfied at the limit point.

- Since $P^{t_s}_{U|Y} \in \mathcal{P}_{U|Y}, \forall s$, and $\mathcal{P}_{U|Y}$ is a closed set, we have $\hat{P}_{U|Y} \in \mathcal{P}_{U|Y}$;
- By taking limit along the subsequence on both sides of the equation in Proposition 2 5), we have

$$
\hat{P}_U = p(y_1)\hat{P}_{U|y_1} + p(y_2)\hat{P}_{U|y_1}; \tag{40}
$$

- Based on (40), we have $\hat{p}(u) > 0, \forall u$, and

$$
\begin{aligned}
\sum_u \hat{p}(u) &= \sum_u\sum_y \hat{p}(u|y)p(y) \\
&= \sum_y p(y)\sum_u \hat{p}(u|y) = \sum_y p(y) = 1,
\end{aligned}
$$

which indicate that $\hat{P}_U \in \mathcal{P}_U$.

Appendix K
Proof of Theorem 1

Since $\mathcal{L}[P^t_{U|Y}, P^t_U, \Lambda^t]$ is non-decreasing between iterations and bounded from above, we have that $\mathcal{L}[P^{t_s}_{U|Y}, P^{t_s}_U, \Lambda^{t_s}]$ is also monotonic non-decreasing and upper-bounded. Then we have $\lim_{s\to\infty}\mathcal{L}[P^{t_s}_{U|Y}, P^{t_s}_U, \Lambda^{t_s}] = \mathcal{L}[\hat{P}_{U|Y}, \hat{P}_U, \hat{\Lambda}]$ as $\mathcal{L}$ is continuous for $P_{U|Y} \in \mathcal{P}_{U|Y}, P_U \in \mathcal{P}_U$, and Theorem 1 is proved following from Proposition 3.

APPENDIX L
PROOF OF LEMMA 9

The optimality condition of v-subproblem yields

$$\begin{aligned}0 =&\nabla_v g\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}\right) - \Lambda^k + \rho\left(p(y_1)x_1^{k+1}\right.\\ &\left.+p(y_2)x_2^{k+1} - v^{k+1}\right) - \nabla\phi\left(v^{k+1}\right) + \nabla\phi\left(v^k\right).\end{aligned}$$

As $\Lambda^{k+1} = \Lambda^k - \rho\left(p(y_1)x_1^{k+1} + p(y_2)x_2^{k+1} - v^{k+1}\right)$, we have $\Lambda^{k+1} = \nabla_v g\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}\right) - \nabla\phi\left(v^{k+1}\right) + \nabla\phi\left(v^k\right)$. Thus,

$$\begin{aligned}&\left\|\Lambda^{k+1} - \Lambda^k\right\|_2^2\\ &= \left\|\nabla_v g\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}\right) - \nabla_v g\left(x_1^k, x_2^k, v^k\right)\right.\\ &\left.-\nabla\phi\left(v^{k+1}\right) + \nabla\phi\left(v^k\right) + \nabla\phi\left(v^k\right) - \nabla\phi\left(v^{k-1}\right)\right\|_2^2\\ &\le 3\left(\left\|\nabla_v g\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}\right) - \nabla_v g\left(x_1^k, x_2^k, v^k\right)\right\|_2^2\right.\\ &\left.+\left\|\nabla\phi\left(v^{k-1}\right) - \nabla\phi\left(v^k\right)\right\|_2^2 + \left\|\nabla\phi\left(v^k\right) - \nabla\phi\left(v^{k+1}\right)\right\|_2^2\right)\\ &\le 3l_g^2\left(\left\|x_1^{k+1} - x_1^k\right\|_2^2 + \left\|x_2^{k+1} - x_2^k\right\|_2^2\right)\\ &+3\left(l_g^2 + l_\phi^2\right)\left\|v^{k+1} - v^k\right\|_2^2 + 3l_\phi^2\left\|v^k - v^{k-1}\right\|_2^2.\end{aligned}$$

APPENDIX M
PROOF OF LEMMA 10

From the update of x_1, we have

$$\begin{aligned}&h_1\left(x_1^{k+1}\right) + \left\langle x_1^{k+1} - x_1^k, \nabla_{x_1} g\left(u^k\right)\right\rangle\\ &\quad+\left\langle\Lambda^k, p(y_1)x_1^{k+1} + p(y_2)x_2^k - v^k\right\rangle\\ &\quad-\frac{\rho}{2}\left\|p(y_1)x_1^{k+1} + p(y_2)x_2^k - v^k\right\|_2^2 - \Delta_{\varphi_1}\left(x_1^{k+1}, x_1^k\right)\\ &\ge h_1\left(x_1^k\right) + \left\langle\Lambda^k, r_k\right\rangle - \frac{\rho}{2}\left\|r_k\right\|_2^2,\end{aligned}$$

where $u^k = (x_1^k, x_2^k, y^k)^T$ and $r_k = p(y_1)x_1^k + p(y_2)x_2^k - v^k$.

From the update of x_2, we have

$$\begin{aligned}&h_2\left(x_2^{k+1}\right) + \left\langle x_2^{k+1} - x_2^k, \nabla_{x_2} g\left(u^k\right)\right\rangle\\ &\quad+\left\langle\Lambda^k, p(y_1)x_1^{k+1} + p(y_2)x_2^{k+1} - v^k\right\rangle\\ &\quad-\frac{\rho}{2}\left\|p(y_1)x_1^{k+1} + p(y_2)x_2^{k+1} - v^k\right\|_2^2 - \Delta_{\varphi_2}\left(x_2^{k+1}, x_2^k\right)\\ &\ge h_2\left(x_2^k\right) + \left\langle\Lambda^k, p(y_1)x_1^{k+1} + p(y_2)x_2^k - v^k\right\rangle\\ &\quad-\frac{\rho}{2}\left\|p(y_1)x_1^{k+1} + p(y_2)x_2^k - v^k\right\|_2^2.\end{aligned}$$

From the update of v, we have

$$\begin{aligned}&g\left(u^{k+1}\right) + \left\langle\Lambda^k, r_{k+1}\right\rangle - \frac{\rho}{2}\left\|r_{k+1}\right\|_2^2 - \Delta_\phi\left(v^{k+1}, v^k\right)\\ &\ge g\left(x_1^{k+1}, x_2^{k+1}, v^k\right) - \frac{\rho}{2}\left\|p(y_1)x_1^{k+1} + p(y_2)x_2^{k+1} - v^k\right\|_2^2\\ &\quad+\left\langle\Lambda^k, p(y_1)x_1^{k+1} + p(y_2)x_2^{k+1} - v^k\right\rangle,\end{aligned}$$

where $u^{k+1} = (x_1^{k+1}, x_2^{k+1}, y^{k+1})^T$ and $r_{k+1} = p(y_1)x_1^{k+1} + p(y_2)x_2^{k+1} - v^{k+1}$.

Adding up the above three inequalities, we have

$$\begin{aligned}&\mathcal{L}\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}, \Lambda^k\right) = h_1\left(x_1^{k+1}\right) + h_2\left(x_2^{k+1}\right)\\ &\quad+g\left(u^{k+1}\right) + \left\langle\Lambda^k, r_{k+1}\right\rangle - \frac{\rho}{2}\left\|r_{k+1}\right\|_2^2\\ &\ge h_1\left(x_1^k\right) + h_2\left(x_2^k\right) + g\left(x_1^{k+1}, x_2^{k+1}, v^k\right) + \left\langle\Lambda^k, r_k\right\rangle\\ &\quad-\left[\left\langle x_1^{k+1} - x_1^k, \nabla_{x_1} g\left(u^k\right)\right\rangle + \left\langle x_2^{k+1} - x_2^k, \nabla_{x_2} g\left(u^k\right)\right\rangle\right]\\ &\quad+\Delta_{\varphi_1}\left(x_1^{k+1}, x_1^k\right) + \Delta_{\varphi_2}\left(x_2^{k+1}, x_2^k\right) + \Delta_\phi\left(v^{k+1}, v^k\right)\\ &\quad-\frac{\rho}{2}\left\|r_k\right\|_2^2\\ &= h_1\left(x_1^k\right) + h_2\left(x_2^k\right) + g\left(u^k\right) + \left\langle\Lambda^k, r_k\right\rangle - \frac{\rho}{2}\left\|r_k\right\|_2^2\\ &\quad-g\left(u^k\right) + g\left(x_1^{k+1}, x_2^{k+1}, v^k\right)\\ &\quad-\left\langle\left(x_1^{k+1} - x_1^k, x_2^{k+1} - x_2^k, 0\right), \nabla g\left(u^k\right)\right\rangle\\ &\quad+\Delta_{\varphi_1}\left(x_1^{k+1}, x_1^k\right) + \Delta_{\varphi_2}\left(x_2^{k+1}, x_2^k\right) + \Delta_\phi\left(v^{k+1}, v^k\right)\\ &= \mathcal{L}\left(w^k\right) + g\left(x_1^{k+1}, x_2^{k+1}, v^k\right) - g\left(u^k\right)\\ &\quad-\left\langle\left(x_1^{k+1} - x_1^k, x_2^{k+1} - x_2^k, 0\right), \nabla g\left(u^k\right)\right\rangle\\ &\quad+\Delta_{\varphi_1}\left(x_1^{k+1}, x_1^k\right) + \Delta_{\varphi_2}\left(x_2^{k+1}, x_2^k\right) + \Delta_\phi\left(v^{k+1}, v^k\right)\\ &\overset{(a)}{\ge}\mathcal{L}\left(w^k\right) - \frac{l_g}{2}\left[\left\|x_1^{k+1} - x_1^k\right\|_2^2 + \left\|x_2^{k+1} - x_2^k\right\|_2^2\right]\\ &\quad+\frac{\delta_{\varphi_1}}{2}\left\|x_1^{k+1} - x_1^k\right\|_2^2 + \frac{\delta_{\varphi_2}}{2}\left\|x_2^{k+1} - x_2^k\right\|_2^2\\ &\quad+\frac{\delta_\phi}{2}\left\|v^{k+1} - v^k\right\|_2^2,\end{aligned}$$

where (a) follows from the assumption 3) and the fact from [37] that if $h : \mathbb{R}^n \longrightarrow \mathbb{R}$ is a continuous differentiable function where gradient ∇h is Lipschitz continuous with the modulus $l_h > 0$, then for any $x, y \in \mathbb{R}^n$, we have $|h(y) - h(x) - \langle\nabla h(x), y - x\rangle| \le \frac{l_h}{2}\|y - x\|_2^2$, and apply this result on g here.

By using the fact that

$$\langle\Lambda^{k+1} - \Lambda^k, r_{k+1}\rangle = -\frac{1}{\rho}\|\Lambda^{k+1} - \Lambda^k\|_2^2,$$

we have

$$\begin{aligned}&\mathcal{L}\left(w^{k+1}\right) - \mathcal{L}\left(w^k\right)\\ =\;&\mathcal{L}\left(w^{k+1}\right) - \mathcal{L}\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}, \Lambda^k\right)\\ &+\mathcal{L}\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}, \Lambda^k\right) - \mathcal{L}\left(w^k\right)\\ =\;&-\frac{1}{\rho}\left\|\Lambda^{k+1} - \Lambda^k\right\|_2^2\\ &+\mathcal{L}\left(x_1^{k+1}, x_2^{k+1}, v^{k+1}, \Lambda^k\right) - \mathcal{L}\left(w^k\right)\\ \ge\;&\left(\frac{\delta_{\varphi_1} - l_g}{2} - \frac{3l_g^2}{\rho}\right)\left\|x_1^{k+1} - x_1^k\right\|_2^2\\ &+\left(\frac{\delta_{\varphi_2} - l_g}{2} - \frac{3l_g^2}{\rho}\right)\left\|x_2^{k+1} - x_2^k\right\|_2^2\\ &+\left(\frac{\delta_\phi}{2} - \frac{3l_g^2 + 3l_\phi^2}{\rho}\right)\left\|v^{k+1} - v^k\right\|_2^2\\ &-\frac{3l_\phi^2}{\rho}\left\|v^k - v^{k-1}\right\|_2^2,\end{aligned}$$

which implies

$$\begin{aligned}
&\left(\mathcal{L}\left(w^{k+1}\right)-\frac{3l_\phi^2}{\rho}\left\|v^{k+1}-v^k\right\|_2^2\right)\\
&\quad-\left(\mathcal{L}\left(w^{k}\right)-\frac{3l_\phi^2}{\rho}\left\|v^{k}-v^{k-1}\right\|_2^2\right)\\
\geq\ &\left(\frac{\delta_{\varphi_1}-l_g}{2}-\frac{3l_g^2}{\rho}\right)\left\|x_1^{k+1}-x_1^k\right\|_2^2\\
&+\left(\frac{\delta_{\varphi_2}-l_g}{2}-\frac{3l_g^2}{\rho}\right)\left\|x_2^{k+1}-x_2^k\right\|_2^2\\
&+\left(\frac{\delta_{\phi}}{2}-\frac{3l_g^2+6l_\phi^2}{\rho}\right)\left\|v^{k+1}-v^k\right\|_2^2.
\end{aligned}$$

REFERENCES

[1] Y. Jin and L. Lai, "Privacy-accuracy trade-off of inference as service," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, (Toronto, Canada), pp. 2645–2649, Jun. 2021.

[2] F. Meneghello, M. Calore, D. Zucchetto, M. Polese, and A. Zanella, "IoT: Internet of threats? A survey of practical security vulnerabilities in real IoT devices," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8182–8201, May. 2019.

[3] A. Ahrabian, S. Kolozali, S. Enshaeifar, C. Cheong-Took, and P. Barnaghi, "Data analysis as a web service: A case study using IoT sensor data," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, (New Orleans, LA), pp. 6000–6004, Mar. 2017.

[4] M. Sun, W. P. Tay, and X. He, "Toward information privacy for the Internet of things: A nonparametric learning approach," *IEEE Transactions on Signal Processing*, vol. 66, no. 7, pp. 1734–1747, Jan. 2018.

[5] J. Wurm, K. Hoang, O. Arias, A. Sadeghi, and Y. Jin, "Security analysis on consumer and industrial IoT devices," in *Proc. Asia and South Pacific Design Automation Conference*, (Macao, China), pp. 519–524, Jan. 2016.

[6] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the Internet of things: A survey," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 414–454, 2013.

[7] C. Zhang, M. Yu, W. Wang, and F. Yan, "Mark: Exploiting cloud services for cost-effective, slo-aware machine learning inference serving," in *Proc. Annual Technical Conference*, (Renton, WA), pp. 1049–1062, Jul. 2019.

[8] A. Gujarati, S. Elnikety, Y. He, K. McKinley, and B. Brandenburg, "Swayam: Distributed autoscaling to meet SLAs of machine learning inference services with resource efficiency," in *Proc. Middleware Conference*, (Las Vegas, NV), pp. 109–120, Dec. 2017.

[9] M. Tebaa, S. El Hajji, and A. El Ghazi, "Homomorphic encryption applied to the cloud computing security," in *Proc. World Congress on Engineering*, vol. 1, (London, U.K), pp. 4–6, Jul. 2012.

[10] F. Boemer, A. Costache, R. Cammarota, and C. Wierzynski, "nGraph-HE2: A high-throughput framework for neural network inference on encrypted data," in *Proc. ACM Workshop on Encrypted Computing & Applied Homomorphic Cryptography*, (London, UK), pp. 45–56, Nov. 2019.

[11] C. Gentry and D. Boneh, *A fully homomorphic encryption scheme*, vol. 20. Stanford university, 2009.

[12] I. Issa, A. Wagner, and S. Kamath, "An operational approach to information leakage," *IEEE Transactions on Information Theory*, vol. 66, no. 3, pp. 1625–1657, Mar. 2019.

[13] Y. Wang, W. Yin, and J. Zeng, "Global convergence of ADMM in nonconvex nonsmooth optimization," *Journal of Scientific Computing*, vol. 78, no. 1, pp. 29–63, Feb. 2019.

[14] M. Chao, Z. Deng, and J. Jian, "Convergence of linear Bregman ADMM for nonconvex and nonsmooth problems with nonseparable structure," *Complexity*, Feb. 2020.

[15] C. Chen, M. Li, X. Liu, and Y. Ye, "Extended ADMM and BCD for nonseparable convex minimization models with quadratic coupling terms: convergence analysis and insights," *Mathematical Programming*, vol. 173, no. 1-2, pp. 37–77, Jan. 2019.

[16] P. Samarati and L. Sweeney, "Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression," 1998.

[17] C. Dwork, "Differential privacy: A survey of results," in *Proc. International Conference on Theory and Applications of Models of Computation*, (Xi'an, China), pp. 1–19, Apr. 2008.

[18] F. Calmon and N. Fawaz, "Privacy against statistical inference," in *Proc. Annual Allerton Conference on Communication, Control, and Computing*, (Monticello, IL), pp. 1401–1408, Oct. 2012.

[19] C. Glackin, G. Chollet, N. Dugan, N. Cannings, J. Wall, S. Tahir, I. G. Ray, and M. Rajarajan, "Privacy preserving encrypted phonetic search of speech data," in *Proc. International Conference on Acoustics, Speech and Signal Processing*, (New Orleans, LA), pp. 6414–6418, Mar. 2017.

[20] X. Wang, H. Ishii, L. Du, P. Cheng, and J. Chen, "Privacy-preserving distributed machine learning via local randomization and ADMM perturbation," *IEEE Transactions on Signal Processing*, vol. 68, pp. 4226–4241, Jul. 2020.

[21] B. Martin, M. Natalia, P. Afroditi, Q. Qiang, R. Miguel, R. Galen, and S. Guillermo, "Adversarially learned representations for information obfuscation and inference," in *Proc. International Conference on Machine Learning*, (Long Beach, CA), pp. 614–623, Jun. 2019.

[22] J. Hamm, "Minimax filter: Learning to preserve privacy from inference attacks," *The Journal of Machine Learning Research*, vol. 18, no. 1, pp. 4704–4734, 2017.

[23] A. Tripathy, Y. Wang, and P. Ishwar, "Privacy-preserving adversarial networks," in *Proc. Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, (Monticello, IL), pp. 495–505, Sep. 2019.

[24] Y. Peng, A. Ganesh, J. Wright, W. Xu, and Y. Ma, "Rasl: Robust alignment by sparse and low-rank decomposition for linearly correlated images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 11, pp. 2233–2246, Jan. 2012.

[25] W. Deng, M. Lai, Z. Peng, and W. Yin, "Parallel multi-block ADMM with o (1/k) convergence," *Journal of Scientific Computing*, vol. 71, no. 2, pp. 712–736, May 2017.

[26] M. Li, D. Sun, and C. Toh, "A convergent 3-block semi-proximal ADMM for convex minimization problems with one strongly convex block," *Asia-Pacific Journal of Operational Research*, vol. 32, no. 04, p. 1550024, Aug. 2015.

[27] T. Lin, S. Ma, and S. Zhang, "On the sublinear convergence rate of multi-block ADMM," *Journal of the Operations Research Society of China*, vol. 3, no. 3, pp. 251–274, Sep. 2015.

[28] X. Wang, M. Hong, S. Ma, and Z. Luo, "Solving multiple-block separable convex minimization problems using two-block alternating direction method of multipliers," *arXiv preprint arXiv:1308.5294*, Aug. 2013.

[29] A. Böttcher and D. Wenzel, "The Frobenius norm and the commutator," *Linear Algebra and Its Applications*, vol. 429, no. 8-9, pp. 1864–1885, May 2008.

[30] C. Chen, B. He, Y. Ye, and X. Yuan, "The direct extension of ADMM for multi-block convex minimization problems is not necessarily convergent," *Mathematical Programming*, vol. 155, no. 1-2, pp. 57–79, Jan. 2016.

[31] X. Cai, D. Han, and X. Yuan, "On the convergence of the direct extension of ADMM for three-block separable convex minimization models with one strongly convex function," *Computational Optimization and Applications*, vol. 66, no. 1, pp. 39–73, Jan. 2017.

[32] Y. Cui, X. Li, D. Sun, and C. Toh, "On the convergence properties of a majorized ADMM for linearly constrained convex optimization problems with coupled objective functions," *arXiv preprint arXiv:1502.00098*, Jan. 2015.

[33] F. Wang, W. Cao, and Z. Xu, "Convergence of multi-block bregman ADMM for nonconvex composite problems," *Science China Information Sciences*, vol. 61, no. 12, pp. 1–12, Dec. 2018.

[34] B. Fuglede and F. Topsoe, "Jensen-Shannon divergence and Hilbert space embedding," in *Proc. International Symposium on Information Theory*, (Parma, Italy), p. 31, Oct. 2004.

[35] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam*, pp. 423–435, Springer, 1997.

[36] L. Le Cam and G. L. Yang, *Asymptotics in statistics: some basic concepts*. Springer Science & Business Media, 2012.

[37] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer Science & Business Media, 2003.