PRIVACY-ACCURACY TRADE-OFF OF INFERENCE AS SERVICE

Yulu Jin and Lifeng Lai

Department of ECE, University of California, Davis
Email:{yuljin,lflai}@ucdavis.edu

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. 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 optimal 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.

***Index Terms*—** statistical inference, privacy-preserving, privacy-accuracy trade-off, iterative algorithm.

1. INTRODUCTION

The Internet of Things (IoT) is an emerging communication paradigm that aims at connecting different kinds of devices to the Internet [1, 2, 3]. Within the past decade, the number of IoT devices being introduced in the market has increased dramatically due to its low cost and convenience [4]. However, as machine learning models with a large number of parameters become much more complex [5], it is difficult to run those models on IoT devices. One of the emerging solutions to this problem is so called inference as service (IAS) [6, 7]. In IAS, the devices will send data to a server in the cloud, who will make inference using powerful machine learning models. However, IAS 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 [8, 9, 10]. 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 optimal privacy-preserving mapping. As the inference accuracy is directly related to the mutual information, we use mutual information to measure the inference accuracy. However, determining the privacy measure is tricky, as there are many existing information leakage measures [11], 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 to measure privacy leakage. The proposed privacy leakage metric is defined by a continuous function f. 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.

If we optimize over the space of privacy-mapping directly, the formulated optimization problem is a complicated non-convex problem. Through various transformations and variable augmentations, we reveal certain nice structures of the optimization problem. We then exploit these structures to design an iterative update algorithm to solve the optimization problem for general f. Compared with solving the optimization problem using gradient ascent in the space of privacy-mapping directly, the proposed 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.

There exist many other privacy-preserving techniques that are based on perturbations of data and thus provide privacy guarantees at the expense of a loss of accuracy [12, 13, 14]. For example, k-anonymity is proposed by Samarati and Sweeney [15], which requires that each record is indistinguishable from at least k–1 other records within the dataset. Differential privacy works by adding a predetermined amount of randomness into a computation per-

This work was supported in part by National Science Foundation under Grants CCF-1717943, ECCS-1711468, CNS-1824553, CCF-1908258 and ECCS-2000415.

formed on a data set [16]. Moreover, various minimax formulations and algorithms have also been proposed to defend against inference attack in different scenarios [17, 18, 19]. 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.

2. PROBLEM FORMULATION

Consider an inference problem, in which one would like to infer the parameter $S \in \mathcal{S}$ of data $Y \in \mathcal{Y}$, where $\mathcal{Y}$ is a finite set. For example, Y could be a picture and S could be the label of the picture. 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, where $\mathcal{U}$ is also a finite set. 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 $U = u$. Furthermore, we use $p(s)$ to denote the prior distribution of S and $p(y|s)$ to denote the conditional distribution Y given S.

Even though this mapping reduces the privacy leakage, it could unfortunately reduce inference accuracy. As the result, to find the optimal map $p(u|y)$, we need to strike a balance between the inference accuracy and data privacy.

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 KL divergence between $p(s)$ and $p(s|u)$, we use it to measure the inference accuracy.

To measure the privacy leakage, we consider a general metric $\mathbb{E}_{Y,U}[d(y,u)]$ to characterize the distance Y and U. The larger the distance, the better the privacy protection. Here $d(y,u) = f(\frac{p(y)}{p(y|u)})$ and f is a continuous function defined on $(0, +\infty)$. Different choices of f will lead to different privacy measures (examples will be provided in the sequel). Let $\phi(y,u) = \frac{p(y)}{p(y|u)}$, which can be viewed as a ratio representing the information provided by the observed sample u.

Note that $\phi(y,u) = \frac{p(y)}{p(y|u)} = \frac{p(u)}{p(u|y)}$. Hence we will also use $\frac{p(u)}{p(u|y)}$ to calculate $\phi(y,u)$ in the sequel.

Using these measures, we formulate an optimization problem to find the optimal mapping $p(u|y)$

$$\max_{p(u|y)} \quad \mathcal{F}[p(u|y)] \triangleq I(S;U) + \beta \mathbb{E}_{Y,U}[d(y,u)], \quad (1)$$

$$\text{s.t.} \quad \sum_u p(u|y) = 1, \forall y \in \mathcal{Y}. \quad (2)$$

Here, $\beta \in (0, \infty)$ is a weight that indicates the relative importance of maximizing $I(S;U)$, which will lead to a better inference accuracy, and maximizing the distance $\mathbb{E}_{Y,U}[d(y,u)]$ between Y and U, which will enhance the privacy protection.

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

$$\begin{aligned}\mathbb{E}_{Y,U}[d(y,u)] &= \sum_{y,u} p(y)p(u|y) \log\left(\frac{p(u)}{p(u|y)}\right) \\ &= -\sum_y p(y) D_{KL}[p(u|y) \parallel p(u)] = -I[U;Y],\end{aligned}$$

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 information leakage, a very common choice in information theory study. More examples will be provided in Section 4.

3. ALGORITHM

In this section, we discuss how to solve the optimization problem defined in (1) for general f. As the objective function is a complicated non-convex function of $p(u|y)$, we only expect to find the local maximal point. One possible approach is to apply the gradient ascent (GA) algorithm, which faces several challenges such as proper step size, computation complexity, convergence speed and the quality of the local optimal point found etc. To overcome these challenges, we propose a new method motivated by the information bottleneck methods [20], which transforms the maximization over single argument to an alternative maximization problem over multiple arguments. We first find three arguments and show that the objective function is concave with respect to each argument. Then based on these properties, we develop an iterative algorithm to find the local maxima.

We first have the following lemma that facilitates further analysis.

Lemma 1.

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

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

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

As the result, the objective function defined in (1) can be viewed as a functional on $p(u|y), p(u), p(s|u)$. By using Lemma 1, we have the following lemma.

Lemma 2. *Suppose that f is a strictly concave function. Then for given $p(s|u)$, $p(u)$, $\mathcal{F}[p(s|u), p(u), p(u|y)]$ is*

concave in $p(u|y)$. Similarly, for given $p(s|u)$, $p(u|y)$, $\mathcal{F}[p(s|u), p(u), p(u|y)]$ is concave in $p(u)$. For given $p(u|y)$, $p(u)$, $\mathcal{F}[p(s|u), p(u), p(u|y)]$ is concave in $p(s|u)$.

Lemma 3. *For a strictly concave function $f(\cdot)$, if $\lim_{t\to+\infty} f'(t) < \infty$, then $\mathcal{F}[p(u|y)]$ is bounded from above.*

Suppose that $f(\cdot)$ is a strictly concave function and $\lim_{t\to+\infty} f'(t) < \infty$, by Lemma 2, the original optimization problem can be converted to

$$\begin{aligned} \max_{p(s|u)} \max_{p(u)} \max_{p(u|y)} \quad & \mathcal{F}[p(s|u), p(u), p(u|y)]. \\ \text{subject to} \quad & \sum_u p(u|y) = 1, \forall y \in \mathcal{Y}, \\ & \sum_u p(u) = 1, \\ & \sum_s p(s|u) = 1, \forall u \in \mathcal{U}. \end{aligned} \tag{3}$$

We can now exploit the structural property of the objective function $\mathcal{F}[p(s|u), p(u), p(u|y)]$ presented in Lemma 2 to iteratively find a solution. In particular, given $p_t(u), p_t(s|u)$ and $p_t(u|y)$ obtained in iteration t, we obtain $p_{t+1}(u)$, $p_{t+1}(s|u)$ and $p_{t+1}(u|y)$ for iteration $t+1$ one by one.

In the first step, given $p_t(s|u), p_t(u)$, we obtain $p_{t+1}(u|y)$ by solving the optimization problem

$$\max_{p(u|y)} \quad \mathcal{F}[p_t(s|u), p_t(u), p(u|y)], \tag{4}$$

which is concave as shown in Lemma 2. We denote the solution obtained as

$$p_{t+1}(u|y) = g(p_t(u), p_{t+1}(s|u)), \tag{5}$$

whose form depends on the choice of $f(\cdot)$ used. For many f's that are widely used, $g(\cdot)$ has a closed form expression. We will provide such examples in Section 4.

In the second step, we obtain $p_{t+1}(u)$ by Bayesian rule

$$p_{t+1}(u) = \sum_y p_{t+1}(u|y)p(y). \tag{6}$$

Finally, after obtaining $p_{t+1}(u|y)$ and $p_{t+1}(u)$, we can obtain $p_{t+1}(s|u)$ by solving

$$\max_{p(s|u)} \quad \mathcal{F}[p(s|u), p_{t+1}(u), p_{t+1}(u|y)],$$

which again is a concave optimization problem as shown in Lemma 2. In fact, we can show that for any $f(\cdot)$, this optimization problem has a simple closed form solution

$$p_{t+1}(s|u) = \frac{\sum_y p_{t+1}(u|y)p(s,y)}{p_{t+1}(u)}. \tag{7}$$

We now intuitively explain the reason for applying such iterative process. Since the objective function (4) is maximized in (5), the update (5) increases the function value $\mathcal{F}$. Similarly update (7) increases the function value $\mathcal{F}$. Furthermore, by Lemma 3, the objective function $\mathcal{F}$ is upper-bounded. Hence, the proposed iterative process can find a local maxima.

The algorithm is summarized in Algorithm 1. We note that the updates are computed element-wise, and hence these updates can be computed efficiently, regardless of the alphabet size. Furthermore, we do not need to choose the step size.

Algorithm 1 Design the optimal privacy-preserving mapping

Input:
Prior distribution $p(s)$ and conditional distribution $p(y|s)$.
Trade-off parameter β.
Converge parameter ϵ.
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 (6) and (7).

1: **while** difference$< \epsilon$ **do**
2: update $p(u|y), p(u), p(s|u)$ by (5)(6)(7)
3: calculate difference
4: **return** $p(u|y)$

4. EXAMPLES

We now apply the framework to different privacy measures by choosing different f.

In the first example, we consider $f(x) = \log(x)$. As shown in Section 2, the privacy measure is then the mutual information. The update equation for $p(u|y)$ has been shown to have the following closed form solution in [20],

$$p_{t+1}(u|y) = \frac{p(u)}{Z(y,\beta)} \exp(-\beta D_{KL}[p(s|y) \parallel p_t(s|u)]),$$

with $Z(y,\beta) = \sum_u p(u) \exp(-\beta D_{KL}[p(s|y) \parallel p_t(s|u)])$ being the normalization function.

In the second example, we consider the following strictly convex function

$$f(x) = x \log \frac{2x}{x+1} + \log \frac{2}{x+1}. \tag{8}$$

This choice leads to the Jensen-Shannon divergence [21]:

$$\mathbb{E}_{Y,U}[d(y,u)] = -\sum_y p(y) JS[p(u|y), p(u)].$$

It is easy to check that the condition in Lemma 3 is satisfied for this choice of f.

The update equation for $p(u|y)$ in (5) has the closed form solution

$$p_{t+1}(u|y) = \frac{p_t(u)}{\exp\{\frac{1}{\beta} D_{KL}[p(s|y) \parallel p_t(s|u)] + \mu(y)\} - 1},$$

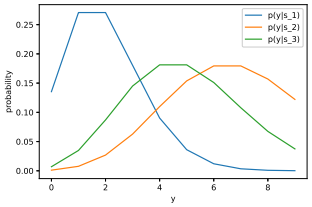

Fig. 1. Conditional distribution $p(y|s)$

in which $\mu(y)$ is determined by the normalization condition $\sum_u p_{t+1}(u|y) = 1$. We can further show that there always exists a unique $\mu(y)$ that satisfies this condition.

In the third example, consider the function

$$f(x) = \frac{1-x}{2x+2}, \tag{9}$$

which leads to the Le Cam divergence [22] as the privacy measure,

$$\mathbb{E}_{Y,U}[d(y,u)] = -\sum_y p(y) LC[p(u|y) \parallel p(u)].$$

For this choice of f, again, the update $g(\cdot)$ has a closed form

$$p_{t+1}(u|y) = p(u)\left\{\left[\frac{1}{\beta} D_{KL}[p(s|y) \parallel p_t(s|u)] + \nu(y)\right]^{-\frac{1}{2}} - 1\right\},$$

in which $\nu(y)$ can be uniquely determined from the normalization equation $\sum_u p_{t+1}(u|y) = 1$.

5. NUMERICAL RESULT

In this section, we provide numerical results to show that the methods proposed here converges much faster than GA, and the local maxima found by our methods has much better quality than the one found by GA.

In our experiment, we set the prior distribution $\boldsymbol{p}_s = \{\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\}$ and let $|\mathcal{Y}| = 10, |\mathcal{U}| = 11$. The conditional distributions $p(y|s)$ under each s are shown in Fig. 5. Under this setup, we will perform both Algorithm 1 and GA to find the optimal transition mapping $p(u|y)$ that maximizes the functional defined in (1) using different f.

We first set f as in (8), which means we use Jensen-Shannon divergence as the privacy metric. The initial mapping $p(u|y)$ is obtained by selecting random numbers conforming to uniform distribution and normalizing them. After applying Algorithm 1, we plot the relationship between the function value and iteration in Fig 2. For comparison purpose, we also plot the corresponding figure for GA in Fig. 3. From these figures, we can see that Algorithm 1 converges after 10 iterations. On the other hand, even after 8000 iterations, the value of the objective function found by GA is very

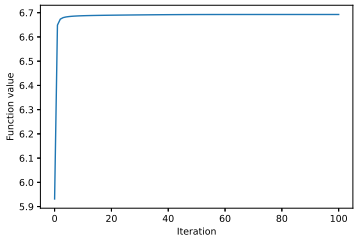

Fig. 2. Convergence process of Algorithm 1 (JS divergence)

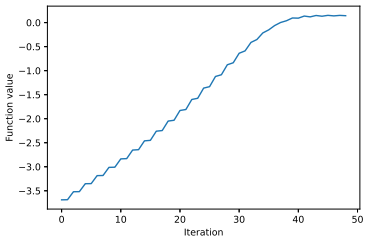

Fig. 3. Convergence process of GA (JS divergence)

far away from that value found by Algorithm 1. In the second example, we set f as in (9), which corresponds to the Le Cam divergence as discussed in Section 4. We again compare Algorithm 1 and GA. The results are shown in Table 1. From the table, we can see that the local maximum value found by our method is larger than the one found by GA. Moreover, since the objective function is quite complex in $p(u|y)$, it is hard to find a proper step size for the GA algorithm while there is no need to do so in our algorithm.

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

Table 1. Convergence results comparison

6. CONCLUSION

We have proposed a general framework to design privacy-preserving mapping to achieve privacy-accuracy trade-off in the inference as services scenarios. We have formulated an optimization problem to find the optimal mapping. We have discussed the structure of the formulated problem and designed an iterative method to solve this complicated optimization problem. We have provided numerical results showing that this method has better performance than GA in the convergence speed, solution quality and algorithm stability.

7. REFERENCES

[1] 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.

[2] 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. IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, Mar. 2017, pp. 6000–6004.

[3] 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.

[4] 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, Jan. 2016, pp. 519–524.

[5] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, MN, Apr. 2019, pp. 4171–4186.

[6] 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, Jul. 2019, pp. 1049–1062.

[7] 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. ACM/IFIP/USENIX Middleware Conference*, Las Vegas, NV, Dec. 2017, pp. 109–120.

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

[9] 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, Nov. 2019, pp. 45–56.

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

[11] 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.

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

[13] 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. IEEE International Conference on Acoustics, Speech and Signal Processing*, New Orleans, LA, Mar. 2017, pp. 6414–6418.

[14] 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.

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

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

[17] 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, Jun. 2019, pp. 614–623.

[18] 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.

[19] A. Tripathy, Y. Wang, and P. Ishwar, "Privacy-preserving adversarial networks," in *Proc. 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Monticello, IL, Sep. 2019, pp. 495–505.

[20] N. Tishby, F. Pereira, and W. Bialek, "The information bottleneck method," *arXiv preprint physics/0004057*, Jan. 2000.

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

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