# The Utility Cost of Robust Privacy Guarantees

Hao Wang*, Mario Diaz*†, Flavio P. Calmon* and Lalitha Sankar†
*Harvard University, {hao_wang,mdiaztor}@g.harvard.edu, flavio@seas.harvard.edu
†Arizona State University, {mdiaztor,lsankar}@asu.edu

***Abstract*—Consider a data publishing setting for a data set with public and private features. The objective of the publisher is to maximize the amount of information about the public features in a revealed data set, while keeping the information leaked about the private features bounded. The goal of this paper is to analyze the performance of privacy mechanisms that are constructed to match the distribution learned from the data set. Two distinct scenarios are considered: (i) mechanisms are designed to provide a privacy guarantee for the learned distribution; and (ii) mechanisms are designed to provide a privacy guarantee for every distribution in a given neighborhood of the learned distribution. For the first scenario, given any privacy mechanism, upper bounds on the difference between the privacy-utility guarantees for the learned and true distributions are presented. In the second scenario, upper bounds on the reduction in utility incurred by providing a uniform privacy guarantee are developed.**

## I. Introduction

The disclosure of data with both privacy and utility guarantees is a recognized objective in many applications. A common approach to this problem is to process the data set through a *privacy mechanism* that seeks to fulfill certain privacy and utility guarantees. Information theoretic methods for designing privacy mechanisms often rely on the implicit assumption that the data distribution is, for the most part, known [1]–[4]. However, in practice, the data distribution may only be accessed through a limited number of observed samples.

In this work, we revisit this assumption, and study the robustness of privacy and utility guarantees of information-theoretic privacy mechanisms to partial knowledge of the input distribution. In practice, this inaccuracy stems from the limited availability of samples which, in turn, produces a discrepancy between the learned and the true data distribution. To mitigate the effect of this discrepancy, we also study the performance of privacy mechanisms that, by design, are robust: they assure privacy for *every* data set drawn from a distribution within a neighborhood. Here, the neighborhood is given by an $\ell_1$-ball of radius $r \geq 0$ around a distribution estimated from a limited number of samples. Our analysis can be applied when privacy and utility are measured in terms of a broad range of metrics based on $f$-divergences, or by probability of correct guessing.

Due to its natural interpretation and simplicity, we start our analysis by letting $r = 0$. This corresponds to the *pointwise* setting, where the privacy mechanism is fixed, and its performance is evaluated in terms of a single distribution learned from data. We provide bounds on the gap between the privacy-utility guarantees computed under the empirical distribution and the *de facto* guarantees for the true data distribution. This gap depends on the number of observed samples and properties of the data (e.g. support size, probability of least likely symbol), and improves and generalizes the results presented by Wang and Calmon in [4].

We then extend our analysis to the more general *uniform* setting. Here, a given level of privacy is uniformly assured for all data sets drawn from distributions within a neighborhood $r > 0$ of a target distribution (potentially learned from data). Using large deviation results, we establish upper bounds on the reduction in utility due to the uniform privacy guarantee which, in turn, depend on the value of $r$.

The paper is organized as follows. In Section II-A we recall the framework of privacy-utility trade-offs. The formal definitions for the uniform privacy guarantees are introduced in Section II-B. In Section II-C we recall basic results from large deviation theory related to the distance between the empirical and true distributions. The definitions of $f$-informations and probability of correct guessing are recalled in Section II-D. Our main results for the pointwise and uniform results are presented in Sections III-A and III-B, respectively.

## II. Problem Setup and Preliminaries

### A. Privacy-Utility Trade-Offs

Suppose that $S$ is a variable to be hidden (e.g. political preference) and $X$ is an observed variable (e.g. movie ratings) that is correlated with $S$. In order to receive some utility (e.g. personalized recommendations), we would like to disclose as much information about $X$ without compromising $S$. An approach with rigorous privacy guarantees is to release a new random variable $Y$ produced by applying a randomized mapping to $X$. This mapping, called the privacy mechanism, is designed to fulfill a certain privacy constraint.

In the sequel we assume that $S$ and $X$ are discrete and let $P_{S,X}$ denote their joint distribution. The support of $Y$ can be any discrete set. We let $\mathcal{L}(P_{S,X}, P_{Y|X})$ and $\mathcal{U}(P_{S,X}, P_{Y|X})$ be the privacy leakage and the utility generated by a mapping $P_{Y|X}$ for the underlying distribution $P_{S,X}$, respectively. Throughout this paper, specific instantiations of $\mathcal{L}$ and $\mathcal{U}$ are $f$-informations and probability of correctly guessing. The following definition captures the fundamental trade-off between privacy and utility in the present setting.

**Definition 1.** For a given joint distribution $P_{S,X}$ and $\epsilon \geq 0$, the *privacy-utility function* is defined as

$$\mathsf{H}(P_{S,X}; \epsilon) \triangleq \sup_{P_{Y|X} \in \mathcal{D}(P_{S,X}; \epsilon)} \mathcal{U}(P_{S,X}, P_{Y|X}), \qquad (1)$$

where $\mathcal{D}(P_{S,X}; \epsilon) \triangleq \{P_{Y|X} : \mathcal{L}(P_{S,X}, P_{Y|X}) \leq \epsilon\}$.

This material is based upon work supported by the National Science Foundation under Grant No. CCF-1350914 and an ASU seed grant.

This type of privacy-utility trade-off (PUT) has been investigated for several measures of privacy and utility, see, for example, [3]–[5]. When the distribution $P_{S,X}$ is known, the privacy-utility function in Definition 1 quantifies the best utility achievable by any privacy mechanism providing the desired privacy guarantee. In practice, the designer may not have access to the true distribution $P_{S,X}$, but only to independent samples $\{(s_i, x_i)\}_{i=1}^n$ drawn from this distribution. In this case, the privacy-utility guarantees for a distribution learned from the samples, say $\hat{P}_{S,X}$, and the true distribution $P_{S,X}$ might be different. For any given privacy mechanism $P_{Y|X}$, these discrepancies are effectively quantified by

$$|\mathcal{L}(P_{S,X}, P_{Y|X}) - \mathcal{L}(\hat{P}_{S,X}, P_{Y|X})|, \tag{2}$$

and

$$|\mathcal{U}(P_{S,X}, P_{Y|X}) - \mathcal{U}(\hat{P}_{S,X}, P_{Y|X})|. \tag{3}$$

### B. Uniform Privacy Guarantees

When privacy is a priority, a specific privacy guarantee for the true distribution $P_{S,X}$ may still be required, even though the designer has only access to a distribution $\hat{P}_{S,X}$ estimated from the samples $\{(s_i, x_i)\}_{i=1}^n$. We propose the following procedure to overcome this difficulty: (a) use large deviation theory results to find an upper bound, say $r$, for the distance between $\hat{P}_{S,X}$ and $P_{S,X}$; (b) provide a privacy guarantee for *all* distributions at distance less or equal than $r$ from the $\hat{P}_{S,X}$. In the sequel, we measure the distance between two probability distributions $P$ and $Q$ by their $\ell_1$-distance,

$$\|P - Q\| \triangleq \sum_{z \in \mathcal{Z}} |P(z) - Q(z)|. \tag{4}$$

With this notation, we introduce the following definition.

**Definition 2.** Given $\hat{P}_{S,X}$, $\epsilon \geq 0$, and $r \geq 0$, we define

$$P^*_{Y|X}(\hat{P}_{S,X}; \epsilon, r) \triangleq \underset{P_{Y|X} \in \mathcal{D}(\hat{P}_{S,X}; \epsilon, r)}{\arg\max} \mathcal{U}_r(\hat{P}_{S,X}, P_{Y|X}), \tag{5}$$

where $\mathcal{D}(\hat{P}_{S,X}; \epsilon, r)$ is the set of all mechanisms $P_{Y|X}$ such that $\mathcal{L}(Q_{S,X}, P_{Y|X}) \leq \epsilon$ for all $Q_{S,X}$ with $\|\hat{P}_{S,X} - Q_{S,X}\| \leq r$, and

$$\mathcal{U}_r(\hat{P}_{S,X}, P_{Y|X}) \triangleq \inf_{Q_{S,X}: \|\hat{P}_{S,X} - Q_{S,X}\| \leq r} \mathcal{U}(Q_{S,X}, P_{Y|X}). \tag{6}$$

For a given privacy mechanism $P_{Y|X}$, the infimum in (6) equals the worst case utility attained by $P_{Y|X}$ over all the distributions $Q_{S,X}$ at a distance less than or equal to $r$ from $\hat{P}_{S,X}$. Thus, by definition, $P^*_{Y|X}(\hat{P}_{S,X}; \epsilon, r)$ is the privacy mechanism with the best worst-case performance among all privacy mechanisms which ensure an $\epsilon$-privacy guarantee for *all* the distributions at a distance less than or equal to $r$ from $\hat{P}_{S,X}$. In this context, it is natural to investigate the utility degradation incurred by providing such a robust privacy guarantee. For this matter, we introduce the *uniform utility-degradation* function as follows.

**Definition 3.** Given $\hat{P}_{S,X}$, $P_{S,X}$, $\epsilon \geq 0$, and $r \geq 0$, we define

$$\Delta(P_{S,X}, \hat{P}_{S,X}; \epsilon, r) \triangleq \mathsf{H}(P_{S,X}; \epsilon) - \mathcal{U}(P_{S,X}, P^*_{Y|X}), \tag{7}$$

where $P^*_{Y|X} = P^*_{Y|X}(\hat{P}_{S,X}; \epsilon, r)$.

Note that when $r = 0$, (7) measures the utility degradation due to the mismatched estimation; while for $r > 0$, (7) quantifies the utility degradation incurred by the uniform privacy guarantee and the mismatched estimation.

### C. Distance between the Estimated and True Distribution

The distance between the learned and true distributions is the superposition of several errors, for example, estimation error, observation and sampling errors, etc. All these effects can be incorporated into the parameter $r$. Due to space constraints, here we deal only with the estimation error.

A result by Devroye [6, Lemma 3] establishes that, for every $\epsilon \geq \sqrt{20k/n}$,

$$\Pr\left(\sum_{i=1}^{k} |V_i - np_i| > n\epsilon\right) \leq 3\exp\left(-\frac{n}{25}\epsilon^2\right), \tag{8}$$

where $(V_1, \ldots, V_k)$ is a multinomial $(n, p_1, \ldots, p_k)$ random vector. Note that the empirical distribution is a (normalized) multinomial random vector. Hence, by taking $k = M \triangleq |\mathcal{S}||\mathcal{X}|$ and $\epsilon = \lambda\sqrt{20M/n}$ with $\lambda \geq 1$, Devroye's lemma implies that, with probability at least $1 - \beta_\lambda$,

$$\|\hat{P}_{S,X} - P_{S,X}\| \leq \lambda\sqrt{\frac{20M}{n}}, \tag{9}$$

where $\beta_\lambda \triangleq 3\exp(-4\lambda^2 M/5)$ and $\hat{P}_{S,X}$ is the empirical distribution obtained from $\{(s_i, x_i)\}_{i=1}^n$. Even though in this paper we focus on large deviation results, it is worth pointing out that the order $O(\sqrt{M/n})$ is present in other fundamental settings, e.g., the minimax expected loss framework in [7, Cor. 9].

### D. $f$-Informations and Probability of Correct Guessing

We briefly introduce a few definitions that will be used for privacy and utility metrics in the rest of the paper. Let $f : (0, \infty) \to \mathbb{R}$ be a convex function with $f(1) = 0$. The $f$-divergence between two probability distributions $P$ and $Q$ with $P \ll Q$ is given by [8]

$$D_f(P\|Q) \triangleq \sum_x Q(x) f\left(\frac{P(x)}{Q(x)}\right). \tag{10}$$

More recent developments about the properties of $f$-divergences can be found in [9], [10] and the references therein. With this notation, the $f$-information between two discrete random variables $U$ and $V$ is defined by

$$\begin{aligned} I_f(P_{U,V}) &\triangleq D_f(P_{U,V}\|P_U P_V) \\ &= \sum_{u,v} P_U(u) P_V(v) f\left(\frac{P_{U,V}(u,v)}{P_U(u)P_V(v)}\right). \end{aligned} \tag{11}$$

Also, the probability of correctly guessing $U$, with no additional information, is given by [11]

$$P_c(U) \triangleq \max_{u \in \mathcal{U}} \Pr(U = u). \tag{12}$$

Similarly, the probability of correctly guessing $U$ given $V$ is

$$P_c(U|V) \triangleq \sum_{v \in \mathcal{V}} \max_{u \in \mathcal{U}} \Pr(U = u, V = v). \tag{13}$$

## III. Main Results

### A. Pointwise Privacy Guarantees

Now we study the discrepancy between the guarantees provided for the empirical and true distributions by any fixed mechanism when both privacy and utility are measured using $f$-informations. For space brevity, all the results in this section are stated using the same $f$-information to measure both privacy and utility. It can be shown, *mutatis mutandis*, that they hold true also when privacy and utility are measured using different $f$-informations.

The main result of this section is based on the following two technical lemmas. Before stating them, we recall the following definitions. For a given function $g : [0,\infty) \to \mathbb{R}$ and $u > 0$, we let

$$K_{g,u} \triangleq \sup\{|g(x)|\colon x \in [0, u^{-1}]\}. \tag{14}$$

The constant $K_{g,u}$ is the so-called supremum norm of $g$ on $[0, u^{-1}]$. In addition, if $g$ is Liptschitz on $[0, u^{-1}]$, we let $L_{g,u}$ be its Lipschitz constant, i.e.,

$$\min\{L \geq 0 : |g(x)-g(y)| \leq L|x-y|, \forall x, y \in [0, u^{-1}]\}. \tag{15}$$

A function $g : [0,\infty) \to \mathbb{R}$ is called locally Lipschitz if, for every $t \geq 0$, it is Lipschitz on $[0, t]$ with a Lipschitz constant that may depend on $t$. For example, the function $g(x) = x^2$ is locally Lipschitz but not Lipschitz.

**Lemma 1.** *Suppose that $S_i \to X_i \to Y_i$ for $i = 1, 2$ and $P_{Y_1|X_1} = P_{Y_2|X_2}$. Let $m_S \triangleq \min\{P_{S_i}(s) : s \in \mathcal{S}, i \in \{1, 2\}\}$ and $m_X \triangleq \min\{P_{X_i}(x) : x \in \mathcal{X}, i \in \{1, 2\}\}$. For notational simplicity, let*

$$\Delta_L \triangleq |I_f(P_{S_1,Y_1}) - I_f(P_{S_2,Y_2})|, \tag{16}$$

$$\Delta_U \triangleq |I_f(P_{X_1,Y_1}) - I_f(P_{X_2,Y_2})|. \tag{17}$$

*If $f : [0,\infty) \to \mathbb{R}$ is locally Lipschitz, then, for all $\delta > 0$,*

$$\Delta_L \leq \begin{cases} A_f|\mathcal{S}|\delta + B_{f,\delta}\|P_{S_1,X_1} - P_{S_2,X_2}\| & m_S < \delta, \\ C_{f,m_S}||P_{S_1,X_1} - P_{S_2,X_2}|| & \delta \leq m_S, \end{cases}$$

$$\Delta_U \leq \begin{cases} A_f|\mathcal{X}|\delta + B_{f,\delta}\|P_{S_1,X_1} - P_{S_2,X_2}\| & m_X < \delta, \\ C_{f,m_X}||P_{S_1,X_1} - P_{S_2,X_2}|| & \delta \leq m_X, \end{cases}$$

*where $A_f = 4K_{f,m_X}$,*

$$B_{f,\delta} = K_{f,m_X} + 2K_{f,\delta} + (2\delta^{-1} + 1)L_{f,\delta}, \tag{18}$$

*and $C_{f,u} = 2K_{f,m_X} + (2u^{-1}+1)L_{f,m_X}$ with $u \in \{m_S, m_X\}$.*

Observe that the previous lemma implicitly assumes that $f(0) = \lim_{x\to 0^+} f(x)$ is finite. Examples of $f$-divergences satisfying the assumptions of Lemma 1 include the total variation distance, the $\chi^2$-distance, and the Hellinger distance of order $\alpha > 1$ (a one-to-one transformation of the Rényi divergence of the same order). See [12] for further examples. Note that, however, KL-divergence cannot be handled by Lemma 1 as $|\log(x)| \to \infty$ as $x \to 0^+$. Indeed, KL-divergence has a different asymptotic behavior than the one obtained in Theorem 1 below, see [13].

Due to limited sample size, not all outcomes of $X$ may be observable in the data set used to design the privacy mechanism, and can significantly impact performance depending on the metric used. Indeed, by taking $P_{S_1,X_1} = P_{S,X}$ and $P_{S_2,X_2} = \hat{P}_{S,X}$, we can see that the upper bounds for $\Delta_L$ and $\Delta_U$ in Lemma 1 become larger as $m_X$ gets smaller. In order to address this issue, we propose a pre-processing technique which combines the symbols with *less* observations. Specifically, for $\gamma \geq 0$ and $x_0$ a symbol not belonging to $\mathcal{X}$, we introduce the pre-processing technique $\Pi_\gamma$ with input alphabet $\mathcal{X}$ and output alphabet

$$\mathcal{X}_\gamma \triangleq \{x \in \mathcal{X} : \hat{P}_X(x) \geq \gamma\} \cup \{x_0\}, \tag{19}$$

determined by

$$\Pi_\gamma(x) = \begin{cases} x & \hat{P}_X(x) \geq \gamma, \\ x_0 & \text{otherwise.} \end{cases} \tag{20}$$

Consider the following lemma regarding this pre-processing technique.

**Lemma 2.** *Let $\gamma \geq 0$. If $X \to X_0 \to Y_0$ is a Markov chain with $X_0 = \Pi_\gamma(X)$. Then, for every $f$-information,*

$$I_f(P_{X,Y_0}) = I_f(P_{X_0,Y_0}). \tag{21}$$

Although this lemma may look counterintuitive at a first glance, its proof relies on the fact that the conditional distributions $P_{Y_0|X}$ and $P_{Y_0|X_0}$ are essentially the same. Specifically,

$$P_{Y_0|X}(y|x) = \begin{cases} P_{Y_0|X_0}(y|x) & x \in \{x \in \mathcal{X} : \hat{P}_X(x) \geq \gamma\}, \\ P_{Y_0|X_0}(y|x_0) & x \in \mathcal{X} \setminus \{x \in \mathcal{X} : \hat{P}_X(x) \geq \gamma\}. \end{cases}$$

The following theorem is the main result of this section. It bounds the discrepancy of the privacy-utility guarantees between the learned and true distributions.

**Theorem 1.** *Let $\gamma \geq 0$ and $\hat{P}_{S,X}$ be the empirical distribution of $n$ i.i.d. samples drawn from $P_{S,X}$. Assume that*

$$S \to X \to X_0 \to Y_0, \tag{22}$$

*where $X_0 = \Pi_\gamma(X)$ and $P_{Y_0|X_0}$ is fixed. Let $P_{S,Y_0}$ and $\hat{P}_{S,Y_0}$ be the joint distributions of $(S, Y_0)$ when the joint distributions of $(S, X)$ are $P_{S,X}$ and $\hat{P}_{S,X}$, respectively. Define $P_{X,Y_0}$ and $\hat{P}_{X,Y_0}$ in an equivalent manner. Let*

$$m_S \triangleq \min\{\{P_S(s) : s \in \mathcal{S}\} \cup \{\hat{P}_S(s) : s \in \mathcal{S}\}\},$$
$$m_X \triangleq \min\{\{P_{X_0}(x) : x \in \mathcal{X}_\gamma\} \cup \{\hat{P}_{X_0}(x) : x \in \mathcal{X}_\gamma\}\}.$$

*If $f : [0,\infty) \to \mathbb{R}$ is locally Lipschitz and $m_X \leq m_S$, then, with probability $1 - \beta_\lambda$,*

$$|I_f(\hat{P}_{S,Y_0}) - I_f(P_{S,Y_0})| \leq C_{f,m_S}\lambda\sqrt{\frac{20M}{n}}, \tag{23}$$

$$|I_f(\hat{P}_{X_0,Y_0}) - I_f(P_{X,Y_0})| \leq C_{f,m_X}\lambda\sqrt{\frac{20M}{n}}, \tag{24}$$

*where $M = |\mathcal{S}||\mathcal{X}|$, $\beta_\lambda = 3\exp(-4\lambda^2 M/5)$ with $\lambda \geq 1$ and $C_{f,u}$ is defined in Lemma 1.*

**Proof of Theorem 1:** We first apply Lemma 1 with $\delta = m_X$, $P_{S_1,X_1} = P_{S,X_0}$, and $P_{S_2,X_2} = \hat{P}_{S,X_0}$. In particular, we obtain that

$$|I_f(\hat{P}_{S,Y_0}) - I_f(P_{S,Y_0})| \leq C_{f,m_S}\|\hat{P}_{S,X_0} - P_{S,X_0}\|, \tag{25}$$

$$|I_f(\hat{P}_{X_0,Y_0}) - I_f(P_{X_0,Y_0})| \leq C_{f,m_X}\|\hat{P}_{S,X_0} - P_{S,X_0}\|. \tag{26}$$

By the data processing inequality, we have

$$\|\hat{P}_{S,X_0} - P_{S,X_0}\| \leq \|\hat{P}_{S,X} - P_{S,X}\|. \tag{27}$$

By the inequality (9), with probability at least $1-\beta_\lambda$,

$$\|\hat{P}_{S,X} - P_{S,X}\| \leq \lambda\sqrt{\frac{20M}{n}}, \tag{28}$$

where $\beta_\lambda = 3\exp(-4\lambda^2 M/5)$ and $\lambda \geq 1$. Hence,

$$|I_f(\hat{P}_{S,Y_0}) - I_f(P_{S,Y_0})| \leq C_{f,m_S}\lambda\sqrt{\frac{20M}{n}}, \tag{29}$$

$$|I_f(\hat{P}_{X_0,Y_0}) - I_f(P_{X_0,Y_0})| \leq C_{f,m_X}\lambda\sqrt{\frac{20M}{n}}. \tag{30}$$

By Lemma 2, we have that

$$I_f(P_{X,Y_0}) = I_f(P_{X_0,Y_0}). \tag{31}$$

The result follows. ∎

### B. Uniform Privacy Guarantees

Let $\mathbb{P}$ denote the set of all probability distributions over $\mathcal{S}\times\mathcal{X}$. Assume that prior information about the joint distribution of $S$ and $X$ is available. In this case, we let $\mathbb{Q}\subseteq\mathbb{P}$ be all the joint distributions compatible with the prior knowledge. For a given $\hat{P}_{S,X}\in\mathbb{P}$ and $r\geq 0$, we define

$$\mathbb{Q}_r(\hat{P}_{S,X}) \triangleq \{Q_{S,X}\in\mathbb{Q} : \|Q_{S,X} - \hat{P}_{S,X}\| \leq r\}. \tag{32}$$

In this setting, a natural modification for the uniform privacy mechanism $P^*_{Y|X}$ is the following. Given $\hat{P}_{S,X}\in\mathbb{Q}$, $\epsilon\geq 0$, and $r\geq 0$, let

$$P^*_{Y|X}(\hat{P}_{S,X};\epsilon,r) \triangleq \underset{P_{Y|X}\in\mathcal{D}_{\mathbb{Q}}(\hat{P}_{S,X};\epsilon,r)}{\arg\max} \mathcal{U}_r(\hat{P}_{S,X},P_{Y|X}) \tag{33}$$

where

$$\mathcal{D}_{\mathbb{Q}}(\hat{P}_{S,X};\epsilon,r) \triangleq \bigcap_{Q_{S,X}\in\mathbb{Q}_r(\hat{P}_{S,X})} \left\{P_{Y|X} : \mathcal{L}(Q_{S,X},P_{Y|X}) \leq \epsilon\right\},$$

$$\mathcal{U}_r(\hat{P}_{S,X},P_{Y|X}) \triangleq \inf_{Q_{S,X}\in\mathbb{Q}_r(\hat{P}_{S,X})} \mathcal{U}(Q_{S,X},P_{Y|X}).$$

Finally, recall that

$$\Delta_{\mathbb{Q}}(P_{S,X},\hat{P}_{S,X};\epsilon,r) \triangleq \mathsf{H}(P_{S,X};\epsilon) - \mathcal{U}(P_{S,X},P^*_{Y|X}). \tag{34}$$

In order to simplify the notation, in what follows we denote $\hat{P}_{S,X}$, $P_{S,X}$, and $Q_{S,X}$ by $\hat{P}$, $P$ and $Q$, respectively. For the privacy measures under consideration, $f$-information and probability of correct guessing, it has been proved that the optimal privacy mechanism for the PUT in Definition 1 requires an alphabet of size $|\mathcal{X}|+1$, see [5], [14] and references therein. With this in mind, let $\mathbb{F}$ be the set of all row stochastic matrices of dimension $|\mathcal{X}|\times(|\mathcal{X}|+1)$. Note that the set $\mathbb{F}$ models all privacy mechanisms $P_{Y|X}$ with $|\mathcal{Y}|\leq|\mathcal{X}|+1$. In this case, the privacy-utility function in Definition 1 equals

$$\mathsf{H}(P;\epsilon) = \sup_{\substack{F\in\mathbb{F}\\ \mathcal{L}(P,F)\leq\epsilon}} \mathcal{U}(P,F), \tag{35}$$

for all $P\in\mathbb{P}$. Note that, in principle, the robust privacy mechanism in Definition 2, or the version in (33), may require the use of more than $|\mathcal{X}|+1$ output symbols. However, the following lower bound for $\mathcal{U}(P,P^*_{Y|X})$, which can be computed using mechanisms with $|\mathcal{X}|+1$ output symbols, will be enough for our purposes,

$$\mathcal{U}(P,P^*_{Y|X}) \geq \sup_{F\in\mathbb{D}_{\mathbb{Q}}(\hat{P};\epsilon,r)} \inf_{Q\in\mathbb{Q}_r(\hat{P})} \mathcal{U}(Q,F), \tag{36}$$

where $\hat{P}\in\mathbb{Q}$ and

$$\mathbb{D}_{\mathbb{Q}}(\hat{P};\epsilon,r) \triangleq \bigcap_{Q\in\mathbb{Q}_r(\hat{P})} \{F\in\mathbb{F} : \mathcal{L}(Q,F)\leq\epsilon\} \subseteq \mathbb{F}. \tag{37}$$

The main result of this section provides an upper bound for $\Delta_{\mathbb{Q}}$ whenever the leakage and utility functions satisfy a Hölder-like condition. Recall that a function $f:\mathbb{R}\to\mathbb{R}$ is said to be Hölder continuous of order $\alpha\in[0,1]$ if there exists $K\geq 0$ such that $|f(x)-f(y)|\leq K|x-y|^\alpha$ for all $x,y\in\mathbb{R}$.

**Theorem 2.** *Assume that $\mathbb{Q}\subseteq\mathbb{P}$ is a closed set and that for every $Q\in\mathbb{Q}$ the functions*

$$F\mapsto\mathcal{L}(Q,F) \quad \textit{and} \quad F\mapsto\mathcal{U}(Q,F), \tag{38}$$

*are continuous and that* (35) *holds true. Furthermore, assume that for a given $\hat{P}\in\mathbb{Q}$ there exist positive constants $r_0$, $\alpha$, $C_L$, and $C_U$ such that*

$$|\mathcal{L}(\hat{P},F) - \mathcal{L}(Q,F)| \leq C_L\|\hat{P}-Q\|^\alpha, \tag{39}$$

$$|\mathcal{U}(\hat{P},F) - \mathcal{U}(Q,F)| \leq C_U\|\hat{P}-Q\|^\alpha, \tag{40}$$

*for all $Q\in\mathbb{Q}_{r_0}(\hat{P})$ and all $F\in\mathbb{F}$. If $P\in\mathbb{Q}_{r_0}(\hat{P})$, then, for all $\epsilon>0$ and all $r^\alpha\leq\min\{r_0^\alpha, (\epsilon-\min_F\mathcal{L}(\hat{P},F))/C_L\}$,*

$$\Delta_{\mathbb{Q}}(P,\hat{P};\epsilon,r) \leq \mathsf{H}(\hat{P};\epsilon+C_Lr^\alpha) - \mathsf{H}(\hat{P};\epsilon-C_Lr^\alpha) + 2C_Ur^\alpha. \tag{41}$$

**Remark 1.** Under the assumptions of Theorem 2, if $\mathsf{H}(\hat{P};\cdot)$ is Lipschitz continuous with Lipschitz constant $L$ then

$$\Delta_{\mathbb{Q}}(P,\hat{P};\epsilon,r) \leq 2(C_U + LC_L)r^\alpha. \tag{42}$$

The assumptions in (38), (39) and (40) might seem restrictive at a first glance. Nonetheless, as shown in the following, they hold true for our measures of interest: $f$-informations and probability of correct guessing.

*1) $f$-Divergences:* Assume that both privacy and utility are measured by an $f$-information for a given convex function $f:(0,\infty)\to\mathbb{R}$ with $f(1)=0$, i.e.,

$$\mathcal{L}(Q_{S,X},P_{Y|X}) \triangleq I_f(Q_{S,Y}), \tag{43}$$

$$\mathcal{U}(Q_{S,X},P_{Y|X}) \triangleq I_f(Q_{X,Y}). \tag{44}$$

A standard convexity argument, see, e.g., [14], shows that

$$\mathsf{H}(Q_{S,X};\epsilon) \triangleq \sup_{\substack{S\to X\to Y\\ I_f(Q_{S,Y})\leq\epsilon}} I_f(Q_{X,Y}) \tag{45}$$

admits the expression in (35), i.e., it is enough to consider privacy mechanisms taking values on $\mathcal{Y}=\{1,\ldots,|\mathcal{X}|+1\}$.

**Example 1.** Consider the parametric case where

$$\mathbb{Q} \triangleq \left\{Q\in\mathbb{P} : \sum_{x\in\mathcal{X}} Q(s,x)\geq\gamma \text{ for all } s\in\mathcal{S}\right\} \cap \left\{Q\in\mathbb{P} : \sum_{s\in\mathcal{S}} Q(s,x)\geq\gamma \text{ for all } x\in\mathcal{X}\right\} \tag{46}$$

for some $\gamma > 0$. Note that this corresponds to the case in which $S$ and $X$ have full support and their marginal distributions are bounded away from zero. In this case, Lemma 1 implies that the assumptions of Theorem 2 are satisfied with $r_0 = \infty$, $\alpha = 1$, and

$$C_L = C_U = 2K_{f,\gamma} + (2\gamma^{-1} + 1)L_{f,\gamma}. \tag{47}$$

In particular, if $f(x) = |x - 1|$, then

$$C_L = C_U \leq 4\gamma^{-1} + 1; \tag{48}$$

and if $f(x) = x^2 - 1$, then

$$C_L = C_U \leq 8\gamma^{-2}. \tag{49}$$

*2) Probability of Correct Guessing:* For ease of notation, let $\mathcal{Y} = \{1, \ldots, |\mathcal{X}|+1\}$. In the setting of the PUT, let

$$\mathcal{L}(Q, F) = \sum_{y\in\mathcal{Y}} \max_{s\in\mathcal{S}} \sum_{x\in\mathcal{X}} Q(s, x)F(x, y), \tag{50}$$

$$\mathcal{U}(Q, F) = \sum_{y\in\mathcal{Y}} \max_{x\in\mathcal{X}} \sum_{s\in\mathcal{S}} Q(s, x)F(x, y). \tag{51}$$

The above choice corresponds to the case when the measures of privacy and utility are $P_c(S|Y)$ and $P_c(X|Y)$, respectively. In particular, for each $\epsilon \in [P_c(S), P_c(S|X)]$,

$$\mathsf{H}(P; \epsilon) = \sup_{\substack{S\to X\to Y \\ P_c(S|Y)\leq\epsilon}} P_c(X|Y). \tag{52}$$

This privacy-utility trade-off based on the probability of correctly guessing was recently studied by Asoodeh et al [5]. In this case, it is possible to verify that $\mathcal{L}(P, \cdot)$ and $\mathcal{U}(P, \cdot)$ are continuous.

For $\hat{P}, Q \in \mathbb{P}$ and $F \in \mathbb{F}$, let $\Delta_L \triangleq |\mathcal{L}(\hat{P}, F) - \mathcal{L}(Q, F)|$. It can be verified that, for $a_i, b_i \geq 0$,

$$|\max_i a_i - \max_i b_i| \leq \max_i |a_i - b_i|, \tag{53}$$

and, in particular,

$$\Delta_L \leq \sum_{y\in\mathcal{Y}} \left| \max_{s\in\mathcal{S}} (\hat{P}F)(s, y) - \max_{s\in\mathcal{S}} (QF)(s, y) \right| \tag{54}$$

$$\leq \sum_{y\in\mathcal{Y}} \max_{s\in\mathcal{S}} \left| (\hat{P}F)(s, y) - (QF)(s, y) \right|. \tag{55}$$

Note that $(QF)(s, y) = \sum_x Q(s, x)F(x, y)$. Thus, a straightforward manipulation shows that

$$\Delta_L \leq \sum_{y\in\mathcal{Y}} \max_{s\in\mathcal{S}} \sum_{x\in\mathcal{X}} |\hat{P}(s, x) - Q(s, x)|F(x, y) \tag{56}$$

$$\leq \sum_{s\in\mathcal{S}} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} |\hat{P}(s, x) - Q(s, x)|F(x, y) \tag{57}$$

$$\leq \|\hat{P} - Q\|. \tag{58}$$

Similarly, it can be shown that

$$|\mathcal{U}(\hat{P}, F) - \mathcal{U}(Q, F)| \leq \|\hat{P} - Q\|. \tag{59}$$

Hence, the probability of correct guessing satisfies the assumptions of Theorem 2 with $r_0 = \infty$, $\alpha = 1$, $C_L = C_U = 1$.

**Example 2.** For $p, q \in [0, 1]$, we let

$$p\#q = \begin{pmatrix} (1-p)(1-q) & (1-p)q \\ pq & p(1-q) \end{pmatrix}, \tag{60}$$

and $\mathbb{Q} = \{p\#q : p \in [1/2, 1], q \in [0, 1/2], p + q \leq 1\}$. This selection of $\mathbb{Q}$ captures the case when $S$ is assumed to be a Bernoulli random variable with $\Pr(S = 1) = p$ and the channel between $S$ and $X$ is a binary symmetric channel with crossover probability $q$. By Theorem 2 in [5], for all $Q \in \mathbb{Q}$,

$$\mathsf{H}(Q; \epsilon) = 1 - \frac{1-q}{p-q}(p + q - 2pq) + \epsilon\frac{p + q - 2pq}{p - q}, \tag{61}$$

whenever $\epsilon \in [p, 1 - q]$. Hence, the bound in (42) becomes

$$\Delta_{\mathbb{Q}}(P, \hat{P}; \epsilon, r) \leq \frac{2\hat{p}(1-\hat{q})}{\hat{p} - \hat{q}} r, \tag{62}$$

where $\hat{P} \triangleq \hat{p}\#\hat{q} \in \mathbb{Q}$. By (9), for $\lambda \geq 1$, with probability at least $1 - \beta_\lambda$,

$$\Delta_{\mathbb{Q}}(P_{S,X}, \hat{P}_{S,X}; \epsilon, 4\lambda\sqrt{5/n}) \leq \frac{8\hat{p}(1-\hat{q})}{\hat{p} - \hat{q}} \lambda\sqrt{\frac{5}{n}}, \tag{63}$$

where $\beta_\lambda \triangleq 3\exp(-16\lambda^2/5)$.

## REFERENCES


[1] F. du Pin Calmon and N. Fawaz, "Privacy against statistical inference," in *Proc. of 50th IEEE Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2012, pp. 1401–1408.

[2] L. Sankar, S. R. Rajagopalan, and H. V. Poor, "Utility-privacy tradeoffs in databases: An information-theoretic approach," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 6, pp. 838–852, 2013.

[3] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Information extraction under privacy constraints," *Information*, vol. 7, no. 1, p. 15, 2016.

[4] H. Wang and F. P. Calmon, "An estimation-theoretic view of privacy," in *Proc. of 55th IEEE Annual Allerton Conference on Communication, Control, and Computing*. IEEE, 2017.

[5] S. Asoodeh, M. Diaz, F. Alajaji, and T. Linder, "Estimation efficiency under privacy constraints," *arXiv preprint arXiv:1707.02409*, 2017.

[6] L. Devroye, "The equivalence of weak, strong and complete convergence in $l_1$ for kernel density estimates," *The Annals of Statistics*, pp. 896–904, 1983.

[7] S. Kamath, A. Orlitsky, D. Pichapati, and A. T. Suresh, "On learning distributions from their samples," in *Conference on Learning Theory*, 2015, pp. 1066–1100.

[8] I. Ciszar, "Information-type measures of difference of probability distributions and indirect observations," *Studia Sci. Math. Hungar.*, vol. 2, pp. 299–318, 1967.

[9] M. Raginsky, "Strong data processing inequalities and Φ-sobolev inequalities for discrete channels," *IEEE Transactions on Information Theory*, vol. 62, no. 6, pp. 3355–3389, 2016.

[10] F. P. Calmon, Y. Polyanskiy, and Y. Wu, "Strong data processing inequalities for input constrained additive noise channels," *IEEE Transactions on Information Theory*, 2017.

[11] S. Fehr and S. Berens, "On the conditional rényi entropy," *IEEE Transactions on Information Theory*, vol. 60, no. 11, pp. 6801–6810, 2014.

[12] I. Sason and S. Verdú, "$f$-divergence inequalities," *IEEE Transactions on Information Theory*, vol. 62, no. 11, pp. 5973–6006, 2016.

[13] O. Shamir, S. Sabato, and N. Tishby, "Learning and generalization with the information bottleneck," *Theoretical Computer Science*, vol. 411, no. 29-30, pp. 2696–2711, 2010.

[14] H. Witsenhausen and A. Wyner, "A conditional entropy bound for a pair of discrete random variables," *IEEE Transactions on Information Theory*, vol. 21, no. 5, pp. 493–501, 1975.

## APPENDIX A
## PROOF OF LEMMA 1

The following auxiliary lemma will be used in the proof of Lemma 1.

**Lemma 3.** *Let $S$, $X$ and $Y$ be random variables supported over finite alphabets $\mathcal{S}$, $\mathcal{X}$ and $\mathcal{Y}$, respectively. Assume that $S \to X \to Y$ form a Markov chain in that order. Then, for all $s \in \mathcal{S}$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$,*

$$\max\left\{\frac{P_{S,Y}(s,y)}{P_S(s)P_Y(y)}, \frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}\right\} \leq \left(\min_{x\in\mathcal{X}} P_X(x)\right)^{-1}. \tag{64}$$

*Proof:* Recall that

$$\frac{\sum_i a_i}{\sum_i b_i} \leq \max_i \frac{a_i}{b_i}, \tag{65}$$

whenever $a_i \geq 0$ and $b_i > 0$. For a given $y \in \mathcal{Y}$, let $\mathcal{X}_y \triangleq \{x \in \mathcal{X} : P_{X,Y}(x,y) > 0\}$. Note that, given $s \in \mathcal{S}$ and $y \in \mathcal{Y}$,

$$\frac{P_{S,Y}(s,y)}{P_S(s)P_Y(y)} = \frac{\sum_{x\in\mathcal{X}_y} P_{S,X,Y}(s,x,y)}{\sum_{x\in\mathcal{X}_y} P_S(s)P_{X,Y}(x,y)} \tag{66}$$

$$\leq \max_{x\in\mathcal{X}_y} \frac{P_{S,X,Y}(s,x,y)}{P_S(s)P_{X,Y}(x,y)} \tag{67}$$

$$= \max_{x\in\mathcal{X}_y} \frac{P_{S,X}(s,x)}{P_S(s)P_X(x)} \tag{68}$$

$$\leq \max_{x\in\mathcal{X}} \frac{1}{P_X(x)} = \left(\min_{x\in\mathcal{X}} P_X(x)\right)^{-1}, \tag{69}$$

where the last inequality follows from the fact that $\mathcal{X}_y \subseteq \mathcal{X}$ and $P_{S,X}(s,x) \leq P_S(s)$ for all $s \in \mathcal{S}$ and $x \in \mathcal{X}$. The rest of the lemma is similar. ∎

**Proof of Lemma 1:** First, let's assume that $m_S < \delta$. Let $\mathcal{S}_i = \{s \in \mathcal{S} : P_{S_i}(s) < \delta\}$ for each $i \in \{1,2\}$ and $\mathcal{S}_+ = \mathcal{S} \setminus (\mathcal{S}_1 \cup \mathcal{S}_2)$. By the definition of $f$-information and the triangle inequality, we have that

$$\Delta_L = |I_f(P_{S_1,Y_1}) - I_f(P_{S_2,Y_2})| \leq \mathrm{I} + \mathrm{II} + \mathrm{III}, \tag{70}$$

where

$$\mathrm{I} = \sum_{s\in\mathcal{S}_1}\sum_{y\in\mathcal{Y}} P_{S_1}(s)P_{Y_1}(y)\left|f\left(\frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)}\right)\right| \tag{71}$$

$$+ \sum_{s\in\mathcal{S}_1}\sum_{y\in\mathcal{Y}} P_{S_2}(s)P_{Y_2}(y)\left|f\left(\frac{P_{S_2,Y_2}(s,y)}{P_{S_2}(s)P_{Y_2}(y)}\right)\right|, \tag{72}$$

$$\mathrm{II} = \sum_{s\in\mathcal{S}_2}\sum_{y\in\mathcal{Y}} P_{S_1}(s)P_{Y_1}(y)\left|f\left(\frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)}\right)\right| \tag{73}$$

$$+ \sum_{s\in\mathcal{S}_2}\sum_{y\in\mathcal{Y}} P_{S_2}(s)P_{Y_2}(y)\left|f\left(\frac{P_{S_2,Y_2}(s,y)}{P_{S_2}(s)P_{Y_2}(y)}\right)\right|, \tag{74}$$

$$\mathrm{III} = \left|\sum_{s\in\mathcal{S}_+}\sum_{y\in\mathcal{Y}}\left(P_{S_1}(s)P_{Y_1}(y)f\left(\frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)}\right)\right.\right. \tag{75}$$

$$\left.\left. - P_{S_2}(s)P_{Y_2}(y)f\left(\frac{P_{S_2,Y_2}(s,y)}{P_{S_2}(s)P_{Y_2}(y)}\right)\right)\right|. \tag{76}$$

By Lemma 3, we have that

$$\max\left\{\frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)}, \frac{P_{S_2,Y_2}(s,y)}{P_{S_2}(s)P_{Y_2}(y)}\right\} \leq m_X^{-1}. \tag{77}$$

In particular, we have that

$$\mathrm{I} \leq K_{f,m_X}(P_{S_1}(\mathcal{S}_1) + P_{S_2}(\mathcal{S}_1)). \tag{78}$$

Since $P_{S_1}(\mathcal{S}_1) + P_{S_2}(\mathcal{S}_1) = P_{S_2}(\mathcal{S}_1) - P_{S_1}(\mathcal{S}_1) + 2P_{S_1}(\mathcal{S}_1)$, the definition of $\mathcal{S}_1$ implies that

$$P_{S_1}(\mathcal{S}_1) + P_{S_2}(\mathcal{S}_1) \leq \frac{1}{2}\|P_{S_1} - P_{S_2}\| + 2|\mathcal{S}|\delta. \tag{79}$$

Note that

$$\max\{\|P_{S_1} - P_{S_2}\|, \|P_{X_1} - P_{X_2}\|\} \leq \|P_{S_1,X_1} - P_{S_2,X_2}\|. \tag{80}$$

Hence, (78) and (79) lead to

$$\mathrm{I} \leq 2K_{f,m_X}|\mathcal{S}|\delta + \frac{K_{f,m_X}}{2}\|P_{S_1,X_1} - P_{S_2,X_2}\|. \tag{81}$$

Using a similar argument, we conclude that

$$\mathrm{II} \leq 2K_{f,m_X}|\mathcal{S}|\delta + \frac{K_{f,m_X}}{2}\|P_{S_1,X_1} - P_{S_2,X_2}\|. \tag{82}$$

By the triangle inequality $\mathrm{III} \leq \mathrm{III}_1 + \mathrm{III}_2$, where

$$\mathrm{III}_1 = \sum_{s\in\mathcal{S}_+}\sum_{y\in\mathcal{Y}} |P_{S_1}(s)P_{Y_1}(y) - P_{S_2}(s)P_{Y_2}(y)| \times \left|f\left(\frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)}\right)\right|, \tag{83}$$

$$\mathrm{III}_2 = \sum_{s\in\mathcal{S}_+}\sum_{y\in\mathcal{Y}} P_{S_2}(s)P_{Y_2}(y) \times \left|f\left(\frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)}\right) - f\left(\frac{P_{S_2,Y_2}(s,y)}{P_{S_2}(s)P_{Y_2}(y)}\right)\right|. \tag{84}$$

By the definition of $\mathcal{S}_+$, we have that, for all $s \in \mathcal{S}_+$ and $y \in \mathcal{Y}$,

$$\frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)} = \frac{P_{S_1|Y_1}(s|y)}{P_{S_1}(s)} \tag{85}$$

$$\leq \frac{1}{P_{S_1}(s)} \leq \delta^{-1}. \tag{86}$$

Recall that $|f(x)| \leq K_{f,\delta}$ for all $x \in [0, \delta^{-1}]$. Hence,

$$\begin{aligned}\mathrm{III}_1 &\leq K_{f,\delta}\sum_{s\in\mathcal{S}_+}\sum_{y\in\mathcal{Y}}|P_{S_1}(s)P_{Y_1}(y) - P_{S_2}(s)P_{Y_2}(y)| \\ &\leq K_{f,\delta}\sum_{s\in\mathcal{S}}\sum_{y\in\mathcal{Y}}(P_{Y_1}(y)|P_{S_1}(s) - P_{S_2}(s)| \\ &\qquad + P_{S_2}(s)|P_{Y_1}(y) - P_{Y_2}(y)|) \\ &= K_{f,\delta}\left(\|P_{S_1} - P_{S_2}\| + \|P_{Y_1} - P_{Y_2}\|\right).\end{aligned} \tag{87}$$

The data processing inequality implies that

$$\|P_{Y_1} - P_{Y_2}\| \leq \|P_{X_1} - P_{X_2}\|, \tag{88}$$

and by (80) we obtain that

$$\mathrm{III}_1 \leq 2K_{f,\delta}\|P_{S_1,X_1} - P_{S_2,X_2}\|. \tag{89}$$

Similarly, for all $s \in \mathcal{S}_+$ and $y \in \mathcal{Y}$ we have that

$$\max\left\{\frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)}, \frac{P_{S_2,Y_2}(s,y)}{P_{S_2}(s)P_{Y_2}(y)}\right\} \leq \delta^{-1}. \tag{90}$$

Recall that $f$ is Lipschitz on $[0, \delta^{-1}]$ and $L_{f,\delta}$ is its Lipschitz constant. In particular,

$$\begin{aligned}\mathrm{III}_2 \leq L_{f,\delta} \sum_{s\in\mathcal{S}_+}\sum_{y\in\mathcal{Y}} &\frac{1}{P_{S_1}(s)P_{Y_1}(y)} \\ &\times |P_{S_2}(s)P_{Y_2}(y)P_{S_1,Y_1}(s,y) \\ &\quad - P_{S_1}(s)P_{Y_1}(y)P_{S_2,Y_2}(s,y)|.\end{aligned} \tag{91}$$

By the triangle inequality,

$$|P_{S_2}(s)P_{Y_2}(y)P_{S_1,Y_1}(s,y) - P_{S_1}(s)P_{Y_1}(y)P_{S_2,Y_2}(s,y)| \tag{92}$$

is upper bounded by

$$\begin{aligned}&P_{S_1,Y_1}(s,y)|P_{S_2}(s)P_{Y_2}(y) - P_{S_1}(s)P_{Y_1}(y)| \\ &\quad + P_{S_1}(s)P_{Y_1}(y)|P_{S_1,Y_1}(s,y) - P_{S_2,Y_2}(s,y)|.\end{aligned} \tag{93}$$

Since $P_{S_1,Y_1}(s,y) \leq P_{Y_1}(y)$ for all $s \in \mathcal{S}$ and $y \in \mathcal{Y}$, (91) leads to

$$\begin{aligned}\mathrm{III}_2 &\leq L_{f,\delta} \sum_{s\in\mathcal{S}_+}\sum_{y\in\mathcal{Y}} \frac{1}{P_{S_1}(s)}|P_{S_2}(s)P_{Y_2}(y) - P_{S_1}(s)P_{Y_1}(y)| \\ &\quad + L_{f,\delta} \sum_{s\in\mathcal{S}_+}\sum_{y\in\mathcal{Y}} |P_{S_1,Y_1}(s,y) - P_{S_2,Y_2}(s,y)| \\ &\leq \delta^{-1} L_{f,\delta}\Big(\|P_{Y_1} - P_{Y_2}\| + \|P_{S_1} - P_{S_2}\|\Big) \\ &\quad + L_{f,\delta}\|P_{S_1,Y_1} - P_{S_2,Y_2}\|.\end{aligned} \tag{94}$$

By assumption, $P_{Y_1|X_1} = P_{Y_2|X_2}$ and hence

$$\|P_{S_1,Y_1} - P_{S_2,Y_2}\| \leq \|P_{S_1,X_1} - P_{S_2,X_2}\|. \tag{95}$$

Therefore,

$$\mathrm{III}_2 \leq \left(2\delta^{-1} + 1\right) L_{f,\delta}\|P_{S_1,X_1} - P_{S_2,X_2}\|. \tag{96}$$

Since $\Delta_L \leq \mathrm{I} + \mathrm{II} + \mathrm{III}_1 + \mathrm{III}_2$, we obtain the upper bound

$$\Delta_L \leq 4K_{f,m_X}|\mathcal{S}|\delta + B_{f,\delta}\|P_{S_1,X_1} - P_{S_2,X_2}\|, \tag{97}$$

where $B_{f,\delta} = K_{f,m_X} + 2K_{f,\delta} + (2\delta^{-1} + 1)L_{f,\delta}$.

Now assume that $\delta \leq m_S$. In particular, we have that $\mathcal{S}_1 = \mathcal{S}_2 = \emptyset$ and hence $\Delta_L \leq \mathrm{III}_1 + \mathrm{III}_2$ with $\mathrm{III}_1$ and $\mathrm{III}_2$ defined as in (83) and (84), respectively. By Lemma 3,

$$\max\left\{\frac{P_{S_1,Y_1}(s,y)}{P_{S_1}(s)P_{Y_1}(y)}, \frac{P_{S_2,Y_2}(s,y)}{P_{S_2}(s)P_{Y_2}(y)}\right\} \leq m_X^{-1}. \tag{98}$$

In particular, we have that

$$\mathrm{III}_1 \leq 2K_{f,m_X}\|P_{S_1,X_1} - P_{S_2,X_2}\|. \tag{99}$$

$$\mathrm{III}_2 \leq \left(2m_S^{-1} + 1\right) L_{f,m_X}\|P_{S_1,X_1} - P_{S_2,X_2}\|. \tag{100}$$

Hence,

$$\Delta_L \leq C_{f,m_S}\|P_{S_1,X_1} - P_{S_2,X_2}\|, \tag{101}$$

where $C_{f,m_S} = 2K_{f,m_X} + (2m_S^{-1} + 1)L_{f,m_X}$.

Mutatis mutandis, setting $\mathcal{X}_i = \{x \in \mathcal{X} : P_{X_i}(x) < \delta\}$ for each $i \in \{1,2\}$ and $\mathcal{X}_+ = \mathcal{X} \setminus (\mathcal{X}_1 \cup \mathcal{X}_2)$, it can be shown that

$$\Delta_U \leq 4K_{f,m_X}|\mathcal{X}|\delta + B_{f,\delta}\|P_{S_1,X_1} - P_{S_2,X_2}\|, \tag{102}$$

when $m_X < \delta$. If $\delta \leq m_X$, it can be shown that

$$\Delta_U \leq C_{f,m_X}\|P_{S_1,X_1} - P_{S_2,X_2}\|, \tag{103}$$

where $C_{f,m_X} = 2K_{f,m_X} + (2m_X^{-1} + 1)L_{f,m_X}$. ∎

## Appendix B
## Proof of Lemma 2

**Proof of Lemma 2:** First, we define $\mathcal{X}_1 \triangleq \{x \in \mathcal{X} : \hat{P}_X(x) \geq \gamma\}$. By the construction of $X_0$, we have that $P_{X_0|X}(x'|x) = 1$ whenever $x \in \mathcal{X}_1$ and $x' = x$, or $x \in \mathcal{X}_1^c$ and $x' = x_0$; in all other cases $P_{X_0|X}(x'|x) = 0$. In particular,

$$P_{X_0}(x_0) = \sum_{x\in\mathcal{X}_1^c} P_X(x) \text{ and } P_{X_0}(x) = P_X(x) \text{ for } x \in \mathcal{X}_1. \tag{104}$$

By the law of total probability and Markovianity

$$P_{X,Y_0}(x,y) = \sum_{x'\in\mathcal{X}_1\cup\{x_0\}} P_X(x)P_{X_0|X}(x'|x)P_{Y_0|X_0}(y|x'), \tag{105}$$

for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. In particular, for all $x \in \mathcal{X}_1$,

$$P_{X,Y_0}(x,y) = P_X(x)P_{Y_0|X_0}(y|x). \tag{106}$$

Similarly, for $x \in \mathcal{X}_1^c$,

$$P_{X,Y_0}(x,y) = P_X(x)P_{Y_0|X_0}(y|x_0). \tag{107}$$

By the definition of $f$-information, (104), (106), and (107),

$$I_f(P_{X,Y_0}) = \sum_{x\in\mathcal{X}_1}\sum_{y\in\mathcal{Y}} P_X(x)P_{Y_0}(y) f\left(\frac{P_{X,Y_0}(x,y)}{P_X(x)P_{Y_0}(y)}\right) \tag{108}$$

$$+ \sum_{x\in\mathcal{X}_1^c}\sum_{y\in\mathcal{Y}} P_X(x)P_{Y_0}(y) f\left(\frac{P_{X,Y_0}(x,y)}{P_X(x)P_{Y_0}(y)}\right) \tag{109}$$

$$= \sum_{x\in\mathcal{X}_1}\sum_{y\in\mathcal{Y}} P_{X_0}(x)P_{Y_0}(y) f\left(\frac{P_{X_0,Y_0}(x,y)}{P_{X_0}(x)P_{Y_0}(y)}\right) \tag{110}$$

$$+ \sum_{y\in\mathcal{Y}} P_{X_0}(x_0)P_{Y_0}(y) f\left(\frac{P_{X_0,Y_0}(x_0,y)}{P_{X_0}(x_0)P_{Y_0}(y)}\right) \tag{111}$$

$$= I_f(P_{X_0,Y_0}), \tag{112}$$

as required. ∎

## Appendix C
## Proof of Theorem 2

**Proof of Theorem 2:** First we show that, for all $r$ and $\epsilon$ with $P \in \mathbb{Q}_r(\hat{P})$,

$$\mathsf{H}(P;\epsilon) \leq \mathsf{H}(\hat{P};\epsilon + C_L r^\alpha) + C_U r^\alpha. \tag{113}$$

Since, for fixed $P$, both $\mathcal{L}(P,\cdot)$ and $\mathcal{U}(P,\cdot)$ are continuous, there exists $F \in \mathbb{F}$ such that $\mathcal{L}(P,F) \leq \epsilon$ and

$$\mathsf{H}(P;\epsilon) = \mathcal{U}(P,F). \tag{114}$$

By assumption, we have that

$$|\mathcal{U}(\hat{P},F) - \mathcal{U}(P,F)| \leq C_U\|\hat{P} - P\|^\alpha \leq C_U r^\alpha. \tag{115}$$

In particular,

$$\mathsf{H}(P;\epsilon) \leq \mathcal{U}(\hat{P},F) + C_U r^\alpha. \tag{116}$$

Similarly, since

$$|\mathcal{L}(\hat{P},F) - \mathcal{L}(P,F)| \leq C_L\|\hat{P} - P\|^\alpha \leq C_L r^\alpha, \tag{117}$$

we have

$$\mathcal{L}(\hat{P}, F) \leq \mathcal{L}(P, F) + C_L r^\alpha \leq \epsilon + C_L r^\alpha. \tag{118}$$

Therefore, from inequality (116) and Definition 1, we have

$$\mathsf{H}(P; \epsilon) \leq \mathsf{H}(\hat{P}; \epsilon + C_L r^\alpha) + C_U r^\alpha. \tag{119}$$

Next, we prove that

$$\mathsf{H}(\hat{P}; \epsilon - C_L r^\alpha) - C_U r^\alpha \leq \mathcal{U}(P, F^*) \tag{120}$$

where we denote $P^*_{Y|X}$ by $F^*$. Let $F_0 \in \mathbb{F}$ be such that $\mathcal{L}(\hat{P}, F_0) \leq \epsilon - C_L r^\alpha$ and

$$\mathcal{U}(\hat{P}, F_0) = \mathsf{H}(\hat{P}; \epsilon - C_L r^\alpha). \tag{121}$$

Since for a fixed $\hat{P}$, both $\mathcal{U}(\hat{P}, \cdot)$ and $\mathcal{L}(\hat{P}, \cdot)$ are continuous, there exists at least one such $F_0$. By assumption, we have for any $Q \in \mathbb{Q}_r(\hat{P})$,

$$|\mathcal{L}(\hat{P}, F_0) - \mathcal{L}(Q, F_0)| \leq C_L \|\hat{P} - Q\|^\alpha \leq C_L r^\alpha. \tag{122}$$

In particular, we have that $\mathcal{L}(Q, F_0) \leq \mathcal{L}(\hat{P}, F_0) + C_L r^\alpha \leq \epsilon$. Hence, $F_0 \in \mathbb{D}_{\mathbb{Q}}(\hat{P}; \epsilon, r)$ and, from the lower bound in (36),

$$\inf_{Q \in \mathbb{Q}_r(\hat{P})} \mathcal{U}(Q, F_0) \leq \mathcal{U}(P, F^*). \tag{123}$$

Since, by assumption,

$$|\mathcal{U}(\hat{P}, F_0) - \mathcal{U}(Q, F_0)| \leq C_U \|\hat{P} - Q\|^\alpha \leq C_U r^\alpha, \tag{124}$$

we have that

$$\mathcal{U}(Q, F_0) \geq \mathcal{U}(\hat{P}, F_0) - C_U r^\alpha \tag{125}$$

$$= \mathsf{H}(\hat{P}; \epsilon - C_L r^\alpha) - C_U r^\alpha. \tag{126}$$

In particular, this implies that

$$\inf_{Q \in \mathbb{Q}_r(\hat{P})} \mathcal{U}(Q, F_0) \geq \mathsf{H}(\hat{P}; \epsilon - C_L r^\alpha) - C_U r^\alpha. \tag{127}$$

Combining (127) and (123), inequality (120) holds. Combining (113) and (120) together, we get the desired conclusion. ■