



Article

# Mechanisms for Robust Local Differential Privacy

Milan Lopuhaä-Zwakenberg \*  and Jasper Goseling \* 

Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente,  
7522 NB Enschede, The Netherlands

\* Correspondence: m.a.lopuhaa@utwente.nl (M.L.-Z.); j.goseling@utwente.nl (J.G.)

**Abstract:** We consider privacy mechanisms for releasing data  $X = (S, U)$ , where  $S$  is sensitive and  $U$  is non-sensitive. We introduce the robust local differential privacy (RLDP) framework, which provides strong privacy guarantees, while preserving utility. This is achieved by providing robust privacy: our mechanisms do not only provide privacy with respect to a publicly available estimate of the unknown true distribution, but also with respect to similar distributions. Such robustness mitigates the potential privacy leaks that might arise from the difference between the true distribution and the estimated one. At the same time, we mitigate the utility penalties that come with ordinary differential privacy, which involves making worst-case assumptions and dealing with extreme cases. We achieve robustness in privacy by constructing an uncertainty set based on a Rényi divergence. By analyzing the structure of this set and approximating it with a polytope, we can use robust optimization to find mechanisms with high utility. However, this relies on vertex enumeration and becomes computationally inaccessible for large input spaces. Therefore, we also introduce two low-complexity algorithms that build on existing LDP mechanisms. We evaluate the utility and robustness of the mechanisms using numerical experiments and demonstrate that our mechanisms provide robust privacy, while achieving a utility that is close to optimal.

**Keywords:** local differential privacy; Rényi divergence; robust optimization



**Citation:** Lopuhaä-Zwakenberg, M.; Goseling, J. Mechanisms for Robust Local Differential Privacy. *Entropy* **2024**, *26*, 233. <https://doi.org/10.3390/e26030233>

Academic Editor: Songze Li

Received: 11 January 2024

Revised: 29 February 2024

Accepted: 3 March 2024

Published: 6 March 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

We consider the setting in which an aggregator collects data from many users with the purpose of, for instance, computing statistics or training a machine learning model. In particular, the data contain sensitive information and users do not trust the aggregator. Therefore, they employ a privacy mechanism that transforms the data before sending it to the aggregator. Users have data  $X = (S, U)$  from a finite alphabet  $\mathcal{X} = \mathcal{S} \times \mathcal{U}$ , where  $s \in \mathcal{S}$  is sensitive information and  $u \in \mathcal{U}$  is non-sensitive. Data are distributed i.i.d. across users according to the distribution  $P^*$ . In order to preserve their privacy, users disclose a sanitized version  $Y$  of  $X$  by using a privacy mechanism  $\mathcal{Q} : \mathcal{X} \rightarrow \mathcal{Y}$ . The aim is that  $Y$  contains as much information about  $X$  as possible without leaking too much information about  $S$ . The challenge that is addressed in this paper is to develop good privacy mechanisms. This scenario and closely related ones were studied in, for instance [1–11]. In this paper, we use the following version of local differential privacy (LDP), as introduced in [3]:

$$\mathbb{P}(Y = y | S = s) \leq e^\epsilon \mathbb{P}(Y = y | S = s'), \quad (1)$$

for all  $s, s' \in \mathcal{S}$  and privacy parameter  $\epsilon > 0$ . In addition, we measure the utility of  $Y$  through the mutual information  $I(X; Y)$ . We discuss differences with related work in Section 2.

Note that if all information is sensitive, i.e., if  $\mathcal{X} = \mathcal{S}$ , (1) reduces to

$$\mathbb{P}(Y = y | X = x) \leq e^\epsilon \mathbb{P}(Y = y | X = x'), \quad (2)$$

which is the traditional LDP constraint [1,2,5]. An important property of (2) is that it does not depend on  $P^*$ , but only on  $\mathcal{Q}$ . The independence of  $P^*$  is a key factor in the success of differential privacy, since it leverages the need to make assumptions about the distribution of the data or on the background/side-knowledge available to the aggregator. As is clear from (1), however, independence from  $P^*$  no longer holds if not all data are sensitive.

Assuming that  $P^*$  is known, one can develop good privacy mechanisms for various settings with partially sensitive information [3,6,12]. In practice, however,  $P^*$  has to be modeled using domain knowledge or estimated from data, leading to errors. The prevalent approach in the literature has been to develop privacy mechanisms based on a (point) estimate  $\hat{P}$  and analyze sensitivity with respect to errors in this estimate. In this work, we follow the approach that was proposed in [13,14], which is to construct a set  $\mathcal{F}$  of probability distributions that we are confident contains  $P^*$ . Subsequently, we construct privacy mechanisms that aim to maximize utility, while satisfying (1) for all probability distributions in  $\mathcal{F}$ . We call the resulting privacy framework robust local differential privacy (RLDP).

In a sense, RLDP is a relaxed form of privacy. Indeed, it may seem appealing, but it is—as we illustrate next—often infeasible to enforce (1) for all possible distributions. To this end, we consider two extreme cases. First, consider a joint distribution of  $S$  and  $U$  under which  $S = U$ . Intuitively, we cannot disclose much information about  $U$ , since this is directly leaking information about  $S$ . As such, the utility of  $Y$  is low. Next, consider a joint distribution under which  $S$  and  $U$  are independent. Intuitively, we can disclose  $U$  without additional precautions, providing a high utility on  $Y$ . The point is that we need to design a single privacy mechanism  $\mathcal{Q}$  that satisfies (1) for all distributions, including the ‘worst case’ in which  $S = U$ , leading to low utility  $Y$ . In this work, we take the mid-ground between, on the one hand, only using a point estimate  $\hat{P}$  and, on the other hand, using all possible distributions. We do so by defining a set of ‘reasonable’ distributions  $\mathcal{F}$ . In particular, we construct  $\mathcal{F}$  based on public side-information. This public side information consists of  $n$  pairs of data  $(s_1, u_1), \dots, (s_n, u_n)$ , which like the data of users are i.i.d. according to unknown distribution  $P^*$ . Our set  $\mathcal{F}$  is constructed as a closed ball under a Rényi divergence around the maximum likelihood point estimate  $\hat{P}$  of  $P^*$ . By doing so, we are (statistically) confident that  $\mathcal{F}$  contains  $P^*$ , with the radius of the ball controlling the confidence level.

The RLDP framework is an instance of the more general Pufferfish framework [15]. In Section 2, we make this connection explicit and use it to describe the semantic privacy guarantees that are offered by RLDP.

The main contributions of this paper are as follows:

1. We use a Rényi divergence to construct  $\mathcal{F}$  and analyze the resulting structure and statistics of  $\mathcal{F}$ . In particular, we demonstrate that projections of  $\mathcal{F}$  are again balls under the same divergence. Moreover, we bound the projected sets in terms of an  $\ell_1$  norm.
2. Using these results we approximate  $\mathcal{F}$  by an enveloping polytope. We then use techniques from robust optimization [16–18] to characterize PolyOpt, the mechanism that is optimal over this polytope.
3. A drawback of this method is that it relies on vertex enumeration and is, therefore, computationally unfeasible for large alphabets. Therefore, we introduce two low-complexity privacy mechanisms. The first is independent reporting (IR), in which  $S$  and  $U$  are reported through separate LDP mechanisms.
4. We characterize the conditions that underlying LDP mechanisms have to satisfy in order for IR to ensure RLDP. Furthermore, while IR can incorporate any LDP mechanism, we show that it is optimal to use randomized response [19]. This drastically reduces the search space and allows us to find the optimal IR mechanism using low-dimensional optimization.
5. The second low-complexity mechanism that we develop is called secret-randomized response (SRR) and is based on randomized response.

6. We show that SRR maximizes mutual information in the low-privacy regime for the case that  $\mathcal{F}$  is the entire probability simplex.
7. We demonstrate the improved utility of RLDP over LDP with numerical experiments. In particular, we compare the performance of our mechanisms with generalized random response [5]. We provide results for both synthetic data sets and real-world census data.

The structure of this paper is as follows: After discussing related work in Section 2, we describe the model in detail in Section 3. In Section 4, we present results on the structure and statistics of projections of  $\mathcal{F}$ . These results are used in Section 5 to develop the PolyOpt privacy mechanism. Low-complexity privacy mechanisms are presented in Sections 6 and 7. In Section 8, we evaluate the discussed methods experimentally. Finally, in Section 9, we provide a discussion of our results and provide an outlook on future work. Most proofs are deferred to Appendix A.

Part of this paper was presented at the IEEE International Symposium on Information Theory 2021 [14]. In this paper, we generalize from a  $\chi^2$ -divergence to an arbitrary Rényi divergence. Moreover, Sections 4 and 6, most of Section 8, and all proofs are new in the current paper.

## 2. Related Work

### 2.1. The Pufferfish Framework

Our RLDP framework is an instance of the more general Pufferfish framework [15]. In this subsection, we make this connection explicit and elaborate on the semantic guarantees offered by RLDP.

A privacy definition following the Pufferfish framework specifies (i) a set of potential secrets, (ii) a set of discriminative pairs of secrets, and (iii) a set of assumptions about how data are generated. In RLDP the potential secrets are the possible values of  $S$ , i.e.,  $\mathcal{S}$ . We want to prevent the aggregator from learning anything about  $S$ . This means that it should not be able to distinguish the case  $S = s$  from  $S = s'$  for all  $s \neq s'$ , so all non-identical pairs are discriminative. Note that this relies on  $\mathcal{S}$  being finite, with extensions to continuous  $\mathcal{S}$  discussed in detail in [15].

The set of assumptions on how data are generated consist, in our setting, of probability distributions over  $\mathcal{X}$ . A key idea in Pufferfish is that this set explicitly models the information that is available to an attacker, i.e., an entity that is trying to infer information about  $S$  by observing  $Y$ . In our setting, the aggregator is the only attacker and a probability distribution  $P$  over  $\mathcal{X}$  captures the beliefs that the attacker has about  $S$  prior to seeing  $Y$ . We can rewrite (1) as

$$\frac{\mathbb{P}_{X \sim P}(S = s | Y = y)}{\mathbb{P}_{X \sim P}(S = s' | Y = y)} \leq e^\epsilon \frac{\mathbb{P}_{X \sim P}(S = s)}{\mathbb{P}_{X \sim P}(S = s')} \quad (3)$$

and see that our local differential privacy constraint (1) can be interpreted as the condition that the posterior distribution of  $S$  after seeing  $Y$  must be very close to the prior distribution. The relevance of  $P$  is that it captures a specific set of beliefs of the attacker. As such, we want (3) to hold for various values of  $P$ , where each  $P$  captures specific background/side-knowledge available to the attacker/aggregator. Note that by doing so we are not making any claims about the actual knowledge available to the aggregator, but instead describing the possible scenarios for which we want to protect the privacy of users. In Pufferfish, these possible scenarios are called the set of assumptions on how data are generated, and in RLDP this is  $\mathcal{F}$ .

Often, side-information in the form of domain knowledge or existing data is publicly available; i.e., to both the users and the aggregator. This public side-information may suggest, for instance, that there is, at most, limited dependence between  $S$  and  $U$ . In that case, protecting against attackers who have the belief that  $S = U$  incurs an enormous penalty in achieved utility. It is true that those attackers gain a lot of information on  $S$  by observing  $Y$ . However, they could have also obtained this information from the public

side-information directly. Therefore, the approach taken in the Pufferfish framework and in this paper is that we only protect against attackers that have beliefs, i.e., distributions  $P$ , that are in line with publicly available side information.

A challenge in working with the Pufferfish framework is that it is often challenging to find good mechanisms. A general mechanism is proposed in [20], but it relies on enumerating over all distributions in  $\mathcal{F}$ , which is an uncountable set in our setting and cannot be used here. A constrained version of Pufferfish that facilitates analysis and a methodology for finding good mechanisms is proposed in [21]. Another interesting line of work is to model correlations between users in the non-local differential privacy setting [22]. Finally, ref. [23] proposed a modeling framework for capturing domain knowledge about the data. In contrast, in the current work, we impose constraints that are learned from data. Our setting does not fit any of the frameworks for which good mechanisms are known in the literature. One of the main contributions of this paper is to develop such mechanisms.

## 2.2. Other Privacy Frameworks

Disclosing  $X$  through a privacy mechanism that protects sensitive information  $S$  has been studied extensively. One line of work starts from differential privacy [24] and imposes the additional challenge that the aggregator cannot be trusted, leading to the concept of local differential privacy [1,2,5]. For this setting, several privacy mechanisms exist, including randomized response [19] and unary encoding [25]. Optimal LDP mechanisms under a variety of utility metrics, including mutual information, are found in [5]. In [1,2,5], all data are sensitive, i.e.,  $X = S$ . The variation of LDP for the case of disclosing  $X = (S, U)$ , where only  $S$  is sensitive, was proposed in [3] and is the setting that we study in this paper. Another line of work connects this setting to the information bottleneck [26], leading to a privacy constraint in terms of mutual information [6,8–10]. In these works, it is shown that approaches to optimizing the information bottleneck also work for finding good privacy mechanisms.

Next to differential privacy and mutual information as privacy measures, a multitude of other privacy frameworks and leakage measures exist [27]. Some of these have been studied in the context of privacy mechanisms. In [7,11], privacy leakage is measured through the improved potential of statistical inference by an attacker after seeing the disclosed information. This measure is formulated through a general cost function, with mutual information resulting as a special case. Perfect privacy, which demands the output to be independent of the sensitive data, was studied in [28], and methods were given to find optimal mechanisms in this setting. An estimation-theoretic framework was studied in [29,30]. Our use of a Rényi divergence in the construction of  $\mathcal{F}$  may suggest considering a generalization of our privacy definition. This could be achieved by considering, for instance, a Rényi divergence in the privacy constraint, as done in [31]. Along a different line, in [32], the maximal leakage measure with a clear operational interpretation is defined. In [33], this measure is generalized to a parametrized measure, enabling interpolating between maximal leakage and mutual information. A stronger, pointwise, version of the maximal leakage measure is proposed in [34]. These are interesting research directions but not pursued in this paper.

Our setting  $X = (S, U)$  is a special case of a Markov chain  $S - X - Y$ , where only  $X$  is observed. This Markov chain is typically studied in the information bottleneck and privacy funnel settings [6,26]. We do not generalize to this setting, because we need observations of  $S$  for the estimate of  $P_{U|S}$ . Without direct observations of  $s$ , we can only make worst-case assumptions on  $P_{U|S}$ , leading to very poor utility. A different type of model, in which only part of the information in  $X$  is sensitive, is proposed in [12]. This is a block-structured model in which  $X$  is partitioned and information about the partition of an element is sensitive but its index in the partition is not. Our setting of  $\mathcal{X} = \mathcal{S} \times \mathcal{U}$  does not fit this model. One can partition  $\mathcal{X}$  according to  $\mathcal{U}$ , but our privacy constraints are different from [12]. We will elaborate on this in Section 6.

### 2.3. Robustness

The distribution  $P_{S,U}^*$  is not available in practice. The approach taken in most works is to estimate  $P_{S,U}^*$  from data and analyze sensitivity with respect to this estimate  $\hat{P}_{S,U}$ . One of the contributions in [7] is to quantify the impact of mismatched priors, i.e., the impact of not knowing  $P_{S,U}^*$  exactly. A bound on the resulting level of privacy is derived in terms of the total variational distance between the actual and the estimated  $\hat{P}_{S,U}$ . The setting in [35] is similar to ours: A ball of probability distributions, centered around a point estimate, was defined that contains  $P_{S,U}^*$  with high probability. It was then shown that a privacy mechanism that was designed based on the empirical distribution was valid for the entire set for a looser privacy constraint. The privacy slack was quantified and shown to approach zero as the size of the data set increased. An important difference with the current work was that we explicitly optimize the privacy mechanism over the uncertainty set. Another difference is that we base our ball on a Rényi divergence, whereas [35] used an  $\ell_1$  norm. The main technical tool used in [35] was large deviations theory, whereas we rely on convex analysis and robust optimization. We also mention [36,37]. In [36] it is assumed that nothing is known about  $P_S^*$  and  $P_{U|S}^*$ . It is shown that good privacy mechanisms can be found through a connection to maximal correlation, see also [38]. In [37], sets of probability distributions are not derived from data but carefully modeled such that optimal mechanisms can be derived analytically.

Using robust optimization [16] to find a good mechanism that satisfies privacy constraints for all  $P_{S,U}$  in uncertainty set  $\mathcal{F}$  was proposed in [13,14]. In this work, we generalize and extend results from [14]. The idea of robust optimization is that constraints in an optimization problem contain uncertain parameters that are known to come from a (a priori defined) uncertainty set. The constraints must hold for possible values of the uncertain parameters. A key result is that, using Fenchel duality, the problem can be expressed in terms of the support function of the uncertainty set and the convex conjugate of the constraint [16,17]. The case where the uncertain parameters are probabilities is known as distributionally robust optimization. Using results from [39], it was shown in [40] how an uncertainty set can be constructed from data using an  $f$ -divergence, providing an approximate confidence set. Confidence sets for parameters that are not necessarily probabilities were constructed in [18] under a  $\chi^2$ -divergence. Convergence of robust optimization based on  $f$ -divergences was studied in [41] and for the case of a KL-divergence in [42]. In [43], it is shown how distributionally robust optimization problems over Wasserstein balls can be reformulated as convex problems. For the regular differential privacy setting, distributionally robust optimization was used in [44] to find optimal additive privacy mechanisms for a general perturbation cost function. In this paper, we show how robust optimization can be applied to the setting of partially sensitive information with local differential privacy.

### 2.4. Miscellaneous

Another line of work on privacy mechanisms builds on recent advances in generative adversarial networks [45]. In [46,47], a generative adversarial framework is used to provide privacy mechanisms that do not use explicit expressions for  $P_X$ . Even though this is not explicitly addressed in [46,47], it is expected that the generalization properties of networks will provide a form of robustness. Closely related approaches are used in the field of face recognition [48,49], with the aim of preventing biometric profiling [50]. The leakage measures that are used in [48,49], however, do not seem to have an operational interpretation.

Disclosing information in a privacy-preserving way is one of the main challenges in official statistics [51,52]. The setting considered in the current paper is closely connected to disclosing a table with microdata, where each record in the table is released independently of the other records. This approach to disclosing microdata was studied in [4] by considering expected error as the utility measure and mutual information as the privacy measure. The resulting optimization problem corresponds to the traditional rate-distortion problem.

### 3. Model and Preliminaries

In this section, we give an overview of the setting and objectives of this paper. The notation used in this section, as well as the rest of the paper, is summarized in Table 1.

**Table 1.** Notation used in this paper. ‘Page’ denotes the page the notation is first defined.

Notation	Meaning	Page
$\mathcal{S}$	sensitive data space	6
$\mathcal{U}$	non-sensitive data space	6
$\mathcal{X}$	$\mathcal{S} \times \mathcal{U}$	6
$a_1, a_2, a, b$	$ \mathcal{S} ,  \mathcal{U} ,  \mathcal{X} ,  \mathcal{Y} $	6
$X = (S, U)$	user data	6
ine $\mathcal{Q}$	privacy mechanism	6
$Q$	matrix of $\mathcal{Q}$	6
$Y$	$\mathcal{Q}(X)$	6
$\mathcal{Y}$	output space	6
$\mathcal{P}_{\mathcal{X}}$	space of prob. dist. on $\mathcal{X}$	6
$P^*$	true distribution	6
$\hat{P}$	estimated distribution	7
I	mutual information	8
$\mathcal{F}$	uncertainty set for $P$	6
$P_{U S}$	condition probability vector	9
$\mathcal{F}_{U S}$	conditional projection of $\mathcal{F}$	9
$L_{U S}(\mathcal{F}), \text{rad}_S(\mathcal{F})$	statistics of $\mathcal{F}$	9
$D_\alpha$	Rényi divergence	10
PolyOpt	PolyOpt	13
SRR $^\epsilon$	Secret Randomized Response	17
IR $_{\mathcal{R}^1, \mathcal{R}^2}$	Independent Reporting	18

The data space is  $\mathcal{X} = \mathcal{S} \times \mathcal{U}$ , where  $\mathcal{S}$  and  $\mathcal{U}$  are finite sets. We write  $|\mathcal{S}| =: a_1$ ,  $|\mathcal{U}| =: a_2$ , and  $|\mathcal{X}| = a_1 a_2 =: a$ . Data items  $X = (S, U)$  are drawn from a probability distribution  $P^*$  in  $\mathcal{P}_{\mathcal{X}}$ , the space of probability distributions on  $\mathcal{X}$ ; here,  $S$  represents sensitive data, while  $U$  represents non-sensitive data. The aggregator’s aim is to create a privacy mechanism  $\mathcal{Q}: \mathcal{X} \rightarrow \mathcal{Y}$  such that  $Y = \mathcal{Q}(X)$  contains as much information about  $X$  as possible, while not leaking too much information about  $S$ .

The mechanism  $\mathcal{Q}$  is a probabilistic map, which we represent by a left stochastic matrix  $(Q_{y|x})_{y \in \mathcal{Y}, x \in \mathcal{X}}$ , and we write  $|\mathcal{Y}| = b$ . Often, we identify  $\mathcal{Y} = \{1, \dots, b\}$ , and likewise for other sets.

The distribution  $P^*$  is not known exactly. Instead, there is a set of possible distributions  $\mathcal{F} \subset \mathcal{P}_{\mathcal{X}}$ , where  $\mathcal{P}_{\mathcal{X}}$  denotes the probability simplex over  $\mathcal{X}$ . We choose  $\mathcal{F}$  in such a way that it is likely that  $P^* \in \mathcal{F}$ . The uncertainty set  $\mathcal{F}$  captures our uncertainty about  $P^*$ , we guarantee privacy for all  $P \in \mathcal{F}$ . We denote this as robust local differential privacy (RLDP).

**Definition 1 (Robust Local Differential Privacy).** Let  $\epsilon \geq 0$  and  $\mathcal{F} \subset \mathcal{P}_{\mathcal{X}}$ . We say that  $\mathcal{Q}$  satisfies  $(\epsilon, \mathcal{F})$ -RLDP if for all  $s, s' \in \mathcal{S}$ , all  $y \in \mathcal{Y}$ , and all  $P \in \mathcal{F}$  we have

$$\mathbb{P}_{X \sim P}(Y = y | S = s) \leq e^\epsilon \mathbb{P}_{X \sim P}(Y = y | S = s'). \tag{4}$$

Note that we use the notation  $\mathbb{P}_{X \sim P}(\bullet)$  to emphasize that  $X$  is distributed according to  $P$ . If no confusion can arise, we often leave out the subscript  $X \sim P$ , to improve readability. Note that we can also write

$$\mathbb{P}_{X \sim P}(Y = y | S = s) = \sum_{u \in \mathcal{U}} Q_{y|s,u} \mathbb{P}_{X \sim P}(U = u | S = s), \tag{5}$$

so Definition 1 depends on the conditional probabilities of  $U$  given  $S = s$  and  $S = s'$ . It does not, however, depend on the realization of  $U$ .

For clarity and use in future sections, we give the definition of regular LDP [1], which is used when the goal is to obfuscate all of  $X$ , rather than just  $S$ .

**Definition 2** (Local Differential Privacy). *Let  $\epsilon \geq 0$ . We say that  $\mathcal{Q}: \mathcal{X} \rightarrow \mathcal{Y}$  satisfies  $\epsilon$ -LDP if for all  $x, x' \in \mathcal{X}$  and all  $y \in \mathcal{Y}$  we have*

$$\mathbb{P}(Y = y|X = x) \leq e^\epsilon \mathbb{P}(Y = y|X = x'). \tag{6}$$

Now, for aggregator uncertainty about  $P^*$ , as captured by  $\mathcal{F}$ , we suppose there is a data base  $\vec{x} = (x_1, \dots, x_n)$  accessible to the user, where each  $x_i = (s_i, u_i)$  is drawn independently from  $P^*$ . Based on this, the user produces an estimate  $\hat{P}$  of  $P^*$ . In the experiments, we consider a maximum likelihood estimator, i.e.,  $\hat{P}_x = |\{i \leq n : x_i = x\}|$ . We construct the uncertainty set  $\mathcal{F}$  as a closed ball around  $\hat{P}$ . In particular, let  $D_\alpha$  be the Rényi divergence of order  $\alpha$  on  $\mathcal{P}_\mathcal{X}$ , i.e., for  $\alpha \in (0, \infty)$

$$D_\alpha(\hat{P}||P) = \begin{cases} \frac{1}{\alpha-1} \log\left(\sum_{x \in \mathcal{X}} \frac{\hat{P}_x^\alpha}{P_x^{\alpha-1}}\right), & \text{if } \alpha \neq 1, \\ \sum_{x \in \mathcal{X}} \hat{P}_x \log \frac{\hat{P}_x}{P_x}, & \text{if } \alpha = 1. \end{cases} \tag{7}$$

The case  $\alpha = 1$  follows, in fact, as a limit from the  $\alpha \neq 1$  case. Similarly, the definition can be extended to  $\alpha \in \{0, \infty\}$  by taking the corresponding limits, but in this paper we restrict our attention to  $\alpha \in (0, \infty)$  to keep the presentation clear. Note that  $D_1 = D_{\text{KL}}$ , the Kullback–Leibler divergence, and  $D_2 = \log \chi^2$ , where the  $\chi^2$ -divergence is  $\chi^2(P_1||P_2) = \sum_x (P_{1,x} - P_{2,x})^2 P_{2,x}^{-1}$ . In general, a Rényi divergence is a continuous increasing function of a power divergence (a.k.a. Hellinger divergence) [39,53,54], an example of an  $f$ -divergence. We omit  $\alpha$  from the notation when it is clear from the context.

We define  $\mathcal{F}$  by fixing a bound  $B \in [0, \infty]$  and letting

$$\mathcal{F} = \{P \in \mathcal{P}_\mathcal{X} : D_\alpha(\hat{P}||P) \leq B\}. \tag{8}$$

Since a Rényi divergence is a continuous increasing function of an  $f$ -divergence, it follows from [39,40] that  $\mathcal{F}$  is a confidence set for  $P^*$ . In particular, for the case of  $\alpha = 2$ , which will be used in our numerical experiments in Section 8, for suitable  $B$ , we have

$$\mathcal{F} = \left\{ P \in \mathcal{P}_\mathcal{X} : \sum_x \frac{(\hat{P}_x - P_x)^2}{P_x} \leq \frac{F_{\chi^2, a-1}^{-1}(1 - \beta)}{n} \right\}, \tag{9}$$

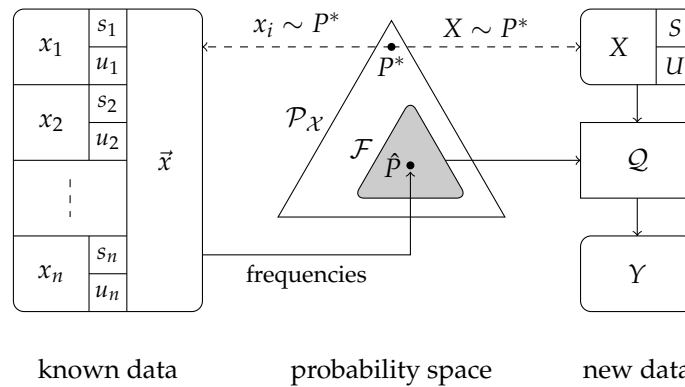
with  $\beta \in (0, 1)$ , where  $F_{\chi^2, a-1}$  is the cumulative density function of the  $\chi^2$ -distribution with  $a - 1$  degrees of freedom, resulting in a set  $\mathcal{F}$  with significance level  $\beta$ . This means that the probability of  $P^* \in \mathcal{F}$  is at least  $1 - \beta$ .

Hence, by designing  $\mathcal{Q}$  based on  $\mathcal{F}$ , we are confident in satisfying (1) for all attackers that have beliefs that are based on the public side-information, as well as for attackers that have beliefs that are closer to  $P^*$ .

As a special case of the above, we will study the case that nothing is known about  $P^*$ . In this case,  $B \rightarrow \infty$  and  $\mathcal{F} = \mathcal{P}_\mathcal{X}$ . Regarding privacy, this is the ‘safest’ choice, as we do not make assumptions about  $P^*$ . Another special case is where  $\mathcal{F}$  is a singleton, which reflects a situation where  $B = 0$  and  $P^*$  is assumed to be known. This setting was studied in [3].

Given  $\mathcal{F}$  and  $\epsilon$ , the goal is now to create a  $\mathcal{Q}: \mathcal{X} \rightarrow \mathcal{Y}$  to be used on new/future data; our setting is depicted in Figure 1. The aim of this paper is to find a satisfactory answer to the following problem:

**Problem 1.** *Given  $\mathcal{F}$  and  $\epsilon$ , find a  $\mathcal{Q}$  satisfying  $(\epsilon, \mathcal{F})$ -RLDP, while maximizing a given utility function.*



**Figure 1.** An overview of the setting of this paper when  $\mathcal{F}$  is a confidence set based on a data set  $\vec{x}$ . Note that it is typically, but not necessarily, true that  $P^* \in \mathcal{F}$ .

Throughout this paper, we follow the original privacy funnel [6] and its LDP counterpart [3] in taking mutual information  $I(X; Y)$  as a utility measure. As is argued in [6], mutual information arises naturally when minimizing log loss distortion in the privacy funnel scenario. As a utility measure of  $\mathcal{Q}$ , we take  $I_{X \sim P}(X; Y)$  (abbreviated to  $I_P(X; Y)$ ), since the aim is to create  $Y$  that reflects  $X$  as faithfully as possible. This utility measure depends on the distribution  $P$  of  $X$  that we choose to evaluate. Ideally, one would like to use  $P = P^*$ , but in practice this is not possible, as  $P^*$  is unknown. In the theoretical part of this paper, we circumvent this issue by proving our results for general  $P$ . In the experiments of Section 8, we take  $P = \hat{P}$  as the best available alternative to  $P = P^*$ . We investigate the effect of this choice by comparing  $I_{P^*}(X; Y)$  to  $I_{\hat{P}}(X; Y)$ .

Another option is to use the robust utility measure  $\min_{P \in \mathcal{F}} I_P(X; Y)$  to ensure good utility for every ‘reasonable’  $P$ , see [13]. We do not explicitly study this measure in this paper, but since our results hold for general  $P$ , they can also be applied to robust utility.

**Example 1.** We set up an example to illustrate the concepts of this paper. Take  $\mathcal{S} = \{s_1, s_2\}$  and  $\mathcal{U} = \{u_1, u_2\}$ , and suppose

$$P^* = \begin{pmatrix} P^*_{s_1, u_1} \\ P^*_{s_1, u_2} \\ P^*_{s_2, u_1} \\ P^*_{s_2, u_2} \end{pmatrix} = \begin{pmatrix} 0.1 \\ 0.1 \\ 0.2 \\ 0.6 \end{pmatrix}. \tag{10}$$

Moreover, suppose we have a publicly known database of  $n = 100$  entries, from which we estimate

$$\hat{P} = \begin{pmatrix} \hat{P}_{s_1, u_1} \\ \hat{P}_{s_1, u_2} \\ \hat{P}_{s_2, u_1} \\ \hat{P}_{s_2, u_2} \end{pmatrix} = \begin{pmatrix} 0.07 \\ 0.10 \\ 0.26 \\ 0.57 \end{pmatrix}. \tag{11}$$

To obtain a 95%-confidence set for  $\mathcal{F}$  according to a  $\chi^2$ -distribution, we take  $\alpha = 2$  and  $B = \log \left( 1 + \frac{F_{\chi^2, 3}^{-1}(0.05)}{100} \right) = 0.0752$ . In this way, we obtain



$$\mathcal{F} = \{P \in \mathcal{P}_{\mathcal{X}} : D_{\alpha}(\hat{P}||P) \leq B\} \tag{12}$$

$$= \left\{ P \in \mathcal{P}_{\mathcal{X}} : \log \left( \sum_x \frac{\hat{P}_x^2}{P_x} \right) \leq \log \left( 1 + \frac{F_{\chi^2,3}^{-1}(0.05)}{100} \right) \right\} \tag{13}$$

$$= \left\{ P \in \mathcal{P}_{\mathcal{X}} : \sum_x \frac{(\hat{P}_x - P_x)^2}{P_x} \leq \frac{F_{\chi^2,3}^{-1}(0.95)}{100} \right\}, \tag{14}$$

which is the desired confidence set (note that the  $\chi^2$ -distribution has  $|\mathcal{X}| - 1 = 3$  degrees of freedom). In this case, we have  $D_2(\hat{P}||P^*) = 0.0281 < B$ , so  $P^* \in \mathcal{F}$ .

#### 4. Conditional Projection of $\mathcal{F}$

In Sections 5 and 7 below, we will introduce privacy mechanisms that provide  $(\epsilon, \mathcal{F})$ -RLDP. These mechanisms depend on the conditional projections of  $\mathcal{F}$  on  $\mathcal{P}_{\mathcal{U}}$  given  $S = s$ , denoted as  $\mathcal{F}_{\mathcal{U}|s}$ . In this section, we analyze the structure and statistics of these sets. To do so, we introduce, for  $s \in \mathcal{S}$ ,  $u \in \mathcal{U}$  and  $P \in \mathcal{P}_{\mathcal{X}}$ .

$$P_s = \sum_{u \in \mathcal{U}} P_{u,s}, \tag{15}$$

$$P_{u|s} = \frac{P_{u,s}}{P_s}, \tag{16}$$

$$P_{\mathcal{U}|s} = (P_{u|s})_{u \in \mathcal{U}} \in \mathcal{P}_{\mathcal{U}}, \tag{17}$$

$$\mathcal{F}_{\mathcal{U}|s} = \{P_{\mathcal{U}|s} : P \in \mathcal{F}\} \subset \mathcal{P}_{\mathcal{U}}, \tag{18}$$

We are interested in the following statistics:

$$L_{u|s}(\mathcal{F}) = \min_{R \in \mathcal{F}_{\mathcal{U}|s}} R_u \quad \text{for a given } u \in \mathcal{U}, \tag{19}$$

$$\text{rad}_s(\mathcal{F}) = \max_{R \in \mathcal{F}_{\mathcal{U}|s}} \|R - \hat{P}_{\mathcal{U}|s}\|_1. \tag{20}$$

In (19),  $R_u$  is the  $u$ -coefficient of  $R \in \mathcal{P}_{\mathcal{U}}$ . It turns out that these statistics give us the information required to construct  $(\epsilon, \mathcal{F})$ -protocols efficiently: In Section 5, we use  $L_{u|s}(\mathcal{F})$  to approximate  $\mathcal{F}_{\mathcal{U}|s}$  by a polytope, to make computation easier, while in Section 7, we use  $\text{rad}_s(\mathcal{F})$  as a measure for the size of  $\mathcal{F}_{\mathcal{U}|s}$ . While these statistics (or bounds for them) are relatively easy to find for  $\mathcal{F}$  itself, the hard part lies in the fact that we have to give bounds for the projection  $\mathcal{F}_{\mathcal{U}|s}$ . The extent to which these bounds can be found explicitly heavily depends on the divergence measure that is used to construct  $\mathcal{F}$ . In this section, we show how these bounds can be obtained for our case where we construct  $\mathcal{F}$  using a Rényi divergence. The reason for this, as we will see below, is that we can give an explicit description of  $\mathcal{F}_{\mathcal{U}|s}$ .

##### 4.1. Structure of $\mathcal{F}_{\mathcal{U}|s}$

Recall that, for a given  $\alpha \in (0, \infty)$ , the Rényi divergence  $D_{\alpha} : \mathcal{P}_{\mathcal{X}} \rightarrow [0, \infty)$  is defined by

$$D_{\alpha}(\hat{P}||P) = \begin{cases} \frac{1}{\alpha-1} \log \left( \sum_{x \in \mathcal{X}} \frac{\hat{P}_x^{\alpha}}{P_x^{\alpha-1}} \right), & \text{if } \alpha \neq 1, \\ \sum_{x \in \mathcal{X}} \hat{P}_x \log \frac{\hat{P}_x}{P_x}, & \text{if } \alpha = 1. \end{cases} \tag{21}$$

The following theorem states that the conditional projections of balls defined by Rényi divergence are themselves Rényi divergence balls:

**Theorem 1.** Let  $s \in \mathcal{S}$  be such that  $\hat{P}_s > 0$ . Let  $\mathcal{F}$  be defined by Rényi divergence, i.e.,

$$\mathcal{F} = \{P \in \mathcal{P}_{\mathcal{X}} : D_{\alpha}(\hat{P}||P) \leq B\} \tag{22}$$

for a given  $\alpha \in (0, \infty)$  and  $B \in \mathbb{R}_{\geq 0}$ . Define the constant  $B_s$  by

$$B_s = \begin{cases} \frac{\alpha}{\alpha-1} \log\left(\frac{e^{(\alpha-1)B/\alpha} - (1-\hat{P}_s)}{\hat{P}_s}\right), & \text{if } \alpha \neq 1, \\ \frac{B}{\hat{P}_s}, & \text{if } \alpha = 1. \end{cases} \tag{23}$$

Then,

$$\mathcal{F}_{\mathcal{U}|s} = \left\{R \in \mathcal{P}_{\mathcal{U}} : D_{\alpha}(\hat{P}_{\mathcal{U}|s}||R) \leq B_s\right\}. \tag{24}$$

This theorem gives us a direct description of the  $\mathcal{F}_{\mathcal{U}|s}$ , which is useful because the  $L_{\mathcal{U}|s}(\mathcal{F})$  of (19) and  $\text{rad}_s(\mathcal{F})$  of (20) are defined in terms of these projection sets. A similar bound could also be found for the limit cases  $\alpha = 0, \infty$ , but this is not pursued in this paper, because it does not provide additional insights.

A key property of the Rényi divergence that allows us to prove Theorem 1 is that we can write

$$\frac{\hat{P}_x^{\alpha}}{P_x^{\alpha-1}} = \frac{\hat{P}_{\mathcal{U}|s}^{\alpha}}{P_{\mathcal{U}|s}^{\alpha-1}} \cdot \frac{\hat{P}_s^{\alpha}}{P_s^{\alpha-1}}. \tag{25}$$

This allows us to express the divergence  $D_{\alpha}(\hat{P}_{\mathcal{U}|s}||P_{\mathcal{U}|s})$  in terms of  $D_{\alpha}(\hat{P}||P)$ . For other divergences, which may depend on  $\hat{P}$  and  $P$  in a more complicated way, this is typically not possible. Therefore, we cannot generalize our results to uncertainty sets constructed from, for instance, arbitrary  $f$ -divergences.

In light of this theorem and the fact that in the following sections we care more about the statistics of  $\mathcal{F}_{\mathcal{U}|s}$  than about those of  $\mathcal{F}$  itself, one might be inclined to think that it is more straightforward to estimate the  $\hat{P}_{\mathcal{U}|s}$  from the data and defining uncertainty sets  $\mathcal{F}_{\mathcal{U}|s}$  around them directly, without going through the intermediate stage  $\mathcal{F}$ . However, projecting these sets back to  $\mathcal{P}_{\mathcal{X}}$  results in a larger set. In other words, there are distributions  $P$  such that each  $P_{\mathcal{U}|s}$  is an element of  $\mathcal{F}_{\mathcal{U}|s}$ , while  $P \notin \mathcal{F}$ . That is, we have  $\mathcal{F} \subsetneq \mathcal{F}' := \{P \in \mathcal{P}_{\mathcal{X}} : \forall s P_{\mathcal{U}|s} \in \mathcal{F}_{\mathcal{U}|s}\}$ . The reason for this is that, in the proof of Theorem 1, it becomes clear that the  $P \in \mathcal{F}$  that project to the boundary points of  $\mathcal{F}_{\mathcal{U}|s}$  satisfy  $P_{\mathcal{U}|s'} = \hat{P}_{\mathcal{U}|s'}$  for  $s' \neq s$ . In other words, elements of  $\mathcal{F}$  can be extremal in, at most, one  $\mathcal{F}_{\mathcal{U}|s}$ . By contrast,  $\mathcal{F}'$  also includes  $P$  that are extremal in multiple  $\mathcal{F}_{\mathcal{U}|s}$ . We conclude that constructing the  $\mathcal{F}_{\mathcal{U}|s}$  directly results in a larger  $\mathcal{F}'$ , which results in a lower utility. We will give an example of this phenomenon in Example 2.

#### 4.2. Statistics of $\mathcal{F}_{\mathcal{U}|s}$

In this section, we analyze statistics of  $\mathcal{F}_{\mathcal{U}|s}$ . More concretely, to find  $L_{\mathcal{U}|s}(\mathcal{F})$  and  $\text{rad}_s(\mathcal{F})$ , fix  $s, \alpha$  and  $B$  and define for  $\rho \in [0, 1]$  and  $\xi \in \mathbb{R}_{\geq 0}$  such that  $\xi(1 - \rho) \leq 1$ ,

$$\varphi_{B_s}(\rho, \xi) = \begin{cases} \frac{1}{\alpha-1} \log\left(\rho \xi^{1-\alpha} + (1-\rho)\left(\frac{1-\rho\xi}{1-\rho}\right)^{1-\alpha}\right) - B_s, & \text{if } \alpha \neq 1 \text{ and } \rho \neq 1, \\ \rho \log \frac{1}{\xi} + (1-\rho) \log \frac{1-\rho}{1-\rho\xi} - B_s, & \text{if } \alpha = 1 \text{ and } \rho \neq 1, \\ \log \frac{1}{\xi} - B_s, & \text{if } \rho = 1, \end{cases} \tag{26}$$

$$\xi_-(\rho) = \inf \left\{ \xi \in (0, 1] : \varphi_{B_s}(\rho, \xi) \leq 0 \right\}, \tag{27}$$

$$\xi_+(\rho) = \sup \left\{ \xi \in [1, (1-\rho)^{-1}) : \varphi_{B_s}(\rho, \xi) \leq 0 \right\}. \tag{28}$$

Note that the case  $\rho = 1$  can be obtained via taking the limit. The expressions for  $\xi_-$  and  $\xi_+$  are a bit complicated, but note that, given  $\rho < 1$ , the function  $\varphi_{B_s}(\rho, \xi)$  is convex in  $\xi$ . Thus,

$\varphi_{B_s}(\rho, \zeta) = 0$  has at most two solutions. Furthermore,  $\varphi_{B_s}(\rho, 1) = -B_s$  and  $\varphi_{B_s}(\rho, \zeta) \rightarrow \infty$  as  $\zeta$  approaches 0 or  $\frac{1}{1-\rho}$ , so for  $\rho < 1$  the values  $\zeta_-(\rho)$  and  $\zeta_+(\rho)$  are the two solutions to  $\varphi_{B_s}(\rho, \zeta) = 0$ .

The following proposition expresses our desired statistics in terms of  $\zeta_-$  and  $\zeta_+$ .

**Proposition 1.** *Let  $u \in \mathcal{U}$ . Then,*

$$L_{u|s}(\mathcal{F}) = \hat{P}_{u|s} \zeta_-(\hat{P}_{u|s}), \tag{29}$$

$$\text{rad}_s(\mathcal{F}) = 2 \max_{\substack{\mathcal{U}_1 \subset \mathcal{U}: \\ \mathcal{U}_1 \neq \emptyset}} \hat{P}_{\mathcal{U}_1|s} (\zeta_+(\hat{P}_{\mathcal{U}_1|s}) - 1). \tag{30}$$

As discussed above,  $\zeta_{\pm}(\rho)$  can be found quickly numerically; however, the calculation of  $\text{rad}_s(\mathcal{F})$  still involves taking the maximum over an exponentially large set.

### 4.3. Special Case $\alpha = 2$

In this section, we show that when  $\alpha = 2$ , we can find explicit expressions for  $\zeta_{\pm}$  and consequently  $L_{u|s}$  and  $\text{rad}_s$ . As discussed in (9), for this  $\alpha$ , the set  $\mathcal{F}$  is a confidence set for a  $\chi^2$ -test. To find  $\zeta_-(\rho), \zeta_+(\rho)$ , we need to solve  $\varphi_{B_s}(\rho, \zeta) = 0$ . For  $\alpha = 2$ , we can write this as a quadratic equation in  $\zeta$ , and solving it leads to the following expression:

**Lemma 1.** *Suppose  $\alpha = 2$ . Then,*

$$\zeta_-(\rho) = \frac{e^{B_s} + 2\rho - 1 - \sqrt{(e^{B_s} - 1)(e^{B_s} - (2\rho - 1)^2)}}{2e^{B_s}\rho}, \tag{31}$$

$$\zeta_+(\rho) = \frac{e^{B_s} + 2\rho - 1 + \sqrt{(e^{B_s} - 1)(e^{B_s} - (2\rho - 1)^2)}}{2e^{B_s}\rho}. \tag{32}$$

Now, we can determine  $L_{u|s}(\mathcal{F})$  and  $\text{rad}_s(\mathcal{F})$  using Lemma 1 and Proposition 1. For  $L_{u|s}(\mathcal{F})$ , we immediately obtain an expression; for  $\text{rad}_s(\mathcal{F})$ , a careful analysis of  $\zeta_+$  shows that the optimal  $\mathcal{U}_1$  of (30) can be found. For large enough  $B_s$ , the optimum is at  $\mathcal{U}_1 = \{u_{\min}\}$ , where  $u_{\min}$  is the  $u$  that minimizes  $\hat{P}_{u|s}$ . Thus, we obtain a concrete expression for  $\text{rad}_s(\mathcal{F})$  without the need for optimization. For smaller  $B_s$ , we do not find an exact expression, but we can still derive a lower bound. The results are summarized in the following proposition.

**Proposition 2.** *Let  $\alpha = 2$ . Then, the following hold:*

1. *One has*

$$L_{u|s}(\mathcal{F}) = \frac{e^{B_s} + 2\hat{P}_{u|s} - 1 - \sqrt{(e^{B_s} - 1)(e^{B_s} - (2\hat{P}_{u|s} - 1)^2)}}{2e^{B_s}}. \tag{33}$$

2. *Let  $u_{\min} = \arg \min_{u \in \mathcal{U}} \hat{P}_{u|s}$ . If  $B_s \geq \log(1 + (1 - \hat{P}_{u_{\min}|s})^2)$ , then*

$$\text{rad}_s(\mathcal{F}) = \frac{-e^{B_s} + 2\hat{P}_{u_{\min}|s} - 1 + \sqrt{(e^{B_s} - 1)(e^{B_s} - (2\hat{P}_{u_{\min}|s} - 1)^2)}}{e^{B_s}}. \tag{34}$$

3. *If  $B_s < \log(1 + (1 - \hat{P}_{u_{\min}|s})^2)$ , one has  $\text{rad}_s(\mathcal{F}) \leq \sqrt{e^{B_s} - 1}$ .*

We note that  $\alpha = 2$  is not the only value of  $\alpha$  for which one can bound  $L_{u|s}$  and  $\text{rad}_s$ . For instance, for  $\alpha \leq 1$ , one can use Pinsker’s inequality [55,56] and its generalizations [57] to bound  $\text{rad}_s(\mathcal{F})$  in terms of  $\|\hat{P}_{\mathcal{U}|s} - P_{\mathcal{U}|s}\|_1$ , which in turn can be used to bound  $L_{u|s}(\mathcal{F})$ . However, unlike  $\alpha = 2$ , these do not result in exact bounds.

**Example 2.** We continue Example 1. We have

$$\begin{aligned} \hat{P}_{s_1} &= 0.17, & \hat{P}_{u_1|s_1} &= 0.4118, & \hat{P}_{u_2|s_1} &= 0.5882, \\ \hat{P}_{s_2} &= 0.83, & \hat{P}_{u_1|s_2} &= 0.3133, & \hat{P}_{u_2|s_2} &= 0.6867. \end{aligned}$$

Inserting our values of  $B$  and  $\hat{P}_s$  into Theorem 2, we find  $B_{s_1} = 0.3782$ ,  $B_{s_2} = 0.0900$ . In other words,

$$\mathcal{P}_{\mathcal{U}|s_1} = \left\{ R = \begin{pmatrix} R_{u_1} \\ R_{u_2} \end{pmatrix} \in \mathcal{P}_{\mathcal{U}} : D_2 \left( \begin{pmatrix} 0.4118 \\ 0.5882 \end{pmatrix} \parallel \begin{pmatrix} R_{u_1} \\ R_{u_2} \end{pmatrix} \right) \leq 0.3782 \right\}, \tag{35}$$

$$\mathcal{P}_{\mathcal{U}|s_2} = \left\{ R = \begin{pmatrix} R_{u_1} \\ R_{u_2} \end{pmatrix} \in \mathcal{P}_{\mathcal{U}} : D_2 \left( \begin{pmatrix} 0.3133 \\ 0.6867 \end{pmatrix} \parallel \begin{pmatrix} R_{u_1} \\ R_{u_2} \end{pmatrix} \right) \leq 0.0900 \right\}. \tag{36}$$

To determine the lower bounds on each  $R_{u_i}$ , we use Proposition 2 to obtain

$$\begin{aligned} L_{u_1|s_1}(\mathcal{F}) &= 0.1620, & L_{u_2|s_1}(\mathcal{F}) &= 0.2829, \\ L_{u_1|s_2}(\mathcal{F}) &= 0.1923, & L_{u_2|s_2}(\mathcal{F}) &= 0.5337. \end{aligned}$$

In principle, we can also use Proposition 2 to determine the  $\text{rad}_s(\mathcal{F})$ . However, in this case, there is a more straightforward approach. Since  $|\mathcal{U}| = 2$ , every element of  $\mathcal{F}_{\mathcal{U}|s}$  is a vector of length two whose coefficients sum to 1; thus  $P_{\mathcal{U}|s}$  is determined by  $P_{u_1|s}$ . Since  $L_{u_1|s}(\mathcal{F}) \leq P_{u_1|s} \leq 1 - L_{u_2|s}(\mathcal{F})$ , it follows that

$$\begin{aligned} \mathcal{F}_{\mathcal{U}|s_1} &\cong [L_{u_1|s_1}(\mathcal{F}), 1 - L_{u_2|s_1}(\mathcal{F})] = [0.1620, 0.7171], \\ \mathcal{F}_{\mathcal{U}|s_2} &\cong [L_{u_1|s_2}(\mathcal{F}), 1 - L_{u_2|s_2}(\mathcal{F})] = [0.1923, 0.4663]. \end{aligned}$$

Under this identification,  $\text{rad}_s(\mathcal{F})$  is only twice the maximal distance from  $\hat{P}_{u_1|s}$  to the endpoint of this interval (the factor two comes from the fact that  $\|P_{\mathcal{U}|s} - \hat{P}_{\mathcal{U}|s}\|_1 = |P_{u_1|s} - \hat{P}_{u_1|s}| + |P_{u_2|s} - \hat{P}_{u_2|s}| = 2|P_{u_1|s} - \hat{P}_{u_1|s}|$ ). Hence,

$$\begin{aligned} \text{rad}_{s_1}(\mathcal{F}) &= 2 \max\{0.4118 - 0.1620, 0.7171 - 0.4118\} = 0.6107, \\ \text{rad}_{s_2}(\mathcal{F}) &= 2 \max\{0.3133 - 0.1923, 0.4663 - 0.3133\} = 0.3061. \end{aligned}$$

We can also construct the set  $\mathcal{F}' = \{P \in \mathcal{P}_{\mathcal{X}} : \forall s P_{\mathcal{U}|s} \in \mathcal{F}_{\mathcal{U}|s}\}$  of Section 4.1. We can write this as

$$\mathcal{F}' = \left\{ \begin{pmatrix} P_{s_1, u_1} \\ P_{s_1, u_2} \\ P_{s_2, u_1} \\ P_{s_2, u_2} \end{pmatrix} \in \mathcal{P}_{\mathcal{X}} : \begin{array}{l} 0.1620 \leq P_{u_1|s_1} \leq 0.7171, \\ 0.1923 \leq P_{u_1|s_2} \leq 0.4663 \end{array} \right\}. \tag{37}$$

The inequality  $0.1620 \leq P_{u_1|s_1}$  can be written as  $0.1620 \leq \frac{P_{s_1, u_1}}{P_{s_1, u_1} + P_{s_1, u_2}}$ , or  $0.1620 P_{s_1, u_2} \leq 0.83830 P_{s_1, u_1}$ ; in other words, this becomes a linear constraint. We can do the same for the other constraints and these, together with inequality constraints of the form  $P_{s, u} \geq 0$  and the equality constraint  $\sum_{s, u} P_{s, u} = 1$ , define the polytope  $\mathcal{F}' \subset \mathbb{R}^4$ . One can calculate that this polytope is a simplex, spanned by the vertices

$$\begin{pmatrix} 0.7171 \\ 0.2829 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.1620 \\ 0.8380 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0.4663 \\ 0.5337 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \\ 0.1923 \\ 0.8077 \end{pmatrix}. \tag{38}$$

The resulting  $\mathcal{F}'$  is considerably larger than  $\mathcal{F}$ : one way to see this is that, for any of these vertices  $P$ , one has  $D_2(\hat{P}||P) = \infty$ . This example shows the importance of working with the set  $\mathcal{F}$ , rather than with just its projections  $\mathcal{F}_{\mathcal{U}|s}$ .

### 5. Polyhedral Approximation: PolyOpt

In this section, we introduce *PolyOpt*, a family of mechanisms  $\mathcal{Q}$  with good utility obtained by enclosing  $\mathcal{F}$  by a polyhedron, and then using robust optimization for polyhedra [16] to describe the space of possible  $\mathcal{Q}$  as a polyhedron; we then maximize the mutual information over this polyhedron. This approach is related to the polyhedral approach of [3], which finds the optimum for this problem in a non-robust setting.

For a mechanism  $\mathcal{Q}$  and  $y \in \mathcal{Y}$ , we define  $Q_y = (Q_{y|x})_{x \in \mathcal{X}} \in \mathbb{R}^{\mathcal{X}}$  to be the  $y$ -th row of the stochastic matrix  $Q$  corresponding to  $\mathcal{Q}$ , but transposed (i.e., viewed as a column vector). Likewise, we define the column vector  $Q_{y|s} = (Q_{y|s,u})_{u \in \mathcal{U}} \in \mathbb{R}^{\mathcal{U}}$ . In this notation, the condition for  $(\epsilon, \mathcal{F})$ -RLDP can be formulated as

$$\forall y \in \mathcal{Y} \forall s_1, s_2 \in \mathcal{S}: \max_{P \in \mathcal{F}} \left[ P_{\mathcal{U}|s_1}^T Q_{y|s_1} - e^\epsilon P_{\mathcal{U}|s_2}^T Q_{y|s_2} \right] \leq 0. \tag{39}$$

Equation (39) boils down to a set of linear constraints in  $Q_y$ . What makes these difficult to satisfy is that every value  $P \in \mathcal{F}$  provides a linear constraint, and each  $Q_y$  has to satisfy all infinitely many of these. In this section, we address this difficulty by making the set  $\mathcal{F}$  slightly larger, so that robust optimization [16] becomes a convenient tool for optimizing over the allowed  $\mathcal{Q}$ . More precisely, for every  $s \in \mathcal{S}$ , let  $\mathcal{D}_s \subset \mathcal{P}_{\mathcal{U}}$  be such that  $\mathcal{F}_{\mathcal{U}|s} \subset \mathcal{D}_s$ . Then, certainly

$$\max_{P \in \mathcal{F}} \left[ P_{\mathcal{U}|s_1}^T Q_{y|s_1} - e^\epsilon P_{\mathcal{U}|s_2}^T Q_{y|s_2} \right] \leq \max_{\substack{R_1 \in \mathcal{D}_{s_1}, \\ R_2 \in \mathcal{D}_{s_2}}} \left[ R_1^T Q_{y|s_1} - e^\epsilon R_2^T Q_{y|s_2} \right]. \tag{40}$$

Thus, we can conclude that  $\mathcal{Q}$  is  $(\epsilon, \mathcal{F})$ -RLDP whenever

$$\forall y \in \mathcal{Y} \forall s_1, s_2 \in \mathcal{S}: \max_{\substack{R_1 \in \mathcal{D}_{s_1}, \\ R_2 \in \mathcal{D}_{s_2}}} \left[ R_1^T Q_{y|s_1} - e^\epsilon R_2^T Q_{y|s_2} \right] \leq 0. \tag{41}$$

The trick is now to choose the  $\mathcal{D}_s$  in such a way that the set of  $\mathcal{Q}$  satisfying (41) has a closed-form description. To this end, we let each  $\mathcal{D}_s$  be a polyhedron; that way, we can use robust optimization for polyhedra [16] to give such a description.

There are multiple ways to create a polyhedron  $\mathcal{D}_s$  that envelops  $\mathcal{F}_{\mathcal{U}|s}$ . Writing  $L_{u|s} = L_{u|s}(\mathcal{F})$  for convenience, we take

$$\mathcal{D}_s = \{R \in \mathcal{P}_{\mathcal{U}} : \forall u R_u \geq L_{u|s}\}. \tag{42}$$

Since  $\mathcal{D}_s$  is described by linear equations, it is a polyhedron, and certainly  $\mathcal{F}_{\mathcal{U}|s} \subset \mathcal{D}_{\mathcal{U}|s}$  for all  $s$ . Robust optimization for polytopes [16] then allows us to describe the set of mechanisms satisfying (41). To formulate this, we first need the following definition:

**Definition 3.** Let  $\epsilon > 0$ . Then, define  $\Gamma_\epsilon$  to be the convex cone consisting of all  $v \in \mathbb{R}_{\geq 0}^{\mathcal{X}}$  that satisfy, for all  $s_1, s_2 \in \mathcal{S}$  and all  $u_1, u_2 \in \mathcal{U}$ :

$$v_{s_1, u_1} - e^\epsilon v_{s_2, u_2} + \sum_u L_{u|s_1} (v_{s_1, u} - v_{s_1, u_1}) - e^\epsilon \sum_u L_{u|s_2} (v_{s_2, u} - v_{s_2, u_2}) \leq 0. \tag{43}$$

Note that, for every choice of  $s_1, s_2, u_1, u_2$ , (3) is a linear inequality in  $T$  and thus defines a half-space in  $\mathbb{R}^{\mathcal{X}}$ . The intersection of these half-spaces, intersected with  $\mathbb{R}_{\geq 0}^{\mathcal{X}}$ , defines the convex cone  $\Gamma_\epsilon$ . This definition allows us to formulate the following result:

**Theorem 2.** Let  $\mathcal{Q}$  be a privacy mechanism, and for  $y \in \mathcal{Y}$ , let  $Q_y$  be the  $y$ -th row of the associated matrix  $Q = (Q_{y|x})_{y \in \mathcal{Y}, x \in \mathcal{X}}$ . Suppose that for all  $y$  we have  $Q_y \in \Gamma_L$ . Then,  $\mathcal{Q}$  satisfies  $(\epsilon, \mathcal{F})$ -RLDP.

The upshot of this theorem is that we have translated the infinitely many constraints of (39) and (41) into the finitely many linear constraints of (3). This makes optimizing utility considerably easier. We perform this optimization by translating it into a linear programming problem. The key inspiration for this optimization is Theorem 4 of [5], where optimal LDP mechanisms are found by translating the problem of optimizing mutual information into linear programming; we use an analogous approach adapted to RLDP. This approach can be sketched as follows: Let  $\hat{\Gamma} = \{v \in \Gamma_\varepsilon : \sum_x v_x = 1\}$ , i.e., the intersection of  $\Gamma_\varepsilon$  with the hyperplane corresponding to  $\sum_x v_x = 1$ . This is a polyhedron, and every  $\mathcal{Q}$  satisfying the conditions of Theorem 2 has  $Q_y = \theta_y v_y$ , for some  $\theta_y \in \mathbb{R}_{\geq 0}$  and  $v_y \in \hat{\Gamma}$ . The authors of [5] made a number of key observations that also apply to our situation. The first is that, in this case, we can write

$$I_{\hat{p}}(X; Y) = \sum_y \theta_y \mu(v_y), \tag{44}$$

where

$$\mu(v) = \sum_{x \in \mathcal{X}} v_x \hat{p}_x \log \frac{v_x}{\sum_{x'} v_{x'} \hat{p}_{x'}}. \tag{45}$$

The second observation is that, in order to maximize (44), one can prove from the convexity of  $\mu$  that it is optimal to have each  $v_y$  be a vertex of  $\hat{\Gamma}$ . Thus, once we know the set of vertices  $\mathcal{V}$  of  $\hat{\Gamma}$ , we find the optimal  $\mathcal{Q}$  by assigning a weight  $\theta_v$  to each  $v \in \mathcal{V}$ , in such a way that the resulting  $Q_y$  form a probabilistic matrix and such that (44) is maximized. Since (44) is linear in  $\theta$ , this is a linear programming problem. This discussion is summarized in the following theorem:

**Theorem 3.** *Let  $\hat{\Gamma}$  be a polyhedron given by  $\{v \in \Gamma_{L,\varepsilon} : \sum_x v_x = 1\}$ . Let  $\mathcal{V}$  be the set of vertices of  $\hat{\Gamma}$ . Define  $\mu$  as in (45). Let  $1_{\mathcal{X}} \in \mathbb{R}^{\mathcal{X}}$  be the constant vector of ones. Let  $\hat{\theta} \in \mathbb{R}_{\geq 0}^{\mathcal{V}}$  be the solution to the optimization problem*

$$\begin{aligned} &\text{maximise}_{\theta} \sum_{v \in \mathcal{V}} \theta_v \mu(v) \\ &\text{satisfying } \theta \in \mathbb{R}_{\geq 0}^{\mathcal{V}}, \\ &\sum_{v \in \mathcal{V}} \theta_v v = 1_{\mathcal{X}}. \end{aligned} \tag{46}$$

*Let the privacy mechanism  $\mathcal{Q}$  be given by  $\mathcal{Y} = \{v \in \mathcal{V} : \hat{\theta}_v > 0\}$  and  $Q_{v|x} = \hat{\theta}_v v_x$ . Then, the mechanism  $\mathcal{Q}$  maximizes  $I_{\hat{p}}(X; Y)$  among all mechanisms satisfying the condition of Theorem 2. One has  $|\mathcal{Y}| \leq a$ .*

Together, Theorems 2 and 3 show that if we can solve a vertex enumeration problem, we can find a mechanism  $\mathcal{Q}$  that maximizes  $I_{\hat{p}}(X; Y)$  among a subset of all  $(\varepsilon, \mathcal{F})$ -RLDP mechanisms; furthermore, we ensure that the output space  $\mathcal{Y}$  is, at most, the size of the input space  $\mathcal{X}$ . The proof of Theorem 3 is analogous to the proof of Theorem 4 of [5] and is given in Appendix A.5. Note that the results of [5] do not run into the vertex enumeration problem, because the relevant polyhedron there is  $[1, e^\varepsilon]^{\mathcal{X}}$ , for which the vertices are known.

We remark that a simplex is not the only possible choice for  $\mathcal{D}_s$ . In general, we can make  $\mathcal{D}_s$  closer to  $\mathcal{F}_{U|s}$  by adding more defining hyperplanes. Doing this allows more  $\mathcal{Q}$  to satisfy Theorem 2 and in turn increases the utility of the  $\mathcal{Q}$  we find via Theorem 3. However, since  $\Gamma$  is related to the  $\mathcal{D}_s$  via duality, adding extra constraints to the  $\mathcal{D}_s$  will increase the dimension of  $\Gamma$  through the addition of auxiliary variables. This makes the vertex enumeration problem of Theorem 3 more computationally involved. Thus, we have a trade-off between utility and computational complexity. Even with the given, ‘simple’ choice of  $\mathcal{D}_s$ , the computational complexity is quite high: recall that we defined  $a = |\mathcal{X}|$ . The polytope  $\hat{\Gamma}$  is  $(a - 1)$ -dimensional and is defined by  $a^2 + a$  inequalities, thus it has  $\mathcal{O}((a^2 + a)^{\frac{a-1}{2}}) = \mathcal{O}(a^a)$  vertices [58]. Since this is the dimension of the linear programming

problem, we find that the total complexity of finding  $\mathcal{Q}$  is  $\mathcal{O}(a^{\omega a} \log(a^{a+1}/\delta))$ , where  $\omega \approx 2.38$  is the exponent of matrix multiplication and  $\delta$  the relative accuracy [58]. Clearly, this becomes infeasible rather quickly for large  $a$ .

It should be noted that, in general, the increasing utility obtained by decreasing  $\mathcal{D}_s$  in size does not approach the optimal utility over all  $(\epsilon, \mathcal{F})$ -RLDP mechanisms. This is because, as we take increasingly finer  $\mathcal{D}_{\mathcal{U}|s}$ , we approach the set of  $\mathcal{Q}$  that satisfy (4) for all  $P$  in  $\mathcal{F}' := \{P: \forall s P_{\mathcal{U}|s} \in \mathcal{F}_{\mathcal{U}|s}\}$ . As discussed in Section 4.1, one has  $\mathcal{F} \subsetneq \mathcal{F}'$ . As a result, the set of  $(\epsilon, \mathcal{F}')$ -RLDP mechanisms is strictly smaller than the set of  $(\epsilon, \mathcal{F})$ -RLDP mechanisms.

**Example 3.** We continue Example 2 by taking  $\epsilon = \log 2$ . To obtain  $\hat{\Gamma}$  in Theorem 3, we need to combine the defining inequalities of  $\Gamma_\epsilon$  in Definition 3, along with the defining equality  $\sum_x P_x = 1$ . Regarding the inequalities, we have  $2^4 = 16$  inequalities of the form (3), as well as 4 inequalities of the form  $v_x \geq 0$ . Together with the equality constraint, we obtain a 3-dimensional polytope in  $\mathbb{R}^{\mathcal{X}} = \mathbb{R}^4$ . Using a vertex enumeration algorithm, one finds that  $\mathcal{V}$  consists of the rows of the matrix  $V$  below, where the order of the columns is the order of the rows of Example 1. For each row  $v$ , we can calculate  $\mu(v)$ , resulting in the vector  $\mu$  below. Solving (46), we obtain the vector  $\hat{\theta}$  below:

$$V = \begin{pmatrix} 0.0744 & 0.3227 & 0.5603 & 0.0426 \\ 0.2426 & 0.2426 & 0.4783 & 0.0364 \\ 0.3333 & 0.3333 & 0.1667 & 0.1667 \\ 0.1091 & 0.4737 & 0.2086 & 0.2086 \\ 0.0993 & 0.4310 & 0 & 0.4697 \\ 0.1121 & 0.4864 & 0 & 0.4015 \\ 0.3404 & 0.3404 & 0 & 0.3191 \\ 0.0770 & 0.3343 & 0.2944 & 0.2944 \\ 0.2234 & 0.2234 & 0 & 0.5531 \\ 0.4875 & 0.1434 & 0 & 0.3690 \\ 0.4360 & 0.1283 & 0 & 0.4358 \\ 0.4758 & 0.1400 & 0.1921 & 0.1921 \\ 0.3437 & 0.1011 & 0.2776 & 0.2776 \\ 0.1602 & 0.1602 & 0.6316 & 0.0481 \\ 0.1667 & 0.1667 & 0.3333 & 0.3333 \\ 0.3325 & 0.0978 & 0.5294 & 0.0403 \end{pmatrix}, \mu = \begin{pmatrix} 0.1152 \\ 0.0942 \\ 0.0087 \\ 0.0135 \\ 0.1097 \\ 0.0968 \\ 0.0723 \\ 0.0080 \\ 0.1240 \\ 0.0878 \\ 0.1014 \\ 0.0106 \\ 0.0076 \\ 0.1240 \\ 0.0075 \\ 0.1083 \end{pmatrix}, \hat{\theta} = \begin{pmatrix} 1.1899 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.7670 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1.4134 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.6297 \end{pmatrix}. \tag{47}$$

We now obtain the privacy mechanism  $\mathcal{Q}_{\text{PolyOpt}}$  as follows: each row of  $\mathcal{Q}_{\text{PolyOpt}}$  corresponds to a non-zero coefficient of  $\hat{\theta}$ , multiplied by its corresponding row of  $V$ . Thus, we obtain

$$\mathcal{Q}_{\text{PolyOpt}} = \begin{pmatrix} Q_{y_1|s_1, u_1} & Q_{y_1|s_1, u_2} & Q_{y_1|s_2, u_1} & Q_{y_1|s_2, u_2} \\ Q_{y_2|s_1, u_1} & Q_{y_2|s_1, u_2} & Q_{y_2|s_2, u_1} & Q_{y_2|s_2, u_2} \\ Q_{y_3|s_1, u_1} & Q_{y_3|s_1, u_2} & Q_{y_3|s_2, u_1} & Q_{y_3|s_2, u_2} \\ Q_{y_4|s_1, u_1} & Q_{y_4|s_1, u_2} & Q_{y_4|s_2, u_1} & Q_{y_4|s_2, u_2} \end{pmatrix} \tag{48}$$

$$= \begin{pmatrix} 0.0885 & 0.3840 & 0.6667 & 0.0507 \\ 0.0860 & 0.3731 & 0 & 0.3080 \\ 0.6162 & 0.1813 & 0 & 0.6159 \\ 0.2094 & 0.0616 & 0.3333 & 0.0254 \end{pmatrix}. \tag{49}$$

Note that indeed we have  $4 = b \leq a = 4$ . As for the utility, we have  $I_{\hat{P}}(X; Y) = \mu \cdot \hat{\theta} = 0.4228$ . However, the true utility is significantly lower, namely  $I_{P^*}(X; Y) = 0.2804$ .

### 6. An Optimal Policy for $\mathcal{F} = \mathcal{P}_{\mathcal{X}}$

As PolyOpt mechanisms are obtained via vertex enumeration in  $a$ -dimensional space, this can be computationally infeasible for larger  $a$ . Thus, there is a need for methods that, given  $\hat{P}$  and  $\mathcal{F}$ , can find  $(\epsilon, \mathcal{F})$ -RLDP mechanisms with reasonable computational complexity.

In this section, we consider the case where  $\mathcal{F}$  is maximal, i.e.,  $\mathcal{F} = \mathcal{P}_{\mathcal{X}}$ . By itself, this represents a situation where we want privacy for every possible probability distribution on  $\mathcal{X}$ . This scenario may not be very relevant in practice, but any protocol that we find in this way is also  $(\epsilon, \mathcal{F})$ -RLDP for any  $\mathcal{F}$ . As we will see below, this allows us to find  $(\epsilon, \mathcal{F})$ -RLDP protocols in a computationally efficient manner.

We show that  $(\epsilon, \mathcal{P}_{\mathcal{X}})$ -RLDP is almost equivalent to LDP. We exploit this to create SRR, the RLDP analogue to GRR [5], the LDP mechanism that is optimal for  $\epsilon \gg 0$ . SRR only depends on  $\epsilon$  and  $\mathcal{X}$  and not on  $\hat{P}$ , and as such does not require an optimization procedure to be found; this makes it a good choice when vertex enumeration is computationally infeasible. The downside is that SRR has a stricter privacy requirement than PolyOpt, as it takes  $\mathcal{F}$  to be maximal; in Section 8, we investigate numerically to what extent this results in a lower utility.

We start by giving a characterization of  $(\epsilon, \mathcal{P}_{\mathcal{X}})$ -RLDP. Like LDP, this can be defined by an inequality constraint on the matrix  $Q$ .

**Proposition 3.**  $Q$  satisfies  $(\epsilon, \mathcal{P}_{\mathcal{X}})$ -RLDP if and only if for all  $y \in \mathcal{Y}$  and  $(s, u), (s', u') \in \mathcal{X}$  with  $s \neq s'$  one has

$$\frac{Q_{y|s,u}}{Q_{y|s',u'}} \leq e^\epsilon. \tag{50}$$

**Proof.** Suppose that  $Q$  satisfies  $(\epsilon, \mathcal{F})$ -RLDP with respect to  $\mathcal{P}_{\mathcal{X}}$ . Let  $(s, u), (s', u') \in \mathcal{X}$  with  $s \neq s'$ . Let  $P$  be given by

$$P_x = \begin{cases} \frac{1}{2}, & \text{if } x \in \{(s, u), (s', u')\}, \\ 0, & \text{otherwise.} \end{cases} \tag{51}$$

Then,  $P_{u|s} = 1$  and  $P_{u''|s} = 0$  for all  $u'' \neq u$ ; an analogous statements holds for  $P_{u'|s'}$ . It follows that

$$\frac{Q_{y|s,u}}{Q_{y|s',u'}} = \frac{Q_{y|s,u}P_{u|s}}{Q_{y|s',u'}P_{u'|s'}} \tag{52}$$

$$= \frac{\sum_{u''} Q_{y|s,u''}P_{u''|s}}{\sum_{u''} Q_{y|s',u''}P_{u''|s'}} \tag{53}$$

$$= \frac{\mathbb{P}_{X \sim P}(Q(X) = y|S = s)}{\mathbb{P}_{X \sim P}(Q(X) = y|S = s')} \leq e^\epsilon. \tag{54}$$

This proves “ $\Rightarrow$ ”. On the other hand, suppose that  $\frac{Q_{y|s,u}}{Q_{y|s',u'}} \leq e^\epsilon$  for all  $s \neq s'$  and  $u, u'$ . Then, for all  $s \neq s'$  and  $P$ , we have

$$\frac{\mathbb{P}_{X \sim P}(Q(X) = y|S = s)}{\mathbb{P}_{X \sim P}(Q(X) = y|S = s')} = \frac{\sum_u Q_{y|s,u}P_{u|s}}{\sum_{u'} Q_{y|s',u'}P_{u'|s'}} \leq e^\epsilon. \tag{55}$$

Hence,  $Q$  satisfies  $(\epsilon, \mathcal{P}_{\mathcal{X}})$ -RLDP with respect to  $\mathcal{F}$ .  $\square$

The proposition demonstrates that RLDP is very similar to LDP. The difference is that the condition “for all  $x, x' \in \mathcal{X}$ ” from Definition 2 is relaxed to only those  $x$  and  $x'$  for which  $s \neq s'$ .

Before moving on and introducing a new mechanism, note that Proposition 3 clearly illustrates the reason that the setting in this paper cannot be modeled using the block-structured approach from [12]. We see that if  $u \neq u'$ , we still have a privacy constraint, whereas in [12] this is not the case.

Next, we will introduce a mechanism that exploits the difference between LDP and RLDP. Recall that  $a = |\mathcal{X}|$ ; then generalized randomized response [19] is the privacy mechanism  $GRR^\epsilon: \mathcal{X} \rightarrow \mathcal{X}$  given by



$$\text{GRR}_{y|x}^\epsilon = \begin{cases} \frac{e^\epsilon}{e^\epsilon + a - 1} & \text{if } x = y, \\ \frac{1}{e^\epsilon + a - 1} & \text{otherwise.} \end{cases} \tag{56}$$

This mechanism has been designed such that  $\frac{\text{GRR}_{y|x}^\epsilon}{\text{GRR}_{y|x'}^\epsilon} = e^{\pm\epsilon}$  for  $x \neq x'$ , the maximal fractional difference that  $\epsilon$ -LDP allows. We will see that for RLDP we can go up to a difference of  $e^{\pm 2\epsilon}$  if  $x = (s, u)$  and  $x' = (s, u')$ , as we typically only need to satisfy

$$Q_{y|s,u} \leq e^\epsilon Q_{y|s',u'} \leq e^{2\epsilon} Q_{y|s,u'}. \tag{57}$$

We capture the intuition from the necessary condition (57) in a new mechanism called *secret randomized response* (SRR). Recall that  $a_1 = |\mathcal{S}|$ ,  $a_2 = |\mathcal{U}|$ .

**Definition 4** (Secret randomized response (SRR)). *Let  $\epsilon > 0$ . Then, the privacy mechanism  $\text{SRR}^\epsilon : \mathcal{X} \rightarrow \mathcal{X}$  is given by*

$$\text{SRR}_{s',u'|s,u}^\epsilon = \begin{cases} \frac{e^\epsilon}{e^\epsilon + e^{-\epsilon}(a_2 - 1) + a - a_2}, & \text{if } (s', u') = (s, u), \\ \frac{e^{-\epsilon}}{e^\epsilon + e^{-\epsilon}(a_2 - 1) + a - a_2}, & \text{if } s' = s \text{ and } u' \neq u, \\ \frac{1}{e^\epsilon + e^{-\epsilon}(a_2 - 1) + a - a_2}, & \text{if } s' \neq s, \end{cases} \tag{58}$$

It is clear that  $\frac{\text{SRR}_{y|s,u}^\epsilon}{\text{SRR}_{y|s',u'}^\epsilon} \in \{e^{-2\epsilon}, e^\epsilon, 1, e^\epsilon, e^{2\epsilon}\}$ , and the two extreme cases are only possible when  $s = s'$ . Thus, we can conclude

**Lemma 2.** *SRR satisfies  $(\epsilon, \mathcal{P}_\mathcal{X})$ -RLDP.*

**Example 4.** *We continue Example 3. Although SRR is closely related to GRR, adopting it can still have a significant impact on utility. For instance, in the setting of Example 3, we obtain*

$$\text{GRR}^\epsilon = \begin{pmatrix} 0.4 & 0.2 & 0.2 & 0.2 \\ 0.2 & 0.4 & 0.2 & 0.2 \\ 0.2 & 0.2 & 0.4 & 0.2 \\ 0.2 & 0.2 & 0.2 & 0.4 \end{pmatrix}, \quad \text{SRR}^\epsilon = \begin{pmatrix} 0.444 & 0.111 & 0.222 & 0.222 \\ 0.111 & 0.444 & 0.222 & 0.222 \\ 0.222 & 0.222 & 0.444 & 0.111 \\ 0.222 & 0.222 & 0.111 & 0.444 \end{pmatrix}. \tag{59}$$

Then,

$$I_{\hat{P}}(X; \text{GRR}^\epsilon(X)) = 0.0419, \quad I_{\hat{P}}(X; \text{SRR}^\epsilon(X)) = 0.1005, \tag{60}$$

$$I_{P^*}(X; \text{GRR}^\epsilon(X)) = 0.0412, \quad I_{P^*}(X; \text{SRR}^\epsilon(X)) = 0.0942. \tag{61}$$

We see that adopting SRR more than doubles the utility. Compared to Example 3, we see that the utility is still significantly lower than that of PolyOpt, but the advantage is that we obtain SRR directly from  $\epsilon$ , without having to take  $\hat{P}$  or  $\mathcal{F}$  into account; this ensures a significantly faster computation.

The power of SRR, beyond slightly improving on GRR, is that we can prove it maximizes  $I_P(X; Y)$  for sufficiently large  $\epsilon$ ; the cutoff point depends on  $P$ . This is proven analogously to the result of [5], where GRR is the optimal LDP mechanism for sufficiently large  $\epsilon$ .

**Theorem 4.** *For every  $P$ , there is an  $\epsilon_0 \geq 0$  such that for all  $\epsilon \geq \epsilon_0$ , SRR is the  $(\epsilon, \mathcal{P}_\mathcal{X})$ -RLDP mechanism maximizing  $I_P(X; Y)$ .*

The proof of this theorem follows the same lines as the proof of Theorem 14 of [5], in which it is proven that GRR is the optimal LDP mechanism for sufficiently large  $\epsilon$ . The

proof is presented in Appendix A.6. This solves the problem of finding the optimal  $(\epsilon, \mathcal{P}_X)$ -mechanism, for sufficiently large  $\epsilon$ . This strategy is similar to the proof of Theorem 3: one can show that the rows  $Q_y$  of the optimal  $(\epsilon, \mathcal{F})$ -RLDP mechanism  $\mathcal{Q}$  correspond to vertices of a polyhedron, and the optimal weights assigned to these vertices are found using a linear programming problem. Unlike in the case of Theorem 3, however, we can give an explicit description of the set of vertices, and we can solve the linear programming problem analytically.

Our result shows that if one wishes to satisfy  $(\epsilon, \mathcal{P}_X)$ -RLDP, then SRR is a solid choice, especially for larger  $\epsilon$ , since it maximizes  $I_{P^*}(X; Y)$  for sufficiently  $\epsilon$ . Thus, we can optimize  $I_{P^*}(X; Y)$  without having to know  $P^*$ , with the caveat that the cutoff point for ‘large enough’ depends on  $P^*$ .

In [5], the optimal LDP mechanism in the high-privacy regime (i.e.,  $\epsilon \ll 1$ ) was also found. In principle, we could also do this for  $(\epsilon, \mathcal{P}_X)$ -RLDP, but this would not be of much use, as the optimal mechanism would depend on  $P^*$ , which we assume to be unknown.

### 7. Independent Reporting

Section 5 demonstrated the need to find efficiently computable  $(\epsilon, \mathcal{F})$ -RLDP mechanisms with decent utility. In Section 6, we approach this problem by considering  $(\epsilon, \mathcal{P}_X)$ -RLDP instead, allowing us to analytically obtain the optimal mechanism. However, when  $\mathcal{F}$  is small, this overapproximation might result in a large loss of utility. In this section, we describe *independent reporting* (IR), a different heuristic that takes the size of  $\mathcal{F}$  into account, while still being significantly less computationally complex than PolyOpt.

The basis of IR is to apply two separate LDP mechanisms  $\mathcal{R}^1$  and  $\mathcal{R}^2$  to  $S$  and  $U$ , respectively, reporting both outputs.

**Definition 5.** Let  $\mathcal{Y}^1, \mathcal{Y}^2$  be sets, and let  $\mathcal{Y} = \mathcal{Y}^1 \times \mathcal{Y}^2$ . Let  $\mathcal{R}^1: \mathcal{S} \rightarrow \mathcal{Y}^1$  and  $\mathcal{R}^2: \mathcal{U} \rightarrow \mathcal{Y}^2$  be probabilistic maps. Then, the independent reporting of  $\mathcal{R}^1$  and  $\mathcal{R}^2$  is the probabilistic map  $\text{IR}_{\mathcal{R}^1, \mathcal{R}^2}: \mathcal{X} \rightarrow \mathcal{Y}$  given by  $\text{IR}_{\mathcal{R}^1, \mathcal{R}^2}(s, u) = (\mathcal{R}^1(s), \mathcal{R}^2(u))$ .

Suppose that  $\mathcal{R}^i$  satisfies  $\epsilon_i$ -LDP. The composition theorem for differential privacy [59] tells us that  $\text{IR}_{\mathcal{R}^1, \mathcal{R}^2}$  satisfies  $(\epsilon_1 + \epsilon_2)$ -LDP. However, in the RLDP setting,  $U$  only indirectly leaks information about  $S$ ; therefore, we can get away with a higher  $\epsilon_2$  compared to the LDP setting. How much higher depends on the degree of relatedness of  $S$  and  $U$ , which is captured by the possible values of  $P$  in  $\mathcal{F}$ . The precise statement is given in the following result:

**Theorem 5.** Let  $\epsilon_1, \epsilon_2 \in \mathbb{R}_{\geq 0}$ . For each  $s$ , let  $d_s \in [0, \infty)$  be such that  $d_s \geq \text{rad}_s(\mathcal{F})$ . Furthermore, define

$$d = \min \left\{ 2, \max_s (2d_s) + \max_{s, s'} \|\hat{P}_{U|s} - \hat{P}_{U|s'}\|_1 \right\}. \tag{62}$$

Let  $\delta_2 = \log \left( 1 + \frac{2(\epsilon_2 - 1)}{d} \right)$ . Suppose that  $\mathcal{R}^1$  is  $\epsilon_1$ -LDP and that  $\mathcal{R}^2$  is  $\delta_2$ -LDP. Then, IR is  $(\epsilon_1 + \epsilon_2, \mathcal{F})$ -RLDP.

If  $S = U$ , then  $\|\hat{P}_{U|s} - \hat{P}_{U|s'}\|_1 = 2$  for  $s \neq s'$ , so  $d = 2$  and  $\delta_2 = \epsilon_2$ . In this case, Theorem 5 is the RLDP analogue to the well-known composition theorem for local differential privacy [59]. In general,  $\delta_2 \geq \epsilon_2$ ; this represents the fact the privacy requirement on  $\mathcal{R}^2$  is less strict when  $S$  and  $U$  are only partially related. At the other extreme, if  $S$  and  $U$  are independent in our observation, we have  $\|\hat{P}_{U|s} - \hat{P}_{U|s'}\|_1 = 0$  for all  $s, s'$ . Still, we cannot fully disclose  $U$ , since  $S$  and  $U$  might be non-independent under  $P^*$ . The term  $d_s$  is present in the definition of  $d$  to account for this possibility.

In order to prove Theorem 5, we need the following lemma:

**Lemma 3.** Let  $\mathcal{Q}: \mathcal{X} \rightarrow \mathcal{Y}$  be an  $\varepsilon$ -LDP mechanism. Then, for all  $y \in \mathcal{Y}$  and all  $P, P' \in \mathcal{P}_{\mathcal{X}}$  we have

$$\frac{\mathbb{P}_{X \sim P}(\mathcal{Q}(X) = y)}{\mathbb{P}_{X \sim P'}(\mathcal{Q}(X) = y)} \leq 1 + \frac{e^\varepsilon - 1}{2} \|P - P'\|_1. \tag{63}$$

**Proof.** Fix  $y$ , and let  $Q_y^{\max} = \max_x Q_{y|x}$  and  $Q_y^{\min} = \min_x Q_{y|x}$ . By the  $\varepsilon$ -LDP property, it holds that  $Q_y^{\max} \leq e^\varepsilon Q_y^{\min}$ . We hence find

$$\mathbb{P}_{X \sim P}(\mathcal{Q}(X) = y) - \mathbb{P}_{X \sim P'}(\mathcal{Q}(X) = y) = \sum_{x \in \mathcal{X}} Q_{y|x}(P_x - P'_x) \tag{64}$$

$$= \sum_{x: P_x \geq P'_x} Q_{y|x}(P_x - P'_x) - \sum_{x: P'_x > P_x} Q_{y|x}(P'_x - P_x) \tag{65}$$

$$\leq \frac{Q_y^{\max}}{2} \|P - P'\|_1 - \frac{Q_y^{\min}}{2} \|P - P'\|_1 \tag{66}$$

$$\leq \frac{(e^\varepsilon - 1)Q_y^{\min}}{2} \|P - P'\|_1 \tag{67}$$

$$\leq \frac{(e^\varepsilon - 1)\mathbb{P}_{X \sim P'}(\mathcal{Q}(X) = y)}{2} \|P - P'\|_1, \tag{68}$$

from which the lemma directly follows.  $\square$

**Proof of Theorem 5.** We start by showing that  $d$  is an upper bound for  $\|P_{\mathcal{U}|s} - P_{\mathcal{U}|s'}\|_1$ . If  $d = 2$ , this is certainly the case. Suppose  $d = \max_s(2d_s) + \max_{s,s'} \|\hat{P}_{\mathcal{U}|s} - \hat{P}_{\mathcal{U}|s'}\|_1$ . Then, for all  $s, s' \in \mathcal{S}$  and  $P \in \mathcal{F}$  we have

$$\|P_{\mathcal{U}|s} - P_{\mathcal{U}|s'}\|_1 \leq \|P_{\mathcal{U}|s} - \hat{P}_{\mathcal{U}|s}\|_1 + \|\hat{P}_{\mathcal{U}|s} - \hat{P}_{\mathcal{U}|s'}\|_1 + \|\hat{P}_{\mathcal{U}|s'} - P_{\mathcal{U}|s'}\|_1 \tag{69}$$

$$\leq d_s + d_{s'} + \|\hat{P}_{\mathcal{U}|s} - \hat{P}_{\mathcal{U}|s'}\|_1 \tag{70}$$

$$\leq d. \tag{71}$$

Combining Lemma 3 with the fact that  $\varepsilon_2 = \log\left(1 + \frac{d(e^{\delta_2} - 1)}{2}\right)$ , it follows that for every  $y_2 \in \mathcal{Y}_2$ , we have

$$\frac{\mathbb{P}_{X \sim P}(\mathcal{R}^2(U) = y_2 | S = s)}{\mathbb{P}_{X \sim P}(\mathcal{R}^2(U) = y_2 | S = s')} \leq 1 + \frac{e^{\delta_2} - 1}{2} \|P_{\mathcal{U}|s} - P_{\mathcal{U}|s'}\|_1 \tag{72}$$

$$\leq 1 + \frac{d(e^{\delta_2} - 1)}{2} \tag{73}$$

$$= e^{\varepsilon_2}. \tag{74}$$

Given  $S$ , the random variables  $\mathcal{R}^1(S)$  and  $\mathcal{R}^2(U)$  are independent. It follows that for every  $y_1 \in \mathcal{Y}_1$  and every  $y_2 \in \mathcal{Y}_2$ , we have

$$\frac{\mathbb{P}(\mathcal{R}^1(S) = y_1, \mathcal{R}^2(U) = y_2 | S = s)}{\mathbb{P}(\mathcal{R}^1(S) = y_1, \mathcal{R}^2(U) = y_2 | S = s')} = \frac{\mathbb{P}(\mathcal{R}^1(S) = y_1 | S = s)}{\mathbb{P}(\mathcal{R}^1(S) = y_1 | S = s')} \cdot \frac{\mathbb{P}(\mathcal{R}^2(U) = y_2 | S = s)}{\mathbb{P}(\mathcal{R}^2(U) = y_2 | S = s')} \tag{75}$$

$$\leq e^{\varepsilon_1 + \varepsilon_2}, \tag{76}$$

where the last equality holds because of (74) and because  $\mathcal{R}^1$  is  $\varepsilon_1$ -LDP. This shows that  $\text{IR}_{\mathcal{R}^1, \mathcal{R}^2}$  is  $(\varepsilon_1 + \varepsilon_2, \mathcal{F})$ -RLDP.  $\square$

Theorem 5 establishes the privacy of independent reporting. To maximize the utility, we need to determine how to divide the privacy budget  $\varepsilon$  between  $\varepsilon_1$  and  $\varepsilon_2$ , and which LDP mechanisms to use for  $\mathcal{R}^1$  and  $\mathcal{R}^2$ . To answer both these questions, we first need an expression for the utility of IR, which is given by the following theorem:

**Theorem 6.** For any  $P \in \mathcal{P}_{\mathcal{X}}$ , one has

$$I_P(\text{IR}_{\mathcal{R}^1, \mathcal{R}^2}(X); X) = I_P(\mathcal{R}^1(S); S) + I_P(\mathcal{R}^2(U); U | \mathcal{R}^1(S)). \tag{77}$$

**Proof.** Since  $\mathcal{R}^1(S)$  and  $U$  are independent given  $S$ , and  $\mathcal{R}^2(U)$  and  $S$  are independent given  $U$  and  $\mathcal{R}^1(S)$ , we have

$$I_P(\mathbb{IR}_{\mathcal{R}^1, \mathcal{R}^2}(X); X) = I_P(\mathcal{R}^1(S), \mathcal{R}^2(U); U, S) \tag{78}$$

$$= I_P(\mathcal{R}^1(S); U, S) + I_P(\mathcal{R}^2(U); U, S | \mathcal{R}^1(S)) \tag{79}$$

$$= I_P(\mathcal{R}^1(S); S) + I_P(\mathcal{R}^2(U); U | \mathcal{R}^1(S)). \tag{80}$$

□

We use Theorems 5 and 6 to find high-utility IR protocols that satisfy  $(\epsilon, \mathcal{F})$ -RLDP, given  $\epsilon$  and  $\mathcal{F}$ . To do so, we need to choose  $\mathcal{R}^1$  and  $\mathcal{R}^2$ , and split the privacy budget between them. Since the expression for the utility of IR in Theorem 6 contains a term  $I_P(\mathcal{R}^1(S); S)$ , the  $\mathcal{R}^1$  that maximizes this is GRR when  $\epsilon$  is large enough; thus, we choose  $\mathcal{R}^1 = \text{GRR}$ . The second term in the utility expression is

$$I_P(\mathcal{R}^2(U); U | \mathcal{R}^1(S)) = \mathbb{E}_r \left[ I_{U \sim P_U | \mathcal{R}^1(S)=r}(\mathcal{R}^2(U); U) \right]. \tag{81}$$

This is the expected value of an expression that is maximized for  $\mathcal{R}^2 = \text{GRR}$ , with the caveat that the maximization only holds when  $\epsilon$  is large enough, and what ‘large enough’ is depends on the distribution of  $U$ . Since this gives us a choice of  $\mathcal{R}^2$  independent of the distribution, we ignore this caveat and take  $\mathcal{R}^2 = \text{GRR}$  as well.

Having chosen  $\mathcal{R}^1$  and  $\mathcal{R}^2$ , we are only left with the division of the privacy budget. If we choose  $\epsilon_2$ , then by Theorem 5 the privacy parameters of  $\mathcal{R}^1$  and  $\mathcal{R}^2$  are  $\epsilon_1 = \epsilon - \epsilon_2$  and  $\delta_2 = \log\left(1 + \frac{2(e^{\epsilon_2}-1)}{d}\right)$ , respectively. It follows that to find a high-utility IR protocol, we have to solve the following optimization problem:

$$\begin{aligned} & \text{maximize}_{\epsilon_2} I_P \left( \text{GRR}_{\epsilon-\epsilon_2}(S), \text{GRR}_{\log\left(1 + \frac{2(e^{\epsilon_2}-1)}{d}\right)}(U); S, U \right) \\ & \text{subject to } \epsilon_2 \in [0, \epsilon]. \end{aligned} \tag{82}$$

This optimization problem is only 1-dimensional. While it is not straightforward to express the complexity of solving this in  $\mathcal{O}$ -notation, our experiments in Section 8 show this can be quickly performed numerically, and significantly faster than PolyOpt.

**Example 5.** We continue Example 4. Having found  $\text{rad}_s(\mathcal{F})$  and  $\hat{P}_{U|s_1}, \hat{P}_{U|s_2}$  in Example 2, we conclude that, in Theorem 5, we have

$$d = \min \left\{ 2, 2 \cdot \max\{0.6107, 0.3061\} + \left\| \begin{pmatrix} 0.4118 \\ 0.5882 \end{pmatrix} - \begin{pmatrix} 0.3133 \\ 0.6867 \end{pmatrix} \right\|_1 \right\} = 1.4591. \tag{83}$$

It follows that  $\delta_2 = \log\left(1 + \frac{2}{1.4591}(e^{\epsilon_2} - 1)\right) = \log(1.3707e^{\epsilon_2} - 0.3707)$ . For a given value of  $\epsilon_2$ , the matrix corresponding to  $\mathbb{IR}(\text{GRR}^{\log(2)-\epsilon_2}, \text{GRR}^{\delta_2})$  is the Kronecker product

$$\begin{aligned} & \begin{pmatrix} \frac{2e^{-\epsilon_2}}{2e^{-\epsilon_2}+1} & \frac{1}{2e^{-\epsilon_2}+1} \\ \frac{1}{2e^{-\epsilon_2}+1} & \frac{2e^{-\epsilon_2}}{2e^{-\epsilon_2}+1} \end{pmatrix} \otimes \begin{pmatrix} \frac{1.3707e^{\epsilon_2}-0.3707}{1.3707e^{\epsilon_2}+0.6293} & \frac{1}{1.3707e^{\epsilon_2}+0.6293} \\ \frac{1}{1.3707e^{\epsilon_2}+0.6293} & \frac{1.3707e^{\epsilon_2}-0.3707}{1.3707e^{\epsilon_2}+0.6293} \end{pmatrix} \\ & = \frac{1}{C} \begin{pmatrix} \frac{2.7414-0.7414e^{-\epsilon_2}}{2e^{-\epsilon_2}} & \frac{2e^{-\epsilon_2}}{2.7414-0.7414e^{-\epsilon_2}} & \frac{1.3707e^{\epsilon_2}-0.3707}{1} & \frac{1}{1.3707e^{\epsilon_2}-0.3707} \\ \frac{1.3707e^{\epsilon_2}-0.3707}{1} & \frac{1}{1.3707e^{\epsilon_2}-0.3707} & \frac{2.7414-0.7414e^{-\epsilon_2}}{2e^{-\epsilon_2}} & \frac{1.3707e^{\epsilon_2}-0.3707}{2e^{-\epsilon_2}} \end{pmatrix}, \end{aligned} \tag{84}$$

where  $C = (2e^{-\epsilon_2} + 1)(1.3707e^{\epsilon_2} + 0.6293)$ . We now wish to optimize its utility, i.e., find the  $\epsilon_2 \in [0, \log 2]$  that maximizes  $I_P(X; Y)$ . The optimum occurs at the boundary  $\epsilon_2 = \log(2)$ , for which  $I_P(X; Y) = 0.0755$ . Notice that now  $\epsilon_1 = 0$ , so  $\mathcal{R}^1 = \text{GRR}^0$  is completely random: its output does not depend on the input. In other words, the optimal IR protocol in this case does not

transmit any direct information about  $S$  at all, only indirectly through  $\text{GRR}^{\delta^2}(U)$ . In this case, we have

$$Q_{\text{IR}} = \begin{pmatrix} 0.3517 & 0.1483 & 0.3517 & 0.1483 \\ 0.1483 & 0.3517 & 0.1483 & 0.3517 \\ 0.3517 & 0.1483 & 0.3517 & 0.1483 \\ 0.1483 & 0.3517 & 0.1483 & 0.3517 \end{pmatrix}. \tag{85}$$

Regarding the ‘true’ utility, we have  $I_{P^*}(X;Y) = 0.0718$ . Interestingly,  $Q_{\text{IR}}$  yields less utility than SRR. As we will see in Section 8, this is typical for small  $S$  and  $U$ .

### 8. Experiments

In order to gain insight into the behavior of the different mechanisms, we performed several experiments, both on synthetic and real data. We compared the three mechanisms introduced in this paper (PolyOpt, SRR, and IR). Throughout, we let  $\mathcal{F}$  be a confidence set for a  $\chi^2$ -test, i.e., for a Rényi divergence with  $\alpha = 2$ . We used the results of Section 4 to find explicit expressions for  $L_{u|s}(\mathcal{F})$  and (an upper bound for)  $\text{rad}_s(\mathcal{F})$ . Recall from Section 3 that

$$\mathcal{F} = \left\{ P \in \mathcal{P}_{\mathcal{X}} : D_2(\hat{P}||P) \leq \log \left( 1 + \frac{F_{\chi^2, a-1}^{-1}(1-\beta)}{n} \right) \right\}, \tag{86}$$

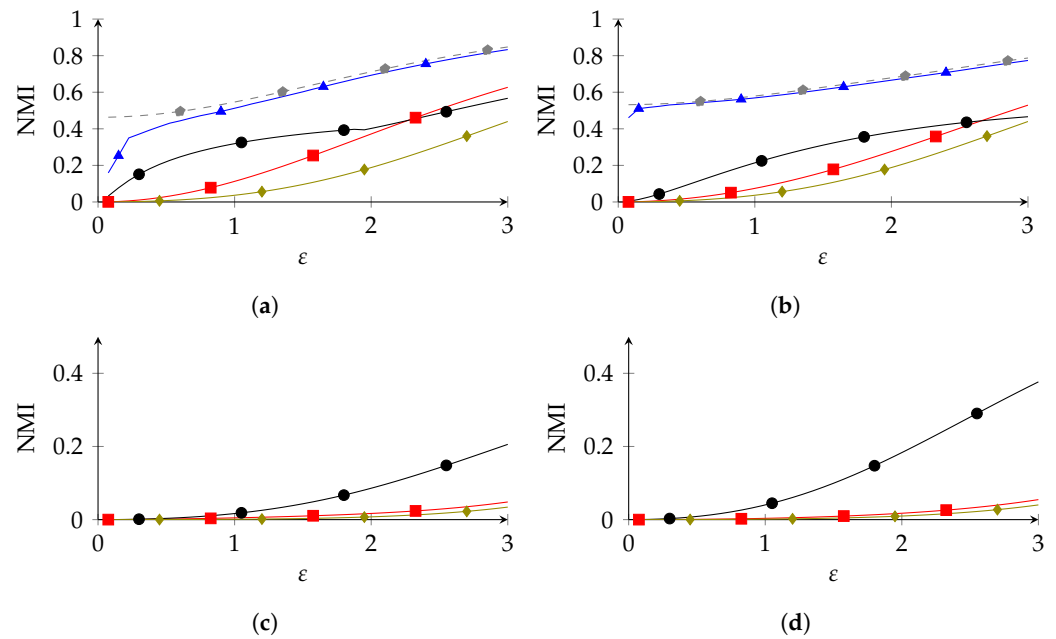
where  $F_{\chi^2, a-1}$  is the cumulative density function of the  $\chi^2$ -distribution with  $a - 1$  degrees of freedom, and  $\beta \in (0, 1)$  is a chosen significance level. Throughout the experiments, we took  $\beta = 0.05$ , unless otherwise specified.

We used  $I_{\hat{P}}(X;Y)$  as a utility metric, divided by  $H(X)$  to obtain the normalized mutual information (NMI). We used this rather than  $I_{P^*}(X;Y)$ , as the aggregator only has access to the former. In fact, while  $P^*$  is known for the synthetic data, this is not the case for real data, so we cannot even use  $I_{P^*}(X;Y)$  as a utility metric.

We compared our methods to two existing approaches, each with a slightly different privacy model. First, we compared to an LDP mechanism, to see to what extent the RLDP framework offered a utility improvement over regular LDP. As the LDP mechanism, we chose GRR, because it optimizes  $I_P(X;Y)$ , our privacy metric, in the low-privacy regime [5]. Second, we compared to the non-robust optimal mechanism of [3]. This mechanism is obtained in a manner similar to PolyOpt, and is the optimal mechanism that satisfies (in our notation)  $(\epsilon, \{\hat{P}\})$ -RLDP. In other words, it is optimal in the scenario where one knows  $P^*$  precisely. We shall refer to this mechanism as NR (non-robust). Typically, we would expect NR to have a higher utility than our RLDP mechanisms, (because it only needs to satisfy privacy with respect to. one distribution) and GRR to have worse a utility (because LDP is stricter than RLDP).

#### 8.1. Adult Data Set

We performed numerical experiments on the adult data set ( $n = 32,561$ ) [60], which contains demographic data from the 1994 US census. Some examples, where we used different categorical attributes from the data set as  $S$  and  $U$ , are depicted in Figure 2. We omitted PolyOpt from the larger two experiments, as the space complexity became unfeasible: for *occupation* vs. *education*, the polyhedron  $\hat{\Gamma}$  was 240-dimensional and was defined by 57,840 inequality constraints; to find its set of vertices Matlab needed to operate on a  $57,840 \times 57,840$  matrix, whose size (24.7 GB) exceeded Matlab’s maximum array size.

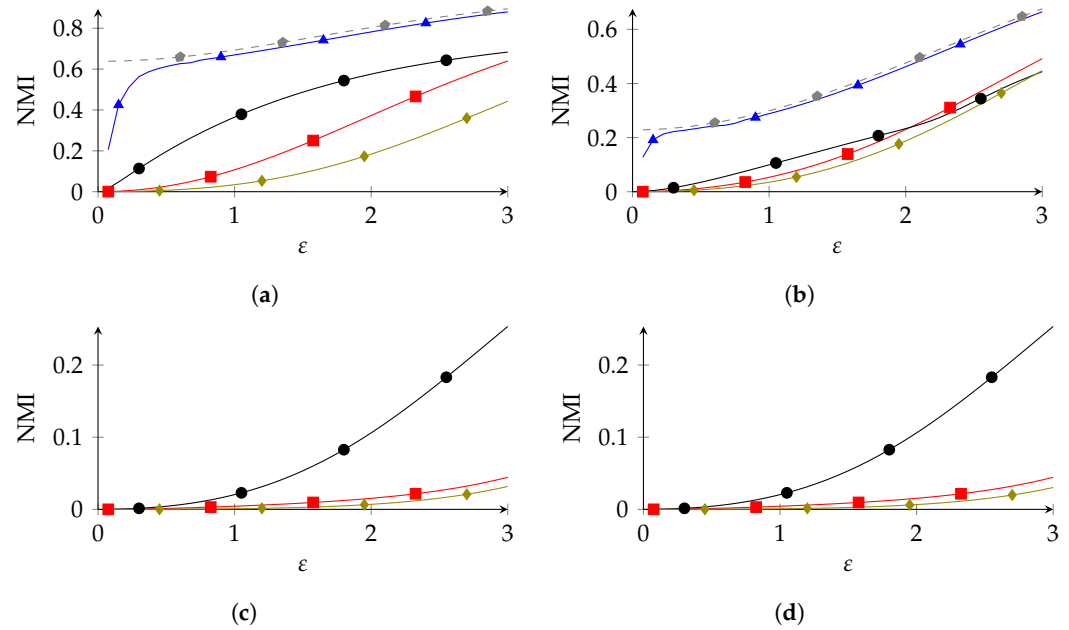


**Figure 2.** Experiments on the categories *sex*, *race*, *education*, *occupation*, *relationship* and *native-country* of the adult data set. Numbers between brackets indicate  $a_1$  and  $a_2$  (—■— SRR, —▲— PolyOpt, —●— IR, —◆— GRR, - -●- NR). (a)  $S = \text{sex}$  (2),  $U = \text{race}$  (5), (b)  $S = \text{race}$  (5),  $U = \text{sex}$  (2), (c)  $S = \text{occ.}$  (15),  $U = \text{edu.}$  (16), (d)  $S = \text{native country}$  (42),  $U = \text{relationship}$  (6).

We can see that PolyOpt clearly outperformed IR and SRR in the first two experiments, especially in the high-privacy regime (low  $\epsilon$ ). Similarly, IR outperformed SRR in the high-privacy regime, but was slightly overtaken for high  $\epsilon$ . This is interesting, since SRR satisfies a stronger privacy guarantee, as it provides privacy for all adversary assumptions, so we expected it to offer less utility than IR. An explanation for this is that IR is forced to transmit  $S$  and  $U$  separately, and so it can be less efficient than SRR, which does not have this restriction. At any rate, the difference between IR and SRR in the low-privacy regime was only marginal compared to the advantage of PolyOpt over both. In the second two experiments, where PolyOpt was infeasible, we can see that IR clearly outperformed SRR. Overall, we see that, especially in the low-privacy regime, PolyOpt was the preferable RLDP mechanism, followed by IR and SRR. Furthermore, we can see that, in all experiments, GRR performed the worst, and the best RLDP mechanism significantly outperformed GRR. This shows that adopting RLDP as a privacy metric results in significantly better utility over LDP. Conversely, NR outperformed the RLDP methods, although the difference between NR and PolyOpt was marginal for higher  $\epsilon$ . As for PolyOpt, NR was computationally out of reach for larger  $|\mathcal{X}|$ .

### 8.2. Synthetic Data

To study the robustness of our method with respect to utility (Section 8.4) and privacy (Section 8.3), we also needed experiments in which  $P^*$  was known. For this, we considered experiments on synthetic data. For this, we first randomly created a probability distribution  $P^*$  on  $\mathcal{X}$ , where  $\mathcal{X}$  was the same as in the experiments on the adult data set. The distribution  $P^*$  was drawn from the Jeffreys prior on  $\mathcal{P}_{\mathcal{X}}$ , i.e., the symmetric Dirichlet distribution with parameter  $\frac{1}{2}$ . From  $P^*$ , we then drew  $n = 32,561$  elements of  $\mathcal{X}$ , which we used to obtain the estimate  $\hat{P}$ ; this estimate was then used to create the privacy mechanisms. We carried this out 100 times, and we averaged the NMI of these 100 distributions. The results are shown in Figure 3. The results were similar to those of the experiments of the adult data set: PolyOpt outperformed IR, which outperformed SRR, for small  $|\mathcal{X}|$  SRR could overtake IR in the low-privacy regime. Furthermore, GRR was the worst overall, while NR was the best overall, but only by a small margin.



**Figure 3.** Synthetic experiments with  $n = 32561$  and  $\beta = 0.05$  (—■— SRR, —▲— PolyOpt, —●— IR, —◆— GRR, —●— NR). (a)  $a_1 = 2, a_2 = 5$ , (b)  $a_1 = 5, a_2 = 2$ , (c)  $a_1 = 15, a_2 = 16$ , (d)  $a_1 = 42, a_2 = 6$ .

### 8.3. Realized Privacy Parameter

In the previous subsections, we saw that NR had a (marginally) better utility than PolyOpt. However, this is not a completely fair comparison, since NR was only designed to give privacy for  $X \sim \hat{P}$  and might result in a larger privacy leakage for  $X \sim P^*$ . For the synthetic data,  $P^*$  was known, and we could measure the true privacy leakage. For a protocol  $\mathcal{Q}$ , we defined the *realized privacy parameter*  $\epsilon^*$  as

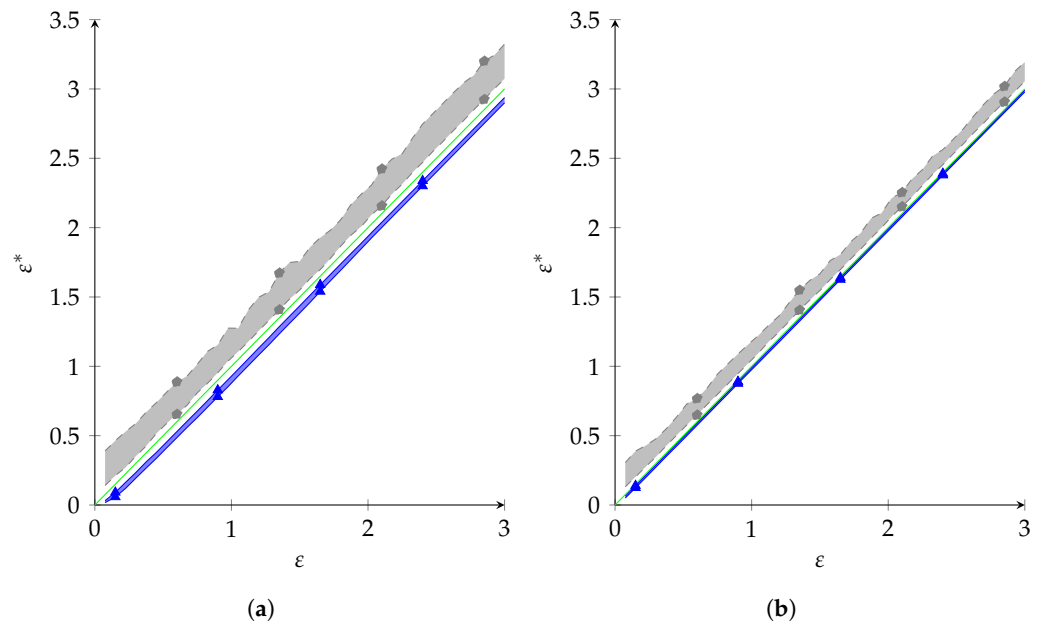
$$\begin{aligned} \epsilon^* &= \max_{\substack{y \in \mathcal{Y}, \\ s_1, s_2 \in \mathcal{S}}} \frac{\mathbb{P}_{X \sim P^*}(Y = y \mid S = s_1)}{\mathbb{P}_{X \sim P^*}(Y = y \mid S = s_2)} \\ &= \max_{\substack{y \in \mathcal{Y}, \\ s_1, s_2 \in \mathcal{S}}} \frac{\sum_u Q_{y|s_1, u} P_{u|s_1}^*}{\sum_u Q_{y|s_2, u} P_{u|s_2}^*}. \end{aligned}$$

Note that this becomes  $\infty$  when there exist  $s, y$  such that  $\mathbb{P}_{X \sim P^*}(Y = y \mid S = s) = 0$ . We compared  $\epsilon^*$  for NR and PolyOpt: the results are shown in Figure 4, where we give the 25% and 75% quantiles for both protocols, out of 100 considered distributions. As one can see, NR's  $\epsilon^*$  was consistently greater than  $\epsilon$ , while PolyOpt's  $\epsilon^*$  was consistently lesser. This is what we expected, as NR does not give privacy guarantees for  $P^*$ , but PolyOpt does when  $P^* \in \mathcal{F}$ , which happens with 95% probability. Note that the privacy leakage was especially bad for low  $\epsilon$ : at  $\epsilon = 0.075$ , the lowest value of  $\epsilon$  we tested, the 75%-quantile of  $\epsilon^*$  of NR was 0.3897, which is more than 5 times the desired privacy parameter. Overall, we can conclude that NR gave marginally better utility, but this came at quite a privacy cost.

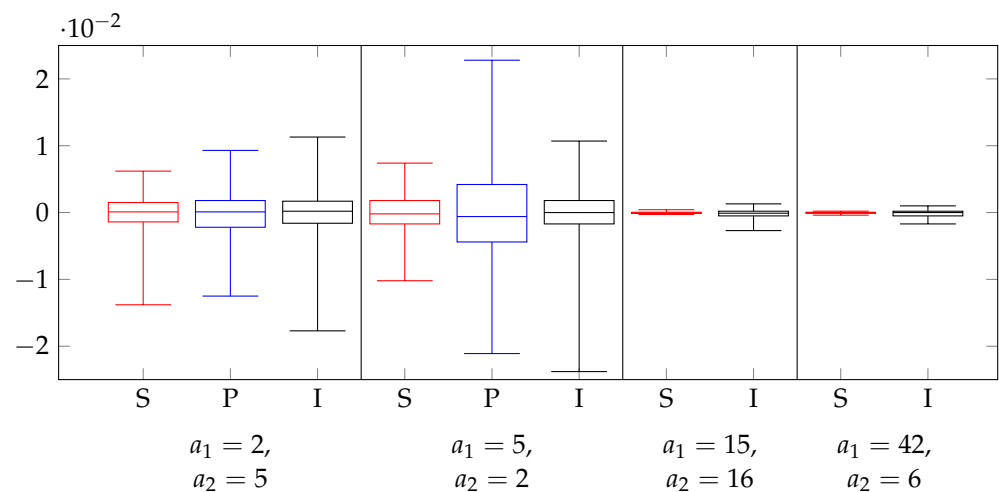
### 8.4. Utility Robustness

For the synthetic data sets (where we knew  $P^*$ ), we also investigated the normalized difference in mutual information  $\frac{I_{\hat{P}}(X; Y) - I_{P^*}(X; Y)}{I_{\hat{P}}(X; Y)}$ , to see to what extent we could use  $I_{\hat{P}}(X; Y)$  as a utility metric in lieu of the true utility  $I_{P^*}(X; Y)$ . This is shown for the three methods in Figure 5, at  $\epsilon = 1.5$ . Overall, we can see that the difference was quite minor: for all three methods, the difference in NMI, even at its most extreme, was less than 3% of the NMI value. Furthermore, the differences were very symmetric, with the difference being

positive and negative approximately equally often. We can conclude that we were justified in using  $I_{\hat{P}}(X; Y)$  as a utility metric in the other experiments.



**Figure 4.** Realized privacy parameter  $\epsilon^*$  on synthetic data. Shaded area is bounded by the 25% and 75% quantiles ( $\blacktriangle$  Polyopt,  $\bullet$  NR). The green line depicts  $\epsilon = \epsilon^*$ . (a)  $a_1 = 2, a_2 = 5$ , (b)  $a_1 = 5, a_2 = 2$ .



**Figure 5.** Normalized difference in NMI for  $\hat{P}$  and  $P^*$  on synthetic data ( $\epsilon = 1.5$ ), measured over 100 runs. Box denotes 25–75%-quantiles, whiskers denote minima and maxima. S = SRR, P = PolyOpt, I = IR.

### 8.5. Impact of $\beta$

We also considered the impact of  $\beta$  on utility for synthetic data (fixing  $\epsilon = 1.5$ ). The results are shown in Table 2, which are averages over 100 runs. Note that SRR does not depend on  $\beta$ , since it assumes  $\mathcal{F} = \mathcal{P}_{\mathcal{X}}$ . Interestingly, we can see that the impact of  $\beta$  was quite limited; changing  $\beta$  by a factor 100 had at most about 4% impact on NMI. This impact was less for PolyOpt than for IR, and less for larger  $\mathcal{X}$ . Overall, we can conclude that by choosing  $\beta$  closer to 0, we can significantly increase the robustness of privacy without making a considerable impact on utility.



**Table 2.** NMI for synthetic data for various values of  $\beta$  ( $\epsilon = 1.5$ ).

	$a_1 = 2, a_2 = 5$			$a_1 = 5, a_2 = 2$		
$\beta$	0.1	0.01	0.001	0.1	0.01	0.001
SRR	0.231	0.231	0.231	0.126	0.126	0.126
PolyOpt	0.727	0.723	0.719	0.374	0.372	0.370
IR	0.512	0.501	0.492	0.169	0.165	0.162
	$a_1 = 15, a_2 = 16$			$a_1 = 42, a_2 = 6$		
$\beta$	0.1	0.01	0.001	0.1	0.01	0.001
SRR	0.009	0.009	0.009	0.005	0.005	0.005
IR	0.055	0.053	0.051	0.052	0.052	0.052

## 9. Conclusions and Future Work

In this paper, we presented a number of algorithms that, given a desired privacy level  $\epsilon$ , an estimated distribution  $\hat{P}$ , and a bound on the Rényi divergence  $D_\alpha(\hat{P}||P)$ , return privacy mechanisms that satisfy a differential privacy-like privacy constraint for the part of the data that is considered sensitive, for all distributions  $P$  within the divergence bound. The first class of privacy mechanisms, PolyOpt, offers high utility, but is computationally complex, as it relies on vertex enumeration. The second class, SRR, satisfies a stronger privacy requirement and is optimal in the low-privacy regime with reference to this requirement, but as a result has less utility than mechanisms that do not satisfy this stronger privacy requirement. The third class, IR, is a general framework for releasing the sensitive and non-sensitive part of the data independently, and the optimal division of the privacy budget between these can be found via 1-dimensional optimization; thus, the optimal IR mechanism can be found quickly, while still offering decent utility. Furthermore, taking RLDP rather than LDP as a privacy constraint, i.e., protecting only the part of the data that is sensitive, significantly improves utility. In particular, we showed that the utility of PolyOpt is close to the utility of the optimal non-robust privacy mechanism. In other words, asking for robustness in privacy comes at only a small performance penalty in utility. At the same time, we showed that *not* asking for robustness comes at a substantial privacy cost.

There are various interesting directions for future research to build upon the results in this paper. One direction is to find analytical bounds on the performance gap between PolyOpt and optimal mechanisms, in particular on the gap with reference to either the non-robust optimal mechanism from [3] or with reference to an optimal robust mechanism. Note, however, that for the moment we do not have any results on optimal robust mechanisms. Another direction is to improve the performance of the low-complexity algorithms that have been proposed. For instance, in independent reporting, one could change the underlying LDP mechanism from GRR to an optimal mechanism. Since GRR is only optimal in the high-privacy regime, we expect that there would be room for improvement in the low-privacy regime. A significant challenge is incorporating optimal mechanisms along the lines of [5]; however, these mechanisms depend on  $P^*$  which is inaccessible in the RLDP framework. Yet another interesting direction would be to incorporate robustness in utility in addition to robustness in privacy. This would require finding a mechanism that maximizes  $\min_{P \in \mathcal{F}} I_P(X; Y)$ . The challenge in this is that  $I_P(X; Y)$  is concave in  $P$ , which makes minimizing it over  $\mathcal{F}$  difficult. Finally, it would be interesting to apply the RLDP framework to other models. In this work, we studied the model where  $X$  splits into a sensitive part  $S$  and a non-sensitive part  $U$ . It would be interesting to also study the more general case where  $X$  is correlated with the sensitive data  $S$ , or to apply RLDP to the models that are studied in [12].

**Author Contributions:** Conceptualization, M.L.-Z. and J.G.; Formal analysis, M.L.-Z. and J.G.; Investigation, M.L.-Z. and J.G.; Methodology, M.L.-Z. and J.G.; Software, M.L.-Z.; Writing, M.L.-Z. and J.G. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was partially funded by the Netherlands Organisation for Scientific Research (NWO) grant 628.001.026, ERC Consolidator grant 864075 CAESAR and the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 101008233.

**Institutional Review Board Statement:** Not applicable.

**Data Availability Statement:** Data are contained within the article.

**Conflicts of Interest:** The authors declare no conflicts of interest.

**Appendix A. Proofs**

*Appendix A.1. Proof of Theorem 1*

This follows from the following four lemmas, where the RHS of (24) is denoted  $\overline{\mathcal{F}}_{\mathcal{U}|s}$ :

**Lemma A1.** *If  $\alpha \neq 1$ , then  $\mathcal{F}_{\mathcal{U}|s} \subset \overline{\mathcal{F}}_{\mathcal{U}|s}$ .*

**Proof.** Assume  $\alpha < 1$ ; the case  $\alpha > 1$  is handled analogously. Then, we rewrite  $D_\alpha(\hat{P}||P) \leq B$  as

$$\sum_x \frac{\hat{P}_x^\alpha}{P_x^{\alpha-1}} \geq e^{B(\alpha-1)}. \tag{A1}$$

Let  $C = e^{B(\alpha-1)}$ . Then,

$$\frac{\hat{P}_s^\alpha}{P_s^{\alpha-1}} \sum_u \frac{\hat{P}_{u|s}^\alpha}{P_{u|s}^{\alpha-1}} = \sum_u \frac{\hat{P}_{s,u}^\alpha}{P_{s,u}^{\alpha-1}} \tag{A2}$$

$$\geq C - \sum_{s' \neq s} \sum_u \frac{\hat{P}_{s',u}^\alpha}{P_{s',u}^{\alpha-1}}. \tag{A3}$$

For  $s' \in \mathcal{S} \setminus \{s\}$  and  $u \in \mathcal{U}$ , define  $P_{s',u|s} = \frac{P_{us'}}{1-P_s}$  and  $\hat{P}_{s',u|s} = \frac{\hat{P}_{us'}}{1-P_s}$ . Then, (A3) can be written as

$$\frac{\hat{P}_s^\alpha}{P_s^{\alpha-1}} \sum_u \frac{\hat{P}_{u|s}^\alpha}{P_{u|s}^{\alpha-1}} \geq C - \frac{(1-\hat{P}_s)^\alpha}{(1-P_s)^{\alpha-1}} \sum_{s' \neq s} \sum_u \frac{\hat{P}_{s',u|s}^\alpha}{P_{s',u|s}^{\alpha-1}}. \tag{A4}$$

Furthermore,  $P_{\bullet|s} = (P_{s',u|s})_{s' \in \mathcal{S} \setminus \{s\}, u \in \mathcal{U}}$  and  $\hat{P}_{\bullet|s} = (\hat{P}_{s',u|s})_{s' \in \mathcal{S} \setminus \{s\}, u \in \mathcal{U}}$  form probability distributions on  $(\mathcal{S} \setminus \{s\}) \times \mathcal{U}$ . As such, we have

$$\sum_u \frac{\hat{P}_{s',u|s}^\alpha}{P_{s',u|s}^{\alpha-1}} = e^{(\alpha-1)D_\alpha(\hat{P}_{\bullet|s}||P_{\bullet|s})} \leq 1. \tag{A5}$$

Applying this to (A4), we obtain

$$\frac{\hat{P}_s^\alpha}{P_s^{\alpha-1}} \sum_u \frac{\hat{P}_{u|s}^\alpha}{P_{u|s}^{\alpha-1}} \geq C - \frac{(1-\hat{P}_s)^\alpha}{(1-P_s)^{\alpha-1}} \tag{A6}$$

or

$$\sum_u \frac{\hat{P}_{u|s}^\alpha}{P_{u|s}^{\alpha-1}} \geq \frac{P_s^{\alpha-1}}{\hat{P}_s^\alpha} \left( C - \frac{(1-\hat{P}_s)^\alpha}{(1-P_s)^{\alpha-1}} \right). \tag{A7}$$

To find the bound on  $\sum_u \frac{\hat{P}_{u|s}^\alpha}{P_{u|s}^{\alpha-1}}$ , we have to minimize the RHS of this inequality. The only unknown on the right is  $P_s$ . We find the minimum value of the right-hand side by differentiating with respect to  $P_s$ , for which we obtain

$$(\alpha - 1) \frac{P_s^{\alpha-2}(1 - \hat{P}_s)^\alpha}{\hat{P}_s^\alpha} \left( \frac{C}{(1 - \hat{P}_s)^\alpha} - \frac{1}{(1 - P_s)^\alpha} \right). \tag{A8}$$

Setting this equal to 0, we find  $P_s = 1 - C^{-1/\alpha}(1 - \hat{P}_s)$ . Substituting this into (A7), we obtain

$$\sum_u \frac{\hat{P}_{u|s}^\alpha}{P_{u|s}^{\alpha-1}} \geq \frac{(C^{1/\alpha} - (1 - \hat{P}_s))^\alpha}{\hat{P}_s^\alpha} \tag{A9}$$

which can be written as

$$D_\alpha(\hat{P}_{\mathcal{U}|s} || P_{\mathcal{U}|s}) \leq \frac{\alpha}{\alpha-1} \log \left( \frac{e^{(\alpha-1)B/\alpha} - (1 - \hat{P}_s)}{\hat{P}_s} \right), \tag{A10}$$

showing that  $P_{\mathcal{U}|s} \in \overline{\mathcal{F}}_{\mathcal{U}|s}$ . Since  $P$  was chosen arbitrarily, we can conclude  $\mathcal{F}_{\mathcal{U}|s} \subset \overline{\mathcal{F}}_{\mathcal{U}|s}$ .  $\square$

**Lemma A2.** *If  $\alpha \neq 1$ , then  $\overline{\mathcal{F}}_{\mathcal{U}|s} \subset \mathcal{F}_{\mathcal{U}|s}$ .*

**Proof.** Again we assume  $\alpha < 1$ . Suppose that  $R \in \mathcal{P}_{\mathcal{U}}$  satisfies  $D_\alpha(\hat{P}_{\mathcal{U}|s} || R) \leq B_s$ . Let  $C$  be as in (A1) and define  $\gamma = 1 - C^{-1/\alpha}(1 - \hat{P}_s)$ ; then,

$$\frac{1}{\alpha - 1} \log \left( \sum_u \frac{\hat{P}_{u|s}^\alpha}{R_u^{\alpha-1}} \right) = D_\alpha(\hat{P}_{\mathcal{U}|s} || R) \tag{A11}$$

$$\leq B_s \tag{A12}$$

$$= \frac{\alpha}{\alpha - 1} \log \left( \frac{e^{(\alpha-1)B/\alpha} - (1 - \hat{P}_s)}{\hat{P}_s} \right) \tag{A13}$$

$$= \frac{\alpha}{\alpha - 1} \log \left( \frac{C^{1/\alpha} \gamma}{\hat{P}_s} \right) \tag{A14}$$

which we can express as

$$\sum_u \frac{\hat{P}_{u|s}^\alpha}{R_u^{\alpha-1}} \geq \frac{C \gamma^\alpha}{\hat{P}_s^\alpha}. \tag{A15}$$

Define  $P \in \mathcal{P}_{\mathcal{X}}$  by

$$P_{u,s'} = \begin{cases} \gamma R_{u,s} & \text{if } s' = s, \\ C^{-1/\alpha} \hat{P}_{u,s'} & \text{otherwise.} \end{cases} \tag{A16}$$

Then,  $P_{\mathcal{U}|s} = R$ , and

$$\sum_{u,s'} \frac{\hat{P}_{u,s'}^\alpha}{P_{u,s'}^{\alpha-1}} = \sum_u \frac{\hat{P}_{u,s}^\alpha}{\gamma^{\alpha-1} R_u^{\alpha-1}} + \sum_u \sum_{s' \neq s} C^{1/\alpha} \hat{P}_{u,s'} \tag{A17}$$

$$= \frac{\hat{P}_s^\alpha}{\gamma^{\alpha-1}} \sum_u \frac{\hat{P}_{u|s}^\alpha}{R_u^{\alpha-1}} + C^{(\alpha-1)/\alpha} (1 - \hat{P}_s) \tag{A18}$$

$$\geq \gamma C + C^{(\alpha-1)/\alpha} (1 - \hat{P}_s) \tag{A19}$$

$$= C. \tag{A20}$$

As in the proof of Lemma A1, the condition  $\sum_{u,s'} \frac{\hat{P}_{u,s'}^\alpha}{P_{u,s'}^{\alpha-1}} \geq C$  is equivalent to  $D_\alpha(\hat{P} || P) \leq B$ . Thus, we can conclude that  $P \in \mathcal{F}$  and so  $R = P_{\mathcal{U}|s} \in \mathcal{F}_{\mathcal{U}|s}$ . Since  $R$  was chosen arbitrary, this shows  $\overline{\mathcal{F}}_{\mathcal{U}|s} \subset \mathcal{F}_{\mathcal{U}|s}$ .  $\square$

**Lemma A3.** If  $\alpha = 1$ , then  $\mathcal{F}_{\mathcal{U}|s} \subset \overline{\mathcal{F}}_{\mathcal{U}|s}$ .

**Proof.** Let  $P \in \mathcal{F}$ , and define  $P_{\bullet|\neg s}, \hat{P}_{\bullet|\neg s}$  as in the proof of Lemma A1. Then,

$$D_1(\hat{P}|P) = \hat{p}_s \sum_u \hat{p}_{u|s} \log \frac{\hat{p}_s \hat{p}_{u|s}}{P_s P_{u|s}} + (1 - \hat{p}_s) \sum_{s',u} \hat{p}_{u,s'|\neg s} \frac{(1 - \hat{p}_s) \hat{p}_{u,s'|\neg s}}{(1 - P_s) P_{u,s'|\neg s}} \tag{A21}$$

$$= \hat{p}_s D_1(\hat{P}_{\mathcal{U}|s} || P_{\mathcal{U}|s}) + (1 - \hat{p}_s) D_1(\hat{P}_{\bullet|\neg s} || P_{\bullet|\neg s}) + \hat{p}_s \log \frac{\hat{p}_s}{P_s} + (1 - \hat{p}_s) \log \frac{1 - \hat{p}_s}{1 - P_s} \tag{A22}$$

$$= \hat{p}_s D_1(\hat{P}_{\mathcal{U}|s} || P_{\mathcal{U}|s}) + (1 - \hat{p}_s) D_1(\hat{P}_{\bullet|\neg s} || P_{\bullet|\neg s}) + D_1(V_{\hat{p}_s} || V_{P_s}), \tag{A23}$$

where for  $p \in [0, 1]$ , the random variable  $V_p$  is defined to follow a Bernoulli distribution with  $\mathbb{P}(V_p = 1) = p$ . Since  $D_1$  is non-negative and  $D_1(\hat{P}|P) \leq B$ , we find

$$D_1(\hat{P}_{\mathcal{U}|s} || P_{\mathcal{U}|s}) = \frac{1}{\hat{p}_s} \left( D_1(\hat{P}|P) - (1 - \hat{p}_s) D_1(\hat{P}_{\bullet|\neg s} || P_{\bullet|\neg s}) - D_1(V_{\hat{p}_s} || V_{P_s}) \right) \tag{A24}$$

$$\leq \frac{D_1(\hat{P}|P)}{\hat{p}_s} \tag{A25}$$

$$\leq \frac{B}{\hat{p}_s}. \tag{A26}$$

Thus,  $P_{\mathcal{U}|s} \in \overline{\mathcal{F}}_{\mathcal{U}|s}$ ; since  $P \in \mathcal{F}$  was chosen arbitrary, we can conclude  $\mathcal{F}_{\mathcal{U}|s} \subset \overline{\mathcal{F}}_{\mathcal{U}|s}$ .  $\square$

**Lemma A4.** If  $\alpha = 1$ , then  $\overline{\mathcal{F}}_{\mathcal{U}|s} \subset \mathcal{F}_{\mathcal{U}|s}$ .

**Proof.** Let  $R \in \mathcal{P}_{\mathcal{U}}$  be such that  $D_1(\hat{P}_{\mathcal{U}|s} || R) \leq \frac{B}{\hat{p}_s}$ . Define  $P \in \mathcal{P}_{\mathcal{X}}$  by

$$P_{u,s'} = \begin{cases} \hat{p}_s R_u, & \text{if } s = s', \\ \hat{p}_{u,s'} & \text{if } s \neq s'. \end{cases} \tag{A27}$$

Then  $P_{\mathcal{U}|s} = R$ . Furthermore, in (A23) one has  $D_1(\hat{P}_{\bullet|\neg s} || P_{\bullet|\neg s}) = D_1(V_{\hat{p}_s} || V_{P_s}) = 0$ , and so  $D_1(\hat{P}|P) \leq B$ . This shows that  $P \in \mathcal{F}$ , and so  $R = P_{\mathcal{U}|s} \in \mathcal{F}_{\mathcal{U}|s}$ . Since  $R$  was chosen arbitrarily, we can conclude that  $\overline{\mathcal{F}}_{\mathcal{U}|s} \subset \mathcal{F}_{\mathcal{U}|s}$ .  $\square$

*Appendix A.2. Proof of Proposition 1*

We first prove the following two auxiliary lemmas. We only prove these for  $\alpha > 1$ ; the other cases are handled analogously.

**Lemma A5.** Let  $x \in \mathcal{X}$ , and define

$$\xi_-(\rho) = \inf \left\{ \xi \in (0, 1] : E_B(\rho, \xi) \leq 0 \right\}, \tag{A28}$$

$$\xi_+(\rho) = \sup \left\{ \xi \in [1, (1 - \rho)^{-1}] : E_B(\rho, \xi) \leq 0 \right\}, \tag{A29}$$

where  $E_B$  is as in Proposition 1. Then,  $\min_{P \in \mathcal{F}} P_x = \hat{P}_x \xi_-(\hat{P}_x)$  and  $\max_{P \in \mathcal{F}} P_x = \hat{P}_x \xi_+(\hat{P}_x)$ .

**Proof.** As in the proof of Lemma A1, define  $C = e^{(\alpha-1)B}$ ; thus

$$\mathcal{F} = \left\{ P \in \mathcal{P}_{\mathcal{X}} : \sum_{x' \in \mathcal{X}} \frac{\hat{p}_{x'}^\alpha}{P_{x'}^{\alpha-1}} \leq C \right\}. \tag{A30}$$

Furthermore, define a function  $F$  by

$$F(\rho, \xi) = \rho \xi^{1-\alpha} + (1 - \rho) \left( \frac{1 - \rho \xi}{1 - \rho} \right)^{1-\alpha} - C, \tag{A31}$$

with  $F(1, \xi) = \xi^{1-\alpha} - C$  the limit as  $\rho \rightarrow 1$ . Then,  $E_B(\rho, \xi) = \frac{1}{\alpha-1} \log(F(\rho, \xi) + e^{(\alpha-1)B}) - B$ , so  $F(\rho, \xi) \leq 0 \Leftrightarrow E_B(\rho, \xi) \leq 0$ . Thus,  $\xi_-(\rho) = \inf\{\xi \in [0, 1]: F(\rho, \xi) \leq 0\}$  and the analogous statement holds for  $\xi_+(\rho)$ . The  $P$  that yield the extremal  $P_x$  lie on the boundary of  $\mathcal{F}$ ; hence, they either satisfy  $P_x \in \{0, 1\}$ , or the equality

$$\sum_{x' \in \mathcal{X}} \frac{\hat{P}_{x'}^\alpha}{P_{x'}^{\alpha-1}} = C. \tag{A32}$$

In the latter case, the extremal values of  $P_x$  have to be stationary points of the Lagrangian expression

$$P_x + \lambda \left( \sum_{x'} \frac{\hat{P}_{x'}^\alpha}{P_{x'}^{\alpha-1}} - C \right) + \mu \left( \sum_{x'} P_{x'} - 1 \right) = 0. \tag{A33}$$

Taking derivatives with respect to all  $P_x$ , we find

$$1 + (1 - \alpha)\lambda \frac{\hat{P}_x^\alpha}{P_x^\alpha} + \mu = 0, \tag{A34}$$

$$\forall x' \neq x: (1 - \alpha)\lambda \frac{\hat{P}_{x'}^\alpha}{P_{x'}^\alpha} + \mu = 0. \tag{A35}$$

It follows that  $P_x = \left(\frac{\alpha-1}{\mu}\lambda\right)^{1/\alpha} \hat{P}_x =: \xi \hat{P}_x$  and  $P_{x'} = \left(\frac{\alpha-1}{\mu}\lambda\right)^{1/\alpha} \hat{P}_{x'} =: \psi \hat{P}_{x'}$  for all  $x' \neq x$ , where  $\xi$  and  $\psi$  do not depend on  $x$  or  $x'$ . We can find  $\xi, \psi \in \mathbb{R}_{\geq 0}$  by solving the joint set of equations

$$C = \sum_{x'} \frac{\hat{P}_{x'}^\alpha}{P_{x'}^{\alpha-1}} \tag{A36}$$

$$= \frac{\hat{P}_x^\alpha}{P_x^{\alpha-1}} + \sum_{x' \neq x} \frac{\hat{P}_{x'}^\alpha}{P_{x'}^{\alpha-1}} \tag{A37}$$

$$= \hat{P}_x \xi^{1-\alpha} + (1 - \hat{P}_x) \psi^{1-\alpha}, \tag{A38}$$

$$1 = \sum_{x'} P_{x'} \tag{A39}$$

$$= \hat{P}_x \xi + (1 - \hat{P}_x) \psi. \tag{A40}$$

Define  $\rho = \hat{P}_x$ . Then, (A40) implies  $\psi = \frac{1-\rho\xi}{1-\rho}$ , and the condition  $\psi \geq 0$  is equivalent to  $\xi \leq \rho^{-1}$ . Substituting this into (A38) shows that we find  $\xi$  by solving  $F(\rho, \xi) = 0$  for  $\xi \in (0, (1 - \rho)^{-1})$ . Since  $F(\rho, 1) = 1 - C < 0$  and  $F$  is strictly convex in  $\xi$ , there exists, at most, one solution in  $(0, 1]$  and, at most, one in  $[1, (1 - \rho)^{-1})$ . It follows that (A33) has, at most, two stationary points, which must correspond to the minimal and maximal value of  $P_x$ . If the solution in  $(0, 1]$  exists, it is equal to  $\xi_-(\rho)$ , and this stationary point of (A33) corresponds to the minimal value of  $P_x$ , which is then equal to  $\hat{P}_x \xi_-(\hat{P}_x)$ . If the solution in  $(0, 1]$  does not exist, then the minimal value of  $P_x$  is not attained on the boundary and is equal to 0, which then is also equal to  $\hat{P}_x \xi_-(\hat{P}_x)$ . Either way, we find

$$\min_{P \in \mathcal{F}} P_x = \hat{P}_x \xi_-(\hat{P}_x). \tag{A41}$$

The proof for the maximal value of  $P_x$  is analogous.  $\square$

**Lemma A6.** For  $\mathcal{X}_1 \subset \mathcal{X}$  define  $\hat{P}_{\mathcal{X}_1} := \sum_{x \in \mathcal{X}_1} \hat{P}_x$ . Then,

$$\sup_{P \in \mathcal{F}} \|P - \hat{P}\|_1 = 2 \max_{\substack{\mathcal{X}_1 \subset \mathcal{X}: \\ \mathcal{X}_1 \neq \emptyset}} \hat{P}_{\mathcal{X}_1} (\zeta_+(\hat{P}_{\mathcal{X}_1}) - 1). \tag{A42}$$

**Proof.** For a given  $P$ , define  $\mathcal{X}_1 = \{x \in \mathcal{X} : P_x \geq \hat{P}_x\}$  and  $\mathcal{X}_2 = \{x \in \mathcal{X} : P_x < \hat{P}_x\}$ . To find the maximal value of  $\|P - \hat{P}\|_1$ , we first maximize it for a given partition  $\mathcal{X}_1, \mathcal{X}_2$  of  $\mathcal{X}$ , and then we maximize over all partitions. Note that  $\mathcal{X}_1 = \emptyset$  is impossible, and for  $\mathcal{X}_1 = \mathcal{X}$ , we have  $P = \hat{P}$ , which is certainly not optimal. Given  $\mathcal{X}_1, \mathcal{X}_2$ , one has

$$\|P - \hat{P}\|_1 = \sum_{x \in \mathcal{X}_1} (P_x - \hat{P}_x) + \sum_{x \in \mathcal{X}_2} (\hat{P}_x - P_x). \tag{A43}$$

As before, the  $P$  maximizing this lies either on the boundary of the probability simplex or it satisfies (A32). For the latter case, we have the Lagrangian expression

$$\sum_{x \in \mathcal{X}_1} (P_x - \hat{P}_x) + \sum_{x \in \mathcal{X}_2} (\hat{P}_x - P_x) + \lambda \left( \sum_x \frac{\hat{P}_x^\alpha}{P_x^{\alpha-1}} - C \right) + \mu \left( \sum_x P_x - 1 \right) = 0. \tag{A44}$$

Taking derivatives, we find, analogously to (A34)–(A35), that there exist  $\zeta, \psi$  such that  $P_x = \zeta \hat{P}_x$  for all  $x \in \mathcal{X}_1$  and  $P_x = \psi \hat{P}_x$  for all  $x \in \mathcal{X}_2$ . By definition of  $\mathcal{X}_1$  and  $\mathcal{X}_2$ , we have  $\zeta \geq 1$  and  $0 \leq \psi < 1$ . Analogously to (A36)–(A40), these have to satisfy

$$\hat{P}_{\mathcal{X}_1} \zeta^{1-\alpha} + (1 - \hat{P}_{\mathcal{X}_1}) \psi^{1-\alpha} = C, \tag{A45}$$

$$\hat{P}_{\mathcal{X}_1} \zeta + (1 - \hat{P}_{\mathcal{X}_1}) \psi = 1. \tag{A46}$$

From this point onward, this proof is analogous to that of Lemma A5. Let  $\rho = \hat{P}_{\mathcal{X}_1}$ . Expressing  $\psi$  in terms of  $\zeta$  and substituting this means that to find  $\zeta$  we have to solve  $F(\rho, \zeta) = 0$  for  $\zeta \in [1, (1 - \rho)^{-1})$ , where  $F$  is as in the proof of Lemma A5. As before, at most, one such solution exists, and when it does, it corresponds to the maximal value of  $\|P - \hat{P}\|_1$  (given  $\mathcal{X}_1$ ). If it does not exist, then the maximal value of  $\|P - \hat{P}\|_1$  is obtained at the boundary where  $\hat{P}_{\mathcal{X}_1} = 1$ . Either way the maximum is obtained when  $\zeta = \zeta_+(\rho)$ , which means that

$$\|P - \hat{P}\|_1 = \sum_{x \in \mathcal{X}_1} (P_x - \hat{P}_x) + \sum_{x \in \mathcal{X}_2} (\hat{P}_x - P_x) \tag{A47}$$

$$= \sum_{x \in \mathcal{X}_1} \hat{P}_x (\zeta_+(\rho) - 1) + \sum_{x \in \mathcal{X}_2} \hat{P}_x \left( 1 - \frac{1 - \rho \zeta_+(\rho)}{1 - \rho} \right) \tag{A48}$$

$$= \rho (\zeta_+(\rho) - 1) + (1 - \rho) \left( 1 - \frac{1 - \rho \zeta_+(\rho)}{1 - \rho} \right) \tag{A49}$$

$$= 2\rho (\zeta_+(\rho) - 1). \tag{A50}$$

This is the maximal value of  $\|P - \hat{P}\|_1$  given  $\mathcal{X}_1$ ; we now find the overall maximum by maximizing over all non-empty  $\mathcal{X}_1$ .  $\square$

**Proof of Proposition 1.** In Lemmas A5 and A6, take  $\mathcal{U}$  instead of  $\mathcal{X}$ ,  $\hat{P}_{\mathcal{U}|s}$  instead of  $\hat{P}$ , and  $B_s$  instead of  $B$ . Then, by Theorem 1, the role of  $\mathcal{F}$  is taken by  $\mathcal{F}_{\mathcal{U}|s}$ . Thus, applying Lemmas A5 and A6 gives us Proposition 1 directly.  $\square$

*Appendix A.3. Proof of Lemma 1 and Proposition 2*

As in the proof of Proposition 1, since by Theorem 1 the projected set  $\mathcal{F}_{\mathcal{U}|s}$  is defined by a Rényi divergence as is  $\mathcal{F}$ , it suffices to prove the analogous statements about  $\mathcal{F}$  rather than  $\mathcal{F}_{\mathcal{U}|s}$ . Concretely, we prove the following:

**Lemma A7.** Suppose  $\alpha = 2$  and define  $\tilde{B} = e^B - 1$ ; let  $\xi_{\pm}$  be as in Lemma A5. Then,

$$\xi_{\pm}(\rho) = \frac{\tilde{B} + 2\rho \pm \sqrt{\tilde{B}^2 + 4\rho\tilde{B} - 4\tilde{B}\rho^2}}{2\rho(\tilde{B} + 1)}. \tag{A51}$$

Furthermore, the following hold:

1. Let  $x_{\min} = \arg \min_{x \in \mathcal{X}} \hat{P}_x$ . If  $\tilde{B} \geq (1 - \hat{P}_{x_{\min}})^2$ , then the maximum in (A6) is attained at  $\mathcal{X}_1 = \{x_{\min}\}$ .
2. If  $\tilde{B} < (1 - \hat{P}_{x_{\min}})^2$  one has  $\sup_{P \in \mathcal{F}} \|P - \hat{P}\|_1 \leq \sqrt{\tilde{B}}$ .

The formulas here look slightly different from those in Lemma 1 and Proposition 2. We use this form because it makes the proof more convenient: replacing  $\tilde{B}$  with  $e^B - 1$  throughout yields exactly the results of Lemma 1 and Proposition 2 for  $\mathcal{F}$  instead of  $\mathcal{F}_{\mathcal{U}|S}$ .

**Proof.** Consider the function  $F(\rho, \xi)$  from (A31) for  $\alpha = 2$  and  $C = \tilde{B} + 1$ , i.e.,

$$F(\rho, \xi) = \frac{\rho}{\xi} + \frac{(1 - \rho)^2}{1 - \rho\xi} - \tilde{B} - 1. \tag{A52}$$

Then,  $F(\rho, \xi) = 0$  can be rewritten to a quadratic equation in  $\xi$ . Its two roots are  $\xi_{\pm}(\rho)$ , and with some rewriting they can be expressed as in (31). For points 1 and 2, we note that

$$2\rho(\xi_+(\rho) - 1) = \frac{\tilde{B} - 2\tilde{B}\rho + \sqrt{\tilde{B}^2 + 4\tilde{B}\rho - 4\tilde{B}\rho^2}}{\tilde{B} + 1}. \tag{A53}$$

We can find its extremal values with respect to  $\rho$  by taking the derivative and setting it to 0, i.e., by solving

$$-\frac{2\tilde{B}}{\tilde{B} + 1} + \frac{2\tilde{B} - 4\tilde{B}\rho}{(\tilde{B} + 1)\sqrt{\tilde{B}^2 + 4\tilde{B}\rho - 4\tilde{B}\rho^2}} = 0, \tag{A54}$$

which has a single solution  $\rho_{\text{opt}} = \frac{1 - \sqrt{\tilde{B}}}{2}$ . Since (A53) is concave in  $\rho$ , this means that this unique extremal value is a maximum. If  $\tilde{B} \geq (1 - 2\hat{P}_{x_{\min}})^2$ , then  $\rho_{\text{opt}} \geq \hat{P}_{x_{\min}}$ , and  $\rho(\xi_+(\rho) - 1)$  is decreasing in  $\rho$  on  $[\hat{P}_{x_{\min}}, 1]$ . Since all possible values of  $\hat{P}_{\mathcal{X}_1}$  lie in this interval, it is optimal to take  $\mathcal{X}_1$  such that  $\hat{P}_{\mathcal{X}_1}$  is minimized, i.e.,  $\mathcal{X}_1 = \{x_{\min}\}$ ; this proves point 1. For point 2 we have (and also for general  $B$ )

$$\sup_{P \in \mathcal{F}} \|P - \hat{P}\|_1 = 2 \max_{\substack{\mathcal{X}_1 \subset \mathcal{X}: \\ \mathcal{X}_1 \neq \emptyset}} \hat{P}_{\mathcal{X}_1} (\xi_+(\hat{P}_{\mathcal{X}_1}) - 1) \tag{A55}$$

$$\leq 2\rho_{\text{opt}} (\xi_+(\rho_{\text{opt}}) - 1) \tag{A56}$$

$$= \frac{\tilde{B} - 2\tilde{B}\frac{1 - \sqrt{\tilde{B}}}{2} + \sqrt{\tilde{B}^2 + 4\tilde{B}\frac{1 - \sqrt{\tilde{B}}}{2} - 4\tilde{B}\frac{(1 - \sqrt{\tilde{B}})^2}{4}}}{\tilde{B} + 1} \tag{A57}$$

$$= \sqrt{\tilde{B}}. \tag{A58}$$

□

#### Appendix A.4. Proof of Theorem 2

Let  $\mathcal{D} = \prod_{s \in \mathcal{S}} \mathcal{D}_s \subset \mathbb{R}^{\mathcal{X}}$ . Thus, an element  $t \in \mathcal{D}$  is of the form  $t = (t_{s,u})_{(s,u) \in \mathcal{X}}$ , and for any  $s$ , we have  $(t_{s,u})_{u \in \mathcal{U}} \in \mathcal{D}_s$ . For  $s_1, s_2 \in \mathcal{S}$ , let  $B^{s_1, s_2} \in \mathbb{R}^{\mathcal{X} \times \mathcal{X}}$  be the matrix given by

$$B_{(s,u);(s',u')}^{s_1, s_2} = \begin{cases} 1, & \text{if } u = u' \text{ and } s = s' = s_1, \\ -e^\epsilon, & \text{if } u = u' \text{ and } s = s' = s_2, \\ 0, & \text{otherwise.} \end{cases} \tag{A59}$$

Then, we can rewrite (41) as

$$\forall y, s_1, s_2 : \max_{t \in \mathcal{D}} ((B^{s_1, s_2})^T Q_y)^T t \leq 0. \tag{A60}$$

Recall that for each  $s$ , we have  $\mathcal{D}_s = \{R \in \mathcal{P}_{\mathcal{U}} : R_u \geq L_{u|s}\}$ . Since  $\mathcal{D} = \prod_s \mathcal{D}_s$ , we can write

$$\mathcal{D} = \left\{ t \in \mathbb{R}^{\mathcal{X}} : \forall s, u : t_{s,u} \geq L_{u|s}, \forall s : \sum_u t_{s,u} = 1 \right\} \tag{A61}$$

$$= \{t \in \mathbb{R}^{\mathcal{X}} : \Phi t + \phi \geq 0, \Psi t + \psi = 0\}, \tag{A62}$$

where  $\Phi \in \mathbb{R}^{\mathcal{X} \times \mathcal{X}}$ ,  $\phi \in \mathbb{R}^{\mathcal{X}}$ ,  $\Psi \in \mathbb{R}^{\mathcal{S} \times \mathcal{X}}$  and  $\psi \in \mathbb{R}^{\mathcal{S}}$  are given, for  $s, s' \in \mathcal{S}$  and  $u \in \mathcal{U}$ , by

$$\Phi = \text{id}_{\mathcal{X}}, \tag{A63}$$

$$\phi_{s,u} = -L_{u|s}, \tag{A64}$$

$$\Psi_{s';(s,u)} = \begin{cases} 1, & \text{if } s = s', \\ 0, & \text{otherwise,} \end{cases} \tag{A65}$$

$$\psi_s = -1. \tag{A66}$$

Combining this with (A60), we find that  $\mathcal{Q}$  satisfies  $(\varepsilon, \mathcal{F})$ -RLDP whenever

$$\forall y, s_1, s_2 : \max_{\substack{t \in \mathbb{R}^{\mathcal{X}}: \\ \Phi t + \phi \geq 0, \\ \Psi t + \psi = 0}} ((B^{s_1, s_2})^T Q_y)^T t \leq 0. \tag{A67}$$

Now fix  $y, s_1, s_2$ , and consider the linear programming problem that forms the LHS of (A67). From the duality of linear programming, we know

$$\max_{\substack{t \in \mathbb{R}^{\mathcal{X}}: \\ \Phi t + \phi \geq 0, \\ \Psi t + \psi = 0}} ((B^{s_1, s_2})^T Q_y)^T t = \min_{\substack{z \in \mathbb{R}^{\mathcal{X}}, w \in \mathbb{R}^{\mathcal{S}}: \\ \Phi^T z + \Psi^T w = -(B^{s_1, s_2})^T Q_y, \\ z \geq 0}} \phi^T z + \psi^T w. \tag{A68}$$

We focus on the linear programming problem of the RHS. The terms of this problem are given by

$$\Phi^T z = z, \tag{A69}$$

$$(\Psi^T w)_{s,u} = w_s, \tag{A70}$$

$$((B^{s_1, s_2})^T Q_y)_{s,u} = \begin{cases} Q_{y|s_1, u}, & \text{if } s = s_1, \\ -e^\varepsilon Q_{y|s_2, u}, & \text{if } s = s_2, \\ 0, & \text{otherwise,} \end{cases} \tag{A71}$$

$$\phi^T z = - \sum_{s,u} L_{u|s} z_{s,u}, \tag{A72}$$

$$\psi^T w = - \sum_s w_s. \tag{A73}$$

The equation  $\Phi^T z + \Psi^T w = -(B^{s_1, s_2})^T Q_y$  can now be rewritten as

$$z_{s,u} = \begin{cases} -Q_{y|s_1, u} - w_{s_1}, & \text{if } s = s_1, \\ e^\varepsilon Q_{y|s_2, u} - w_{s_2}, & \text{if } s = s_2, \\ -w_s, & \text{otherwise.} \end{cases} \tag{A74}$$



Thus, the restriction  $z \geq 0$  translates to

$$\begin{aligned} w_{s_1} &\leq -\max_{u \in \mathcal{U}} Q_{y|s_1,u}, \\ w_{s_2} &\leq e^\epsilon \min_{u \in \mathcal{U}} Q_{y|s_2,u}, \\ \forall s \neq s_1, s_2: w_s &\leq 0. \end{aligned}$$

Furthermore, the objective function  $\phi^T z + \psi^T w$  becomes

$$-\sum_s \left(1 - \sum_u L_{u|s}\right) w_s + \sum_u Q_{y|s_1,u} L_{u|s_1} - e^\epsilon \sum_u Q_{y|s_2,u} L_{u|s_2}. \tag{A75}$$

Combining this with (A67) and (A68), we see that a sufficient condition for  $Q$  to be  $(\epsilon, \mathcal{F})$ -RLDP is if there exists a  $w \in \mathbb{R}^S$  such that

$$-\sum_s \left(1 - \sum_u L_{u|s}\right) w_s + \sum_u Q_{y|s_1,u} L_{u|s_1} - e^\epsilon \sum_u Q_{y|s_2,u} L_{u|s_2} \leq 0, \tag{A76}$$

$$w_{s_1} \leq -\max_{u \in \mathcal{U}} Q_{y|s_1,u}, \tag{A77}$$

$$w_{s_2} \leq e^\epsilon \min_{u \in \mathcal{U}} Q_{y|s_2,u}, \tag{A78}$$

$$\forall s \neq s_1, s_2 : w_s \leq 0. \tag{A79}$$

Since  $\sum_u L_{u|s} \leq 1$  for all  $s$ , it follows that the left-hand side of (A76) is minimal if each  $w_s$  attains its maximal value, subject to the constraints (A77)–(A79). Substituting this, we find that the minimum of the left-hand side is equal to

$$\begin{aligned} &\left(1 - \sum_u L_{u|s_1}\right) \left(\max_{u_1} Q_{y|u_1,s_1}\right) - e^\epsilon \left(1 - \sum_u L_{u|s_2}\right) \left(\min_{u_2} Q_{y|u_2,s_2}\right) \\ &+ \sum_u Q_{y|s_1,u} L_{u|s_1} - e^\epsilon \sum_u Q_{y|s_2,u} L_{u|s_2} \end{aligned} \tag{A80}$$

$$= \max_{u_1, u_2 \in \mathcal{U}} \left[ \left(1 - \sum_u L_{u|s_1}\right) Q_{y|u_1,s_1} - e^\epsilon \left(1 - \sum_u L_{u|s_2}\right) Q_{y|u_2,s_2} \right] + \sum_u Q_{y|s_1,u} L_{u|s_1} - e^\epsilon \sum_u Q_{y|s_2,u} L_{u|s_2} \tag{A81}$$

$$= \max_{u_1, u_2 \in \mathcal{U}} \left[ Q_{y|u_1,s_1} - e^\epsilon Q_{y|u_2,s_2} + \sum_u L_{u|s_1} (Q_{y|s_1,u} - Q_{y|s_1,u_1}) - e^\epsilon \sum_u L_{u|s_2} (Q_{y|s_2,u} - Q_{y|s_2,u_2}) \right]. \tag{A82}$$

This has to be nonpositive for all choices of  $u_1, u_2, s_1, s_2, y$ ; but this is true precisely if  $Q_y \in \Gamma_{L,\epsilon}$  for all  $y$ .

### Appendix A.5. Proof of Theorem 3

This is essentially analogous to the proof of Theorem 4 in [5]; the main difference is that the equivalent of  $\hat{\Gamma}$  is a hypercube, for which a vertex enumeration step is not needed. Let  $Q$  be a mechanism such that  $Q_y \in \Gamma$  for all  $y$ ; then there exist  $\alpha_y \in \mathbb{R}_{\geq 0}, \gamma_y \in \hat{\Gamma}$  such that  $Q_y = \alpha_y \gamma_y$ . One has

$$I_{\hat{\mathcal{P}}}(X; Y) = \sum_y \mu(Q_y) = \sum_y \alpha_y \mu(\gamma_y). \tag{A83}$$

Since  $\hat{\Gamma}$  is the convex hull of  $\mathcal{V}$ , we can write  $\gamma_y = \sum_v \lambda_{y,v} v$  for suitable constants  $\lambda_{y,v}$ . Define  $\theta \in \mathbb{R}_{\geq 0}^{\mathcal{V}}$  by  $\theta_v = \sum_y \lambda_{y,v} \alpha_y$ . Then,

$$\sum_v \theta_v v = \sum_y Q_y = 1_{\mathcal{X}}. \tag{A84}$$

As such, the matrix  $Q' \in \mathbb{R}^{\mathcal{V} \times \mathcal{X}}$  defined by  $Q'_v = \theta_v v$  defines a privacy mechanism  $Q'$ . One has

$$I_p(X; Q'(X)) = \sum_v \mu(Q'_v) \tag{A85}$$

$$= \sum_v \theta_v \mu(v) \tag{A86}$$

$$= \sum_y \alpha_y \sum_v \lambda_{y,v} \mu(v) \tag{A87}$$

$$\geq \sum_y \alpha_y \mu\left(\sum_v \lambda_{y,v} v\right) \tag{A88}$$

$$= I_p(X; Q(X)), \tag{A89}$$

where we use the fact that  $\mu$  is convex. This shows that the  $Q_y$  of the optimal mechanism satisfying Theorem 2 are all of the form  $\theta_v \cdot v$ ; hence, (46) yields the optimal mechanism. To see that  $|\mathcal{Y}| \leq a$ , observe that the polyhedron described in (46) is defined by  $a$  equality constraints, and  $|\mathcal{V}|$  inequality constraints of the form  $\theta_v \geq 0$ . Hence, any vertex of this polyhedron has at most  $a$  nonzero coefficients. Since the optimal mechanism corresponds to such a vertex, and its output space  $\mathcal{Y}$  corresponds to its nonzero coefficients, we conclude that  $|\mathcal{Y}| \leq a$ .  $\square$

Appendix A.6. Proof of Theorem 4

We follow the proof of Theorem 14 in [5]; however, we first need the following auxiliary lemma.

**Lemma A8.** Let  $\epsilon > 0$ , and let  $\mathcal{C} \subset \mathbb{R}_{\geq 0}^{\mathcal{X}}$  be the positive cone defined by

$$\mathcal{C} = \{C \in \mathbb{R}_{\geq 0}^{\mathcal{X}} : C_{s,u} \leq e^\epsilon C_{s',u'} \text{ for all } s \neq s' \in \mathcal{S}, u \in \mathcal{U}\}. \tag{A90}$$

Define the sets  $\mathcal{V}_1, \mathcal{V}_2, \mathcal{V} \subset \mathbb{R}_{\geq 0}^{\mathcal{X}}$  by

$$\mathcal{V}_1 = \left\{v \in \mathbb{R}_{\geq 0}^{\mathcal{X}} : \exists s \text{ s.t. } \begin{matrix} \forall u: v_{s,u} \in \{e^{-\epsilon}, e^\epsilon\}; \\ \forall s' \neq s, \forall u: v_{s',u} = 1 \end{matrix} \right\}, \tag{A91}$$

$$\mathcal{V}_2 = \left\{v \in \mathbb{R}_{\geq 0}^{\mathcal{X}} : \begin{matrix} \forall x: v_x \in \{1, e^\epsilon\}, \\ |\{s: \exists u \text{ s.t. } v_{s,u} = e^\epsilon\}| \geq 2 \end{matrix} \right\}, \tag{A92}$$

$$\mathcal{V} = \mathcal{V}_1 \cup \mathcal{V}_2. \tag{A93}$$

Then  $\mathcal{V}$  spans  $\mathcal{C}$  as a positive cone, i.e.,

$$\mathcal{C} = \left\{ \sum_{v \in \mathcal{V}} \theta_v v : \theta \in \mathbb{R}_{\geq 0}^{\mathcal{V}} \right\}. \tag{A94}$$

**Proof.** For every  $s \in \mathcal{S}$  and  $u, u' \in \mathcal{U}$ , we have

$$C_{s,u} \leq e^\epsilon C_{s',u} \leq e^{2\epsilon} C_{s,u'}, \tag{A95}$$

where  $s' \in \mathcal{S} \setminus \{s\}$  is arbitrary. Thus, in every  $C \in \mathcal{C}$  two coefficients can differ by at most a factor  $e^\epsilon$  if they have different  $s$ , and at most a factor  $e^{2\epsilon}$  if they have the same  $s$ . On the extremal rays of  $\mathcal{C}$ , the inequalities become equalities. By rescaling by a positive scalar, if necessary, we see that  $\mathcal{C}$  is spanned by vectors of which each coefficient is in the set  $\{e^{-\epsilon}, 1, e^\epsilon\}$ . In other words, if  $\mathcal{V}' = \{e^{-\epsilon}, 1, e^\epsilon\}^{\mathcal{X}} \cap \mathcal{C}$ , then

$$\mathcal{C} = \text{Span}(\mathcal{V}'), \tag{A96}$$

where Span refers to the span as in (A94). To determine  $\mathcal{V}'$  we consider two situations: either  $v$  contains both  $e^{-\varepsilon}$  and  $e^\varepsilon$  as coefficients, or not.

Suppose  $v$  contains  $e^{-\varepsilon}$  and  $e^\varepsilon$ , say  $v_{s,u} = e^{-\varepsilon}$  and  $v_{s',u'} = e^\varepsilon$ . By (A90), we must have  $s = s'$ , and by (A95), this means that  $v_{s'',u''} = 1$  for  $s'' \neq s$  and any  $u''$ . Thus, we define, for any  $s \in \mathcal{S}$ , the set

$$\mathcal{V}'_s = \left\{ v \in \mathbb{R}_{\geq}^{\mathcal{X}} : \forall s' \neq s \forall u : v_{s',u} = 1 \right\}. \tag{A97}$$

It is straightforward to show that  $\mathcal{V}'_s \subset \mathcal{C}$ , and by the discussion above any  $v \in \mathcal{V}'$  containing both  $e^{-\varepsilon}$  and  $e^\varepsilon$  is in  $\bigcup_s \mathcal{V}'_s$ .

Suppose  $v$  does not contain both  $e^{-\varepsilon}$  and  $e^\varepsilon$ , then  $v \in \mathcal{V}'_2 \cup \mathcal{V}'_3$  where

$$\mathcal{V}'_2 = \{1, e^\varepsilon\}^{\mathcal{X}}, \tag{A98}$$

$$\mathcal{V}'_3 = \{e^{-\varepsilon}, 1\}^{\mathcal{X}}. \tag{A99}$$

Furthermore, it is easy to see that  $\mathcal{V}'_2 \cup \mathcal{V}'_3 \subset \mathcal{V}'$ . Thus, we conclude that

$$\mathcal{V}' = \left( \bigcup_{s \in \mathcal{S}} \mathcal{V}'_s \right) \cup \mathcal{V}'_2 \cup \mathcal{V}'_3. \tag{A100}$$

To obtain from  $\mathcal{V}'$  to  $\mathcal{V}$ , we throw out some vectors that are not needed to span  $\mathcal{C}$ . We start with  $\mathcal{V}'_s$ . Given  $s$ , define the set

$$\mathcal{V}_s = \left\{ v \in \mathbb{R}_{\geq 0}^{\mathcal{X}} : \begin{array}{l} \forall u : v_{s,u} \in \{e^{-\varepsilon}, e^\varepsilon\}; \\ \forall s' \neq s, \forall u : v_{s',u} = 1 \end{array} \right\}. \tag{A101}$$

It is clear that  $\mathcal{V}_s \subset \mathcal{V}'_s$ ; we claim that

$$\text{Span}(\mathcal{V}_s) = \text{Span}(\mathcal{V}'_s). \tag{A102}$$

To see this, let  $v \in \mathcal{V}'_s \setminus \mathcal{V}_s$ , and define  $v^-, v^+ \in \mathbb{R}_{\geq 0}^{\mathcal{X}}$  by

$$v_{s',u}^+ = \begin{cases} e^{-\varepsilon}, & \text{if } s' = s \text{ and } v_{s',u} = e^{-\varepsilon}, \\ 1, & \text{if } s' \neq s, \\ e^\varepsilon, & \text{if } s' = s \text{ and } v_{s',u} \in \{1, e^\varepsilon\}, \end{cases} \tag{A103}$$

$$v_{s',u}^- = \begin{cases} e^{-\varepsilon}, & \text{if } s' = s \text{ and } v_{s',u} \in \{e^{-\varepsilon}, 1\}, \\ 1, & \text{if } s' \neq s, \\ e^\varepsilon, & \text{if } s' = s \text{ and } v_{s',u} = e^\varepsilon. \end{cases} \tag{A104}$$

In other words,  $v^\pm$  takes all  $s$ -coefficients of  $v$  that are equal to 1 and changes them to  $e^{\pm\varepsilon}$ . Then,  $v^+, v^- \in \mathcal{V}'_s$  and

$$v = \frac{1}{e^\varepsilon + 1} v^+ + \frac{e^\varepsilon}{e^\varepsilon + 1} v^-. \tag{A105}$$

Thus,  $v \in \text{Span}(\mathcal{V}'_s)$ , proving (A102). We now consider  $\mathcal{V}'_2$  and  $\mathcal{V}'_3$ . First note that  $\mathcal{V}'_3 = e^{-\varepsilon} \mathcal{V}'_2$ , so

$$\text{Span}(\mathcal{V}'_2) = \text{Span}(\mathcal{V}'_3). \tag{A106}$$

We furthermore claim that

$$\text{Span} \left( \mathcal{V}_2 \cup \bigcup_{s \in \mathcal{S}} \mathcal{V}_s \right) = \text{Span} \left( \mathcal{V}'_2 \cup \bigcup_{s \in \mathcal{S}} \mathcal{V}'_s \right), \tag{A107}$$

where  $\mathcal{V}_2$  is as in (A92). Note that clearly  $\mathcal{V}_2 \subset \mathcal{V}'_2$ . To see (A107), let  $v \in \mathcal{V}'_2 \setminus \mathcal{V}_2$ ; this means that there is at most a single  $(s, u)$  such that  $v_{s,u} = e^\varepsilon$ . If no such  $(s, u)$  exists, then  $v = 1_{\mathcal{X}}$ ,

the constant vector with all ones. This implies that  $e^\epsilon v \in \mathcal{V}_2$ , showing that  $v \in \text{Span}(\mathcal{V}_2)$ . Now suppose that there is exactly one  $(s, u)$  such that  $v_{s,u} = e^\epsilon$ . Then,

$$v_{s',u'} = \begin{cases} e^\epsilon, & \text{if } s = s' \text{ and } u = u', \\ 1, & \text{otherwise.} \end{cases} \tag{A108}$$

But then we can construct  $v^+$  as in (A103) and  $v^-$  as in (A104), and again we find

$$v = \frac{1}{e^\epsilon + 1} v^+ + \frac{e^\epsilon}{e^\epsilon + 1} v^- \in \text{Span}(\mathcal{V}_s). \tag{A109}$$

This proves (A107). Combining (A102), (A106) and (A107) we obtain

$$\mathcal{C} = \text{Span}\left(\mathcal{V}'_2 \cup \mathcal{V}'_3 \cup \bigcup_{s \in \mathcal{S}} \mathcal{V}'_s\right) \tag{A110}$$

$$= \text{Span}\left(\mathcal{V}_2 \cup \bigcup_{s \in \mathcal{S}} \mathcal{V}_s\right) \tag{A111}$$

$$= \text{Span}(\mathcal{V}_2 \cup \mathcal{V}_1). \tag{A112}$$

□

**Proof of Theorem 4.** We follow the proof of Theorem 14 in [5]. For  $C \in \mathbb{R}_{\geq 0}^{\mathcal{X}}$ , define

$$\mu(C) = \sum_x P_x C_x \log \frac{C_x}{\sum_{x'} P_{x'} C_{x'}}. \tag{A113}$$

For  $y \in \mathcal{Y}$ , let  $Q_y = (Q_{y|x})_x \in \mathbb{R}^{\mathcal{X}}$ ; then the utility of a mechanism  $\mathcal{Q}: \mathcal{X} \rightarrow \mathcal{Y}$  is given by  $I_P(X; Y) = \sum_y \mu(Q_y)$ . Furthermore,  $\mu$  is a sublinear function in the sense of Definition 1 of [5].

We fix an  $\epsilon > 0$ . Furthermore, let  $\mathcal{C} \subset \mathbb{R}_{\geq 0}^{\mathcal{X}}$  be as in Lemma A8. Then, a mechanism  $\mathcal{Q}$  satisfies  $(\epsilon, \mathcal{P}_{\mathcal{X}})$ -RLDP if and only if each  $Q_y$  is an element of  $\mathcal{C}$ . Let  $\mathcal{V}$  be the spanning set of  $\mathcal{V}$  of Lemma A8, and let  $\mathcal{D}$  be the polytope spanned by  $\mathcal{V}$ . If  $\mathcal{Q}$  satisfies  $\epsilon$ -SLDP, then every column  $Q_y$  is of the form  $\theta_y \cdot d_y$ , where  $d_y \in \mathcal{D}$  and  $\theta_y \in \mathbb{R}_{\geq 0}$  are such that  $\sum_y \theta_y d_y = 1_{\mathcal{X}}$ . Analogously to the proof of Theorems 2 and 4 in Section 7 of [5] (or, for that matter, our proof of Theorem 3), one proves that the optimal  $\mathcal{Q}$  is found by taking  $b = a$ , and taking  $d_y \in \mathcal{V}$  for all  $d$ . Since

$$I(X; Y) = \sum_y \mu(Q_y) = \sum_y \theta_y \mu(d_y) \tag{A114}$$

we can find the optimal  $\mathcal{Q}$  by solving the following optimization problem, where  $m \in \mathbb{R}^{\mathcal{V}}$  is the vector  $(\mu(v))_{v \in \mathcal{V}}$ , and where  $A \in \mathbb{R}^{\mathcal{X} \times \mathcal{V}}$  is the matrix whose  $v$ -th column is  $v$ :

$$\begin{aligned} & \text{maximize}_{\theta \in \mathbb{R}^{\mathcal{V}}} m \cdot \theta \\ & \text{such that } A \cdot \theta = 1_{\mathcal{X}}, \\ & \theta \geq 0. \end{aligned}$$

From here, we follow Section 9.5 of [5]. The dual to the above problem is

$$\begin{aligned} & \text{minimize}_{\alpha \in \mathbb{R}^{\mathcal{X}}} (1_{\mathcal{X}}) \cdot \alpha \\ & \text{such that } A^T \cdot \alpha \geq m, \\ & \alpha \geq 0. \end{aligned}$$

By duality, we have  $\max_{\theta} m \cdot \theta = \min_{\alpha} (1_{\mathcal{X}}) \cdot \alpha$ . We describe  $\alpha^* \geq 0$  and  $\theta^* \geq 0$ , depending on  $\epsilon$ , such that for sufficiently large  $\epsilon$  one has  $A^T \cdot \alpha^* \geq m$ , such that  $m \cdot \theta^* = (1_{\mathcal{X}}) \cdot \alpha^*$  and  $A\theta^* = 1_{\mathcal{X}}$ , and such that  $\theta^*$  corresponds to SRR, i.e., for each  $y \in \mathcal{Y} = \mathcal{X}$

there is a  $\hat{v}_y \in \mathcal{V}$  such that  $\text{SRR}_y^\varepsilon = \theta_{\hat{v}_y}^* \hat{v}_y$ . Together, this proves that SRR is optimal for  $\varepsilon \gg 0$ .

More concretely, for  $y = (s, u) \in \mathcal{X}$ , define  $\hat{v}_y$  by

$$(\hat{v}_y)_{s',u'} = \begin{cases} e^\varepsilon, & \text{if } (s', u') = (s, u), \\ e^{-\varepsilon}, & \text{if } s' = s \text{ and } u' \neq u, \\ 1, & \text{if } s' \neq s. \end{cases} \tag{A115}$$

Note that  $\hat{v}_y \in \mathcal{V}$ . Furthermore, let  $\theta^* \in \mathbb{R}^{\mathcal{V}}$  be given by

$$\theta_v^* = \begin{cases} \frac{1}{e^\varepsilon + e^{-\varepsilon}(a_2 - 1) + a - a_2}, & \text{if there is a } y \in \mathcal{X} \text{ such that } v = \hat{v}_y, \\ 0, & \text{otherwise;} \end{cases} \tag{A116}$$

Then, SRR satisfies  $\text{SRR}_y^\varepsilon = \theta_{\hat{v}_y}^* \hat{v}_y$  for all  $y \in \mathcal{X}$ , and for each  $x \in \mathcal{X}$  one has

$$(A\theta^*)_x = \sum_v A_{x,v} \theta_v^* \tag{A117}$$

$$= \sum_v v_x \theta_v^* \tag{A118}$$

$$= \frac{\sum_y (\hat{v}_y)_x}{e^\varepsilon + e^{-\varepsilon}(a_2 - 1) + a - a_2} \tag{A119}$$

$$= 1, \tag{A120}$$

which shows that  $A\theta^* = 1_{\mathcal{X}}$ . Furthermore, define  $\alpha^* \in \mathbb{R}^{\mathcal{X}}$  by

$$\alpha_{s,u}^* = c_1 \mu(\hat{v}_{s,u}) + c_2 \sum_{u' \neq u} \mu(\hat{v}_{s,u'}) + c_3 \sum_{\substack{s' \neq s, \\ u'}} \mu(\hat{v}_{s',u'}), \tag{A121}$$

where

$$c_1 = \frac{-(a_2 - 2)(a_2 - 1) + (a - a_2 + 1)(a_2 - 2)e^\varepsilon + (a - 2a_2 + 1)e^{2\varepsilon} + e^{3\varepsilon}}{(e^\varepsilon - 1)(e^\varepsilon + 1)(e^\varepsilon - a_2 + 1)(e^\varepsilon + (a_2 - 1)e^{-\varepsilon} + a - a_2)}, \tag{A122}$$

$$c_2 = \frac{a_2 - 1 + (a - a_2 + 1)e^\varepsilon}{(e^\varepsilon - 1)(e^\varepsilon + 1)(e^\varepsilon - a_2 + 1)(e^\varepsilon + (a_2 - 1)e^{-\varepsilon} + a - a_2)}, \tag{A123}$$

$$c_3 = \frac{-e^{2\varepsilon}}{(e^\varepsilon - 1)(e^\varepsilon - a_2 + 1)(e^\varepsilon + (a_2 - 1)e^{-\varepsilon} + a - a_2)}. \tag{A124}$$

A cumbersome but straightforward calculation shows that for all  $x$ , we have

$$m \cdot \theta^* = (1_{\mathcal{X}}) \cdot \alpha^* = \frac{1}{e^\varepsilon + e^{-\varepsilon}(a_2 - 1) + a - a_2} \sum_x \mu(\hat{v}_x), \tag{A125}$$

$$\hat{v}_x \cdot \alpha^* = m_{\hat{v}_x} = \mu(\hat{v}_x). \tag{A126}$$

Furthermore,  $c_1, c_2, c_3 \geq 0$ , so  $\alpha^* \geq 0$ . It remains to be shown that  $\alpha^*$  satisfies the dual problem for  $\varepsilon \gg 0$ , i.e.,  $A^T \alpha \geq m$  for sufficiently large  $\varepsilon$ . To this end, for  $v \in \mathcal{V}$ , set

$$F_v = \{x \in \mathcal{X} : v_x = e^\varepsilon\}, \tag{A127}$$

$$G_v = \{x \in \mathcal{X} : v_x = 1\}, \tag{A128}$$

$$H_v = \{x \in \mathcal{X} : v_x = e^{-\varepsilon}\}, \tag{A129}$$

From the description of  $\mathcal{V}$  in Lemma A8, we find that  $|F_v| \geq 1$  for all  $v$ , and  $|F_v| = 1$  if and only if there exist  $s, u$  such that  $v = \hat{v}_{s,u}$ . Now, write  $P_{F_v} = \sum_{x \in F_v} P_x$  and likewise for  $G_v, H_v$ . For large  $\varepsilon$ , we have

$$m_v = \mu(v) = e^\varepsilon \sum_{x \in F_v} P_x \log \frac{1}{P_{F_v} + e^{-\varepsilon} P_{G_v} + e^{-2\varepsilon} P_{H_v}} \quad (\text{A130})$$

$$+ \sum_{x \in G_v} P_x \log \frac{1}{e^\varepsilon P_{F_v} + P_{G_v} + e^{-\varepsilon} P_{H_v}} \quad (\text{A131})$$

$$+ e^{-\varepsilon} \sum_{x \in H_v} P_x \log \frac{1}{e^{2\varepsilon} P_{F_v} + e^\varepsilon P_{G_v} + P_{H_v}} \quad (\text{A132})$$

$$= (-P_{F_v} \log P_{F_v}) e^\varepsilon + \mathcal{O}(\varepsilon) \quad (\text{A133})$$

and furthermore

$$c_1 = e^{-\varepsilon} + \mathcal{O}(e^{-2\varepsilon}), \quad (\text{A134})$$

$$c_2, c_3 = \mathcal{O}(e^{-2\varepsilon}). \quad (\text{A135})$$

From this, it follows that

$$\alpha_x^* = c_1 \mu(\hat{v}_x) + (c_2 + c_3) \mathcal{O}(e^\varepsilon) \quad (\text{A136})$$

$$= -P_x \log P_x + \mathcal{O}(\varepsilon e^{-\varepsilon}), \quad (\text{A137})$$

Hence,

$$v^T \alpha^* = \left( - \sum_{x \in F_v} P_x \log P_x \right) e^\varepsilon + \mathcal{O}(\varepsilon). \quad (\text{A138})$$

For  $|F_v| \geq 2$ , one has  $P_{F_v} \log P_{F_v} > \sum_{x \in F_v} P_x \log P_x$ . This means that if  $v$  is not of the form  $\hat{v}_x$ , one has  $v^T \alpha^* \geq m_v$  for sufficiently large  $\varepsilon$ . Together with (A126), this shows that  $A^T \alpha^* \geq m$  for sufficiently large  $\varepsilon$ ; this concludes the proof.  $\square$

## References

1. Kasiviswanathan, S.P.; Lee, H.K.; Nissim, K.; Raskhodnikova, S.; Smith, A. What can we learn privately? *SIAM J. Comput.* **2011**, *40*, 793–826. [\[CrossRef\]](#)
2. Duchi, J.C.; Jordan, M.I.; Wainwright, M.J. Local privacy and statistical minimax rates. In Proceedings of the 2013 IEEE 54th Annual Symposium on Foundations of Computer Science (FOCS), Berkeley, CA, USA, 26–29 October 2013; pp. 429–438.
3. Lopuhaä-Zwakenberg, M.; Tong, H.; Škorić, B. Data Sanitisation for the Privacy Funnel with Differential Privacy Guarantees. *Int. J. Adv. Secur.* **2020**, *13*, 162–174.
4. Rebollo-Monedero, D.; Forne, J.; Domingo-Ferrer, J. From t-closeness-like privacy to postrandomization via information theory. *IEEE Trans. Knowl. Data Eng.* **2010**, *22*, 1623–1636. [\[CrossRef\]](#)
5. Kairouz, P.; Oh, S.; Viswanath, P. Extremal mechanisms for local differential privacy. *J. Mach. Learn. Res.* **2016**, *17*, 492–542.
6. Makhdoumi, A.; Salamatian, S.; Fawaz, N.; Médard, M. From the information bottleneck to the privacy funnel. In Proceedings of the 2014 IEEE Information Theory Workshop (ITW 2014), Hobart, TAS, Australia, 2–5 November 2014; pp. 501–505.
7. Salamatian, S.; Zhang, A.; du Pin Calmon, F.; Bhamidipati, S.; Fawaz, N.; Kveton, B.; Oliveira, P.; Taft, N. Managing your private and public data: Bringing down inference attacks against your privacy. *IEEE J. Sel. Top. Signal Process.* **2015**, *9*, 1240–1255. [\[CrossRef\]](#)
8. Asoodeh, S.; Diaz, M.; Alajaji, F.; Linder, T. Information extraction under privacy constraints. *Information* **2016**, *7*, 15. [\[CrossRef\]](#)
9. Kung, S. A compressive privacy approach to generalized information bottleneck and privacy funnel problems. *J. Frankl. Inst.* **2018**, *355*, 1846–1872. [\[CrossRef\]](#)
10. Ding, N.; Sadeghi, P. A submodularity-based clustering algorithm for the information bottleneck and privacy funnel. In Proceedings of the 2019 IEEE Information Theory Workshop (ITW), Visby, Sweden, 25–28 August 2019; pp. 1–5.
11. Salamatian, S.; Calmon, F.P.; Fawaz, N.; Makhdoumi, A.; Médard, M. Privacy-Utility Tradeoff and Privacy Funnel. 2020. Available online: <https://api.semanticscholar.org/CorpusID:210927663> (accessed on 10 January 2024).
12. Acharya, J.; Bonawitz, K.; Kairouz, P.; Ramage, D.; Sun, Z. Context aware local differential privacy. In Proceedings of the International Conference on Machine Learning, PMLR, Virtual, 13–18 July 2020; pp. 52–62.
13. Goseling, J.; Lopuhaä-Zwakenberg, M. Robust optimization for local differential privacy. In Proceedings of the 2022 IEEE International Symposium on Information Theory (ISIT), Espoo, Finland, 26 June–1 July 2022; pp. 1629–1634.

14. Lopuhaä-Zwakenberg, M.; Goseling, J. Robust Local Differential Privacy. In Proceedings of the 2021 IEEE International Symposium on Information Theory (ISIT), Melbourne, VIC, Australia, 12–20 July 2021; pp. 557–562.
15. Kifer, D.; Machanavajjhala, A. Pufferfish: A framework for mathematical privacy definitions. *ACM Trans. Database Syst.* **2014**, *39*, 1–36. [[CrossRef](#)]
16. Ben-Tal, A.; El Ghaoui, L.; Nemirovski, A. *Robust Optimization*; Princeton University Press: Princeton, NJ, USA, 2009; Volume 28.
17. Ben-Tal, A.; Den Hertog, D.; Vial, J.P. Deriving robust counterparts of nonlinear uncertain inequalities. *Math. Program.* **2015**, *149*, 265–299. [[CrossRef](#)]
18. Bertsimas, D.; Gupta, V.; Kallus, N. Data-driven robust optimization. *Math. Program.* **2018**, *167*, 235–292. [[CrossRef](#)]
19. Warner, S.L. Randomized response: A survey technique for eliminating evasive answer bias. *J. Am. Stat. Assoc.* **1965**, *60*, 63–69. [[CrossRef](#)] [[PubMed](#)]
20. Song, S.; Wang, Y.; Chaudhuri, K. Pufferfish privacy mechanisms for correlated data. In Proceedings of the 2017 ACM International Conference on Management of Data, Chicago, IL, USA, 14–19 May 2017; pp. 1291–1306.
21. Nuradha, T.; Goldfeld, Z. Pufferfish Privacy: An Information-Theoretic Study. *IEEE Trans. Inf. Theory* **2023**, *69*, 7336–7356. [[CrossRef](#)]
22. Yang, B.; Sato, I.; Nakagawa, H. Bayesian differential privacy on correlated data. In Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data, Melbourne, VIC, Australia, 31 May–4 June 2015; pp. 747–762.
23. He, X.; Machanavajjhala, A.; Ding, B. Blowfish privacy: Tuning privacy-utility trade-offs using policies. In Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, Snowbird, UT, USA, 22–27 June 2014; pp. 1447–1458.
24. Dwork, C.; Roth, A. The algorithmic foundations of differential privacy. *Found. Trends<sup>®</sup> Theor. Comput. Sci.* **2014**, *9*, 211–407. [[CrossRef](#)]
25. Wang, T.; Blocki, J.; Li, N.; Jha, S. Locally differentially private protocols for frequency estimation. In Proceedings of the 26th {USENIX} Security Symposium ({USENIX} Security 17), Vancouver, BC, Canada, 16–18 August 2017; pp. 729–745.
26. Tishby, N.; Pereira, F.C.; Bialek, W. The information bottleneck method. *arXiv* **2000**, arXiv:physics/0004057.
27. Wagner, I.; Eckhoff, D. Technical privacy metrics: A systematic survey. *ACM Comput. Surv.* **2018**, *51*, 1–38. [[CrossRef](#)]
28. Rassouli, B.; Gunduz, D. On perfect privacy. In Proceedings of the 2018 IEEE International Symposium on Information Theory (ISIT), Vail, CO, USA, 17–22 June 2018; pp. 2551–2555.
29. Asoodeh, S.; Diaz, M.; Alajaji, F.; Linder, T. Estimation efficiency under privacy constraints. *IEEE Trans. Inf. Theory* **2018**, *65*, 1512–1534. [[CrossRef](#)]
30. Wang, H.; Calmon, F.P. An estimation-theoretic view of privacy. In Proceedings of the 2017 55th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 3–6 October 2017; pp. 886–893.
31. Mironov, I. Rényi differential privacy. In Proceedings of the 2017 IEEE 30th Computer Security Foundations Symposium (CSF), Barbara, CA, USA, 21–25 August 2017; pp. 263–275.
32. Issa, I.; Wagner, A.B.; Kamath, S. An operational approach to information leakage. *IEEE Trans. Inf. Theory* **2019**, *66*, 1625–1657. [[CrossRef](#)]
33. Liao, J.; Kosut, O.; Sankar, L.; du Pin Calmon, F. Tunable Measures for Information Leakage and Applications to Privacy-Utility Tradeoffs. *IEEE Trans. Inf. Theory* **2019**, *65*, 8043–8066. [[CrossRef](#)]
34. Saeidian, S.; Cervia, G.; Oechtering, T.J.; Skoglund, M. Pointwise maximal leakage. *IEEE Trans. Inf. Theory* **2023**, *69*, 8054–8080. [[CrossRef](#)]
35. Diaz, M.; Wang, H.; Calmon, F.P.; Sankar, L. On the robustness of information-theoretic privacy measures and mechanisms. *IEEE Trans. Inf. Theory* **2019**, *66*, 1949–1978. [[CrossRef](#)]
36. Makhdoumi, A.; Fawaz, N. Privacy-utility tradeoff under statistical uncertainty. In Proceedings of the 2013 51st Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 2–4 October 2013; pp. 1627–1634.
37. Kalantari, K.; Sankar, L.; Sarwate, A.D. Robust privacy-utility tradeoffs under differential privacy and hamming distortion. *IEEE Trans. Inf. Forensics Secur.* **2018**, *13*, 2816–2830. [[CrossRef](#)]
38. Asoodeh, S.; Alajaji, F.; Linder, T. On maximal correlation, mutual information and data privacy. In Proceedings of the 2015 IEEE 14th Canadian Workshop on Information Theory (CWIT), St. John’s, NL, Canada, 6–9 July 2015; pp. 27–31.
39. Pardo, L. *Statistical Inference Based on Divergence Measures*; CRC Press: Boca Raton, FL, USA, 2018.
40. Ben-Tal, A.; Den Hertog, D.; De Waegenaere, A.; Melenberg, B.; Rennen, G. Robust solutions of optimization problems affected by uncertain probabilities. *Manag. Sci.* **2013**, *59*, 341–357. [[CrossRef](#)]
41. Duchi, J.C.; Glynn, P.W.; Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *Math. Oper. Res.* **2021**, *46*, 946–969. [[CrossRef](#)]
42. Wang, Z.; Glynn, P.W.; Ye, Y. Likelihood robust optimization for data-driven problems. *Comput. Manag. Sci.* **2016**, *13*, 241–261. [[CrossRef](#)]
43. Mohajerin Esfahani, P.; Kuhn, D. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Math. Program.* **2018**, *171*, 115–166. [[CrossRef](#)]
44. Selvi, A.; Liu, H.; Wiesemann, W. Differential Privacy via Distributionally Robust Optimization. *arXiv* **2023**, arXiv:2304.12681.
45. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [[CrossRef](#)]

46. Huang, C.; Kairouz, P.; Chen, X.; Sankar, L.; Rajagopal, R. Context-aware generative adversarial privacy. *Entropy* **2017**, *19*, 656. [[CrossRef](#)]
47. Tripathy, A.; Wang, Y.; Ishwar, P. Privacy-preserving adversarial networks. In Proceedings of the 2019 57th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 24–27 September 2019; pp. 495–505.
48. Mirjalili, V.; Raschka, S.; Namboodiri, A.; Ross, A. Semi-adversarial networks: Convolutional autoencoders for imparting privacy to face images. In Proceedings of the 2018 International Conference on Biometrics (ICB), Gold Coast, QLD, Australia, 20–23 February 2018; pp. 82–89.
49. Bortolato, B.; Ivanovska, M.; Rot, P.; Križaj, J.; Terhörst, P.; Damer, N.; Peer, P.; Štruc, V. Learning privacy-enhancing face representations through feature disentanglement. In Proceedings of the 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020) (FG), Buenos Aires, Argentina, 16–20 November 2020; pp. 45–52.
50. Stoker, J.I.; Garretsen, H.; Spreeuwiers, L.J. The facial appearance of CEOs: Faces signal selection but not performance. *PLoS ONE* **2016**, *11*, e0159950. [[CrossRef](#)]
51. Willenborg, L.; De Waal, T. *Elements of Statistical Disclosure Control*; Springer Science & Business Media: Berlin, Germany, 2012; Volume 155.
52. Hundepool, A.; Domingo-Ferrer, J.; Franconi, L.; Giessing, S.; Nordholt, E.S.; Spicer, K.; De Wolf, P.P. *Statistical Disclosure Control*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
53. Liese, F.; Vajda, I.  $f$ -divergences: Sufficiency, deficiency and testing of hypotheses. In *Advances in Inequalities from Probability Theory and Statistics*; Nova Publishers: New York, NY, USA, 2008; p. 113.
54. van Erven, T.; Harremoës, P. Rényi divergence and majorization. In Proceedings of the 2010 IEEE International Symposium on Information Theory, Austin, TX, USA, 13–18 June 2010; pp. 1335–1339.
55. Csiszár, I. Information-type measures of difference of probability distributions and indirect observation. *Stud. Sci. Math. Hung.* **1967**, *2*, 229–318.
56. Kullback, S. A lower bound for discrimination information in terms of variation (corresp.). *IEEE Trans. Inf. Theory* **1967**, *13*, 126–127. [[CrossRef](#)]
57. Gilardoni, G.L. On Pinsker’s and Vajda’s type inequalities for Csiszár’s  $f$ -divergences. *IEEE Trans. Inf. Theory* **2010**, *56*, 5377–5386. [[CrossRef](#)]
58. Toth, C.D.; O’Rourke, J.; Goodman, J.E. *Handbook of Discrete and Computational Geometry*; CRC Press: Boca Raton, FL, USA, 2017.
59. Dwork, C.; McSherry, F.; Nissim, K.; Smith, A. Calibrating noise to sensitivity in private data analysis. In Proceedings of the Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, 4–7 March 2006; pp. 265–284.
60. Dua, D.; Graff, C. UCI Machine Learning Repository. 2017. Available online: <https://archive.ics.uci.edu/> (accessed on 10 January 2024).

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.