Journal of Machine Learning Research 22 (2021) 1-56 Submitted 10/18; Revised 4/21; Published 7/21

Locally Differentially-Private Randomized Response for Discrete Distribution Learning

Adriano Pastore ADRIANO.PASTORE@CTTC.CAT
Department of Statistical Inference for Communications and Positioning
Centre Tecnològic de Telecomunicacions de Catalunya (CTTC/CERCA)
Castelldefels, Barcelona, Spain

Michael Gastpar MICHAEL.GASTPAR@EPFL.CH
School for Computer and Communication Sciences
Ecole Polytechnique Fédérale de Lausanne
Lausanne, Switzerland

Editor: Stefan Wrobel

Abstract

We consider a setup in which confidential i.i.d. samples $X_1, \ldots, X_n$ from an unknown finite-support distribution $\boldsymbol{p}$ are passed through n copies of a discrete privatization channel (a.k.a. *mechanism*) producing outputs $Y_1, \ldots, Y_n$. The channel law guarantees a local differential privacy of ϵ. Subject to a prescribed privacy level ϵ, the optimal channel should be designed such that an estimate of the source distribution based on the channel outputs $Y_1, \ldots, Y_n$ converges as fast as possible to the exact value $\boldsymbol{p}$. For this purpose we study the convergence to zero of three distribution distance metrics: f-divergence, mean-squared error and total variation. We derive the respective normalized first-order terms of convergence (as $n \to \infty$), which for a given target privacy ϵ represent a rule-of-thumb factor by which the sample size must be augmented so as to achieve the same estimation accuracy as that of a non-randomizing channel. We formulate the privacy–fidelity trade-off problem as being that of minimizing said first-order term under a privacy constraint ϵ. We further identify a scalar quantity that captures the essence of this trade-off, and prove bounds and data-processing inequalities on this quantity. For some specific instances of the privacy–fidelity trade-off problem, we derive inner and outer bounds on the optimal trade-off curve.

Keywords: differential privacy, randomized response, distribution estimation, privacy–utility trade-off

1. Introduction

In the statistical analysis of privacy-sensitive data, the key challenge consists in randomizing database queries or post-processing (sanitizing) the data set so as to render inferences about the *data* (values or labels) as difficult as possible while at the same time preserving the usefulness of the data for estimating *parameters* of the underlying distribution. The inherent trade-off between the conflicting goals of privacy and utility arises in a broad variety of situations, notably in medical surveys, customer profiling, consumer studies, population census, opinion polls or surveys in social sciences.

Specifically, we will be concerned with the *randomized response* (RR) technique. An early inspiration for the basic setup, which is depicted in Figure 1 further below, dates back to Warner (1965): a common task in social sciences is to conduct surveys in which some answers might be stigmatizing (e.g., questions on drug use, sexual behavior, etc.). To overcome the respondent's potential reticence to answering faithfully, Warner proposed to perturb the interviewee's answers by having him/her secretly randomize the answers, in such way that not even the interviewer would learn the true answer.

In more recent years, a substantial body of work has developed around the celebrated notion of *differential privacy* (DP) introduced by Dwork (2006); Dwork et al. (2006) building on precursor work by Evfimievski et al. (2003), amongst others. The original purpose of DP was to provide strong privacy guarantees against a resourceful attacker who can access queries to a central database (see for example De (2012); Vadhan (2017)). By contrast, Kasiviswanathan et al. (2011) and Wainwright et al. (2012) considered a decentralized privatization model referred to as *local privacy*, in which each data point is independently perturbed by a randomizing mechanism. Warner's RR scheme can be inscribed in this local privacy framework.

Notable other proposals in the spirit of this local, non-interactive privatization mechanism include Agrawal and Srikant (2000), in which the authors propose a procedure to build a decision-tree classifier from perturbed training data that performs close to a classifier built from the original non-perturbed data. Wasserman and Zhou (2010) study the exponential rate at which certain estimates of continuous distributions from noisy samples concentrate in a small ball (in mean-square loss or Kolmogorov-Smirnov distance) centered around the true distribution. Wainwright et al. (2012) study minimax learning rates of distribution parameters from privatized samples from the viewpoint of statistics. In the context of locally private hypothesis testing, Gaboardi and Rogers (2017) study privatized chi-square tests for goodness of fit and independence testing. A more recent publication by Nageswaran and Narayan (2020) considers a similar setting as ours, but with reversed targets: the true distribution is to be kept secret from the querier, whereas the data (or some function thereof) is to be disclosed with best possible accuracy. Also worth mentioning as a new research direction is the work by Huang et al. (2017), who introduced the concept of *generative adversarial privacy* which is inspired by the recent invention of generative adversarial networks: the confidential data set (in a *non-local* privacy context) is used to train a generative neural network (against an adversary), which then creates plausible new data samples that emulate the original distribution without disclosing any sample from the training set.

For discrete and finitely supported distributions, the work of Kairouz et al. (2014) has shown that a finite class of privatization mechanisms called *staircase mechanisms* are optimal among all ϵ-private mechanisms for a constrained f-divergence maximization problem related to private hypothesis testing. In particular, they show that a simple mechanism, which we call the *step mechanism*—defined in Equation (23) in the present article—which they refer to simply as *randomized response* (RR), is optimal for their problem in the low privacy regime. Interestingly, this mechanism appears in other contexts as well: for instance in the distributed computation problem studied by Kairouz et al. (2015), this mechanism turns out to be optimal in a fairly general sense.

An alternative (and complementary) privatization scheme to RR has been more recently introduced under the name of Randomized Aggregatable Privacy-Preserving Ordinal Response (RAPPOR) scheme (Erlingsson et al., 2014; Duchi et al., 2013). While the conventional RR scheme assumes that the source variable and the privatized share a same finite alphabet, the privatized representation in a RAPPOR scheme lives on an alphabet whose size grows exponentially in the source alphabet size, thus harshly impacting on storage or bandwidth requirements. It was shown in Kairouz et al. (2016a,b) that under ℓ_1 (total variation distance) and ℓ_2 (mean-squared error) losses alike, the RAPPOR scheme is order-optimal in the high privacy regime ($\epsilon \downarrow 0$) and strictly suboptimal in the low privacy regime ($\epsilon \uparrow \infty$). Conversely, the RR scheme is order-optimal in the low privacy, and sub-optimal in the high-privacy regime.

Recently, Ye and Barg (2018) introduced a novel privatization scheme[1] which can be viewed as a generalization of both RR and RAPPOR, and is sometimes referred to as *subset selection*. They prove that under ℓ_1 and ℓ_2 metrics, these schemes are order-optimal in the medium to high privacy regime ($e^\epsilon \ll K$, where K denotes the cardinality of the discrete source's support). Subsequently in Ye and Barg (2017), the same authors strengthen this result for the ℓ_2 metric, by proving asymptotic optimality in all regimes. The result was recently extended to more general metrics (including the ℓ_1 metric) by the same authors (Ye and Barg, 2019). Subsequent work by Acharya et al. (2019) has instead focused on the distributed scenario, in which the Hadamard response is proposed as a more suitable mechanism than RR, RAPPOR or subset selection, in that it substantially drives down communication complexity while still being sample optimal.

In the present work, rather than studying RAPPOR and its generalization by Ye and Barg, we turn our attention to the conventional, more storage efficient RR scheme.[2] Specifically, we consider the situation where an *interviewer* or *curator* observes independent and identically distributed (i.i.d.) realizations of potentially sensitive data. Instead of publishing the records in the clear, the curator is required to process this source data in such way that inferences on the individual source *realizations* are rendered hard, but an accurate estimation of the source *distribution* is rendered easy. In addition, the curator is constrained to using a memoryless privatization strategy, sometimes referred to as *non-interactive mechanism* (Leoni, 2012; Duchi et al., 2013, 2014).

Although non-interactive mechanisms are provably optimal in certain setups (Kairouz et al., 2015), the trade-off under study may in general benefit from more complex, interactive channel structures. However, there might be justified reasons to use a stationary and memoryless privatization channel:

Randomized survey In certain situations such as randomized response surveys (Warner, 1965), the truthful answer to an interviewer's question might be stigmatizing (e.g., drug consumption, sexual behavior, etc.). In such setups, a transparent and non-interactive randomization of answers is necessary as a participatory incentive.

1. Ye and Barg acknowledge in their publication that Wang et al. (2016) independently introduced the same privatization scheme.
2. Note that several key results known in the literature that facilitate further analysis, such as optimality of "extremal" mechanisms (Kairouz et al., 2014; Ye and Barg, 2018, Lemma IV.3), cease to hold under the restriction of pure RR that we take here.

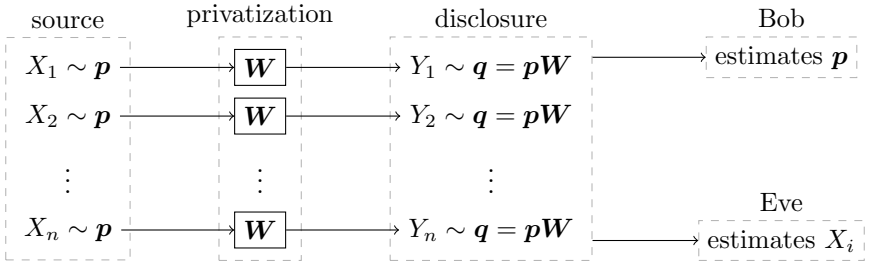


Figure 1: Non-interactive mechanism: the curator sees n samples of an i.i.d. discrete random source $\{X_i\}$ with distribution $\boldsymbol{p}$ and processes them individually through n copies of a privatization channel $\boldsymbol{W}$ prior to publication. From the outputs $\{Y_i\}$, the legitimate observer Bob tries to infer $\boldsymbol{p}$, whereas the adversarial observer Eve tries to infer one (or more) source sample X_i.

Timeliness Strict delay constraints may outrule the possibility of batch processing. Low-latency sensoring or time-critical surveys may be examples of such situations. For instance, a medical survey could be conducted over a timespan of several years, but there might be an urge to publish partial information at much more frequent intervals so as to help gain statistical insights in a timely manner.

Finite horizon In applications where the curator has no control over the eventual size n of the data collection because it may be interrupted anytime, a non-interactive mechanism seems a more viable and robust approach.

Privacy fairness Applying one and the same privatization channel onto each data sample enforces full uniformity of privacy guarantees across samples in a simple and transparent way.

Our privacy requirement will be based on the notion of ϵ-local differential privacy (Sarwate and Sankar, 2014), which is inherited from the celebrated concept of differential privacy proposed by Dwork (see comprehensive surveys by Dwork (2008); Leoni (2012); Ji et al. (2014)) by removing the adjacency relationship between data sets. The ϵ parameter in local differential privacy gives an appreciation of how uniformly hard it is to make inferences on the source realizations, regardless of the source distribution.

On the other hand, the fidelity will be linked to three alternative loss metrics—a family of Csiszár f-divergence metrics (notably including Kullback–Leibler (KL) divergence), mean-squared error (MSE) and total variation (TV)—between the exact source distribution and an estimate thereof from the privatized samples.[3] More specifically, the figure of merit will be the *speed of convergence* to zero of the expected fidelity loss (as measured by the metric of choice among the three metrics under study). For this purpose, we will derive asymptotic expressions of the expected MSE and TV losses for large sample sizes. As to the f-divergence metric, we will generalize (to the effect of including randomization) an asymptotic expansion of the expected KL divergence between the empirical distribution

3. Note that in some publications, the TV and MSE metrics are respectively referred to as ℓ_1 and ℓ_2-loss (Ye and Barg, 2018, 2017)

$\boldsymbol{t}(\boldsymbol{x}_n)$ of n i.i.d. samples of a random variable $X \in [K]$ gathered in a vector $\boldsymbol{x}_n \sim \boldsymbol{p}^{\otimes n}$, and its exact distribution $\boldsymbol{p} = (p_1, \ldots, p_K)$, as n tends to infinity (Abe, 1996):

$$\mathsf{E}\big[D\big(\boldsymbol{t}(\boldsymbol{x}_n)\big\|\boldsymbol{p}\big)\big] = \frac{K-1}{2n} + \left(\sum_{k=1}^{K} \frac{1}{p_k} - 1\right) \frac{1}{12n^2} + O(n^{-3}). \tag{1}$$

The first-order term of this expansion suggests that a low support set cardinality K will be beneficial for the speed of convergence, the second-order term suggests that for a given cardinality, the uniform distribution is most beneficial.

The main contributions of this article are a substantial expansion upon our predecessor conference paper (Pastore and Gastpar, 2016), and can be summarized as follows:

- While previous publications have already studied large-sample size asymptotic expansions of standard loss metrics, we have sought to generalize these derivations in several ways: (i) we force the distribution estimate to be a valid probability distribution and rigorously treat the error term that arises from this projection operation[4]; (ii) in addition to TV and MSE loss metrics (elsewhere referred to as ℓ_1 and ℓ_2 risk) we also cover a large class of f-divergence metrics; (iii) we highlight that maximum-likelihood and MMSE distribution estimators acquire a similar form and argue that they yield the same asymptotic loss.

- We identify the non-negative matrix $\boldsymbol{\Phi}(\boldsymbol{W}) = \boldsymbol{W}(\boldsymbol{W}^{-1} \odot \boldsymbol{W}^{-1})$ (where '$\odot$' denotes entrywise multiplication) as well as the sum of its entries $\varphi(\boldsymbol{W}) = \sum_{ij} \Phi_{ij}(\boldsymbol{W})$ as representative proxies for a larger class of fidelity metrics. These quantities essentially capture the fidelity metric's dependency on the random mechanism (row-stochastic matrix) $\boldsymbol{W}$. We study some of their properties, such as data-processing inequalities.

- We give a partial answer to the question as to which ϵ-private mechanism is optimal in the sense of minimizing the quantity $\varphi(\boldsymbol{W})$. What we prove is that among the class of *circulant* mechanisms, the so-called *step mechanism* (already widely studied in other publications on local differential privacy) minimizes $\varphi(\boldsymbol{W})$ (Theorem 9).

- We derive upper and lower bounds on the fundamental privacy–fidelity tradeoff for all three loss metrics, and for both problem formulations under consideration: the so-called *feasibility problem* and *minimax problem* (defined in Section 3.3). The lower bounds are all based on an important lower bound on $\varphi(\boldsymbol{W})$ which only depends on the privacy level ϵ and the source's support size K.

2. Notation

By convention, all vectors are row vectors unless transposed by $(\cdot)^{\mathrm{T}}$. We occasionally denote the inner product between two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ as $\langle \boldsymbol{a}, \boldsymbol{b} \rangle = \boldsymbol{a}\boldsymbol{b}^{\mathrm{T}}$. The product signs '$\odot$' and '$\otimes$' stand for the Hadamard product (entrywise multiplication) and the Kronecker product,

4. Kairouz et al. (2016a,b), for example, discusses projection operations to some detail and notices by simulation that a projected estimator tends to outperform its unprojected counterpart, but provides no analytic treatment of the error term. Similarly, Ye and Barg (2018) correctly point out that the impact of the projection operation is exponentially small, but omit the detailed analysis.

respectively. The exponent notation $\boldsymbol{a}^{\otimes n}$ stands for the n-fold Kronecker product $\boldsymbol{a}\otimes\ldots\otimes\boldsymbol{a}$. The bracket $[K]$ is shorthand for $\{1,2,\ldots,K\}$. The *type* or *empirical distribution* of a sample sequence $\boldsymbol{x}_n=(X_1,\ldots,X_n)\in[K]^n$ shall be denoted as $\boldsymbol{t}(\boldsymbol{x}_n)=[t_1(\boldsymbol{x}_n),\ldots,t_K(\boldsymbol{x}_n)]$ where $t_k(\boldsymbol{x}_n)=\frac{1}{n}\sum_{i=1}^n\mathbb{1}\{X_i=k\}$ with $\mathbb{1}\{\cdot\}$ standing for the indicator function. The all-ones row vector of dimension N is written as $\mathbf{1}_N$, or simply $\mathbf{1}$ if its dimension is clear from context. The probability simplex is denoted as $\mathbb{P}$ (whose dimension is always clear from context). We denote by $\delta_{k,\ell}$ the Dirac delta function, which equals one if $k=\ell$ and zero otherwise.

3. Problem description

A *privatization channel* or *mechanism* $\boldsymbol{W}$ with finite source alphabet $[K]$ and finite output alphabet $[L]$, is a discrete stochastic mapping (or *Markov kernel*) described by a matrix of conditional probabilities $\boldsymbol{W}$ with (k,ℓ)-th entry

$$W_{k,\ell}=\mathsf{Pr}\{Y=\ell|X=k\}.$$

The matrix $\boldsymbol{W}$ is row-stochastic, meaning that all its entries belong to the unit interval and that each row sums to one, i.e., $\boldsymbol{W}\mathbf{1}_L^{\mathrm{T}}=\mathbf{1}_K^{\mathrm{T}}$. This article is only concerned with square channels, hence $L=K$ throughout.

The privatization channel acts independently upon each of the n i.i.d. source symbols $\boldsymbol{x}_n=(X_1,\ldots,X_n)\sim\boldsymbol{p}^{\otimes n}$ to generate a sequence of i.i.d. *privatized* observations $\boldsymbol{y}_n=(Y_1,\ldots,Y_n)\sim\boldsymbol{q}^{\otimes n}$. The output distribution $\boldsymbol{q}$ induced by the source distribution $\boldsymbol{p}$ is given by right-multiplication of $\boldsymbol{p}$ with the channel matrix $\boldsymbol{W}$. Moreover, we require that $\boldsymbol{W}$ be *full-rank*. Hence,

$$\boldsymbol{q}=\boldsymbol{p}\boldsymbol{W}\qquad\qquad\boldsymbol{p}=\boldsymbol{q}\boldsymbol{W}^{-1}.$$

The curator is cognizant of $\boldsymbol{W}$, has access to the output sequence $\boldsymbol{y}_n$ and seeks to generate an estimate of $\boldsymbol{p}$, which we denote as $\hat{\boldsymbol{p}}_n$. Since $\boldsymbol{W}$ is square full-rank and the source is i.i.d., the quantity

$$\check{\boldsymbol{p}}_n\triangleq\boldsymbol{t}(\boldsymbol{y}_n)\boldsymbol{W}^{-1}\tag{2}$$

is a complete and minimally sufficient statistic for $\boldsymbol{p}$. Consequently, any estimator $\hat{\boldsymbol{p}}_n$ can be defined as a function of $\check{\boldsymbol{p}}_n$ (and possibly $\boldsymbol{W}$) without loss of optimality nor generality. Additionally, we require the estimator $\hat{\boldsymbol{p}}_n$ to be *consistent*, which means that

$$\lim_{n\to\infty}\mathsf{Pr}\{\|\hat{\boldsymbol{p}}_n-\boldsymbol{p}\|>\delta\}=0\tag{3}$$

for any $\delta>0$. Clearly, the speed at which this convergence takes place will depend on how "noisy" the mechanism $\boldsymbol{W}$ is designed to be. We would wish the convergence to be fast (for the sake of fidelity) while the mechanism should allow as little inference on X_i from Y_i as possible (for the sake of privacy), irrespective of the (unknown) source distribution. We now introduce the metrics for characterizing this privacy–fidelity trade-off.

3.1 Fidelity

To measure the accuracy of any given estimator $\hat{\boldsymbol{p}}_n$, we define three expected loss metrics: one based on Csiszár's f-divergence, one based on mean-squared error (MSE) and one based

on total variation distance (TV), namely[5]

$$\mathscr{L}^{(n)}_{f\text{-DIV}}(\boldsymbol{p}, \boldsymbol{W}) \triangleq \mathsf{E}\big[D_f\big(\hat{\boldsymbol{p}}_n \big\| \boldsymbol{p}\big)\big] \tag{4a}$$

$$\mathscr{L}^{(n)}_{\text{MSE}}(\boldsymbol{p}, \boldsymbol{W}) \triangleq \mathsf{E}\Big[\big\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\big\|_2^2\Big] \tag{4b}$$

$$\mathscr{L}^{(n)}_{\text{TV}}(\boldsymbol{p}, \boldsymbol{W}) \triangleq \mathsf{E}\Big[\big\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\big\|_1\Big]. \tag{4c}$$

Here, the f-divergence between two same-sized probability vectors $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{P}$ and for a convex function f satisfying $f(1) = 0$, is defined as

$$D_f(\boldsymbol{p}\|\boldsymbol{q}) = \sum_{k\in[K]} p_k f\left(\frac{q_k}{p_k}\right).$$

Note that TV distance is actually an f-divergence (for the function $f(x) = |x-1|$) whereas the MSE distance is not. However, we choose to single out TV distance as a separate metric, since the class of f-divergences that we shall focus on requires differentiability of the function $f(x)$ at $x = 1$, which excludes TV distance.

More specifically than the loss metrics (4a)–(4c) themselves, we will consider the *limiting ratios* between the loss metrics achieved *with* the privatizing mechanism $\boldsymbol{W}$ against the corresponding value that would be obtained from a clear view on the samples, i.e., *without* privatization. These asymptotic normalized loss metrics are defined as

$$\alpha_{f\text{-DIV}}(\boldsymbol{p}, \boldsymbol{W}) \triangleq \lim_{n\to\infty} \frac{\mathscr{L}^{(n)}_{f\text{-DIV}}(\boldsymbol{p}, \boldsymbol{W})}{\mathscr{L}^{(n)}_{f\text{-DIV}}(\boldsymbol{p}, \mathbf{I})} \tag{5a}$$

$$\alpha_{\text{MSE}}(\boldsymbol{p}, \boldsymbol{W}) \triangleq \lim_{n\to\infty} \frac{\mathscr{L}^{(n)}_{\text{MSE}}(\boldsymbol{p}, \boldsymbol{W})}{\mathscr{L}^{(n)}_{\text{MSE}}(\boldsymbol{p}, \mathbf{I})} \tag{5b}$$

$$\alpha_{\text{TV}}(\boldsymbol{p}, \boldsymbol{W}) \triangleq \lim_{n\to\infty} \left(\frac{\mathscr{L}^{(n)}_{\text{TV}}(\boldsymbol{p}, \boldsymbol{W})}{\mathscr{L}^{(n)}_{\text{TV}}(\boldsymbol{p}, \mathbf{I})}\right)^2, \tag{5c}$$

where $\mathbf{I}$ denotes the identity (non-privatizing) channel.[6] Since, as we shall see, the metrics (4a) and (4b) decay as $O(\frac{1}{n})$ when n tends to infinity, whereas (4c) decays as $O\big(\frac{1}{\sqrt{n}}\big)$ (notice the square in (5c) introduced to compensate for this fact), we can view the normalized quantities $\alpha_{f\text{-DIV}}$, α_{MSE} and α_{TV} as rules of thumb for the factor by which the sample size has to be increased if we want the accuracy of the privatized estimation to approximately match that of the non-privatized case.

5. Besides $(\boldsymbol{p}, \boldsymbol{W})$, the fidelity loss metrics (4a)–(4c) also depend on the estimator function, but we choose to omit this dependency for notational brevity. The same applies to the asymptotic metrics presented further below, in (5a)–(5c).

6. The identity can be replaced by any permutation matrix, since a permutation amounts to a relabeling of symbols. The estimators $\hat{\boldsymbol{p}}_n$ introduced in the next Subsection are indeed consistent with this permutation invariance, in the sense that they give $\mathscr{L}^{(n)}(\boldsymbol{p}, \boldsymbol{W}) = \mathscr{L}^{(n)}(\boldsymbol{p}, \boldsymbol{W}\boldsymbol{\Pi})$ for any permutation matrix $\boldsymbol{\Pi}$.

3.2 Privacy

Our definition of privacy is based on the concept of *local differential privacy*. A privatization channel $\boldsymbol{W} = [W_{k,\ell}]_{k,\ell}$ as defined above is said to be ϵ-locally differentially private (or ϵ-private) if for all index triples $(k, k', \ell) \in [K]^3$, we have

$$W_{k,\ell} \leq e^{\epsilon} W_{k',\ell}. \tag{6}$$

For a given channel $\boldsymbol{W}$, we denote by $\epsilon(\boldsymbol{W})$ the smallest value of ϵ such that (6) holds for all (k, k', ℓ), i.e.,

$$\epsilon(\boldsymbol{W}) = \log\left(\max_{k,k',\ell} \frac{W_{k,\ell}}{W_{k',\ell}}\right). \tag{7}$$

Let $\mathscr{W}$ denote the set of all $K \times K$ full-rank row-stochastic matrices and let $\mathscr{W}_\epsilon \subset \mathscr{W}$ denote the set of ϵ-private mechanisms.

3.3 Two problem formulations for the privacy–fidelity trade-off

Now that we have introduced the fidelity loss and privacy metrics in (5) and (7) respectively, we can formulate the privacy–fidelity trade-off problem. We propose two different problem formulations, which we will refer to as the *feasibility problem* and the *minimax problem*, respectively. In the following, the generic notation α may refer to either $\alpha_{f\text{-DIV}}$, α_{MSE} or α_{TV}.

3.3.1 Feasibility problem

Assume that the source distribution $\boldsymbol{p}$ is fixed. We seek to characterize the set $\mathscr{F}(\boldsymbol{p})$ of all (ϵ, α) pairs which are jointly feasible, i.e.,

$$\mathscr{F}(\boldsymbol{p}) \triangleq \Big\{ \big(\alpha(\boldsymbol{p}, \boldsymbol{W}), \epsilon(\boldsymbol{W})\big) : \boldsymbol{W} \in \mathscr{W} \Big\}.$$

Specifically, we seek to characterize the optimal ϵ-α trade-off curve

$$\alpha^\star(\epsilon; \boldsymbol{p}) \triangleq \min_{\boldsymbol{W} \in \mathscr{W}_\epsilon} \alpha(\boldsymbol{p}, \boldsymbol{W}) \tag{8}$$

Note that, as we have stressed previously, the curator (i.e., the designer of $\boldsymbol{W}$) has no knowledge of $\boldsymbol{p}$. Hence, the optimal trade-off curve $\alpha^\star(\epsilon; \boldsymbol{p})$ is achieved in the event that the curator makes the best guess about the optimal $\boldsymbol{W}$ for a given $\boldsymbol{p}$, as if aided by a genie who hands over the knowledge of $\boldsymbol{p}$. In general though, no design strategy for $\boldsymbol{W}$ can leverage knowledge about $\boldsymbol{p}$, and thus will fall short of achieving $\alpha^\star(\epsilon; \boldsymbol{p})$.

3.3.2 Minimax problem

Assume that the source distribution $\boldsymbol{p}$ can be any among a continuous subset $\mathscr{P}$ of the probability simplex $\mathbb{P}$. The curator seeks to optimize $\boldsymbol{W}$ based on this knowledge of the continuous candidate set $\mathscr{P}$. Hence, we define the minimax problem as being that of determining

$$\alpha^\star(\epsilon; \mathscr{P}) \triangleq \min_{\boldsymbol{W} \in \mathscr{W}_\epsilon} \sup_{\boldsymbol{p} \in \mathscr{P}} \alpha(\boldsymbol{p}, \boldsymbol{W}). \tag{9}$$

Reducing $\mathscr{P}$ to a singleton set $\{\boldsymbol{p}\}$ would make both problem formulations mathematically identical, in the sense that (9) would equal (8). However, one has to bear in mind that the interpretations of both problem formulations are rather different. Indeed, the minimax problem makes sense mostly for a non-singleton set $\mathscr{P}$. Besides, assuming a singleton set would violate the assumption that $\mathscr{P}$ is continuous, which is important for another reason: if $\mathscr{P}$ were discrete, it would be more adequate to consider the problem within a guessing or multiple hypothesis testing framework. In this case, the guessing error would be represented by error probabilities of different types (e.g., false alarm, missed detection, etc.). For finite $\mathscr{P}$, these error probabilities would typically decay exponentially in the sample size n, so a natural candidate for the fidelity loss metric would be the error exponent, rather that the quantities α studied in this article. Besides the feasibility and the minimax problem, one can think of other questions about fundamental limits which might be of independent interest as well, but will not be addressed in this article. Let us mention only one example:

3.3.3 Best-case feasibility problem

The best-case trade-off $\inf_{\boldsymbol{p}} \alpha^\star(\epsilon;\boldsymbol{p})$ delimits the union $\bigcup_{\boldsymbol{p}} \mathscr{F}(\boldsymbol{p})$ and characterizes the most optimistic performance limit, in the sense that the source distribution is most benevolent, and the curator $\boldsymbol{W}$ guesses the best $\boldsymbol{W}$. This limiting curve will only depend on the alphabet dimension K (and possibly on the fidelity metric of choice, be it $\alpha_{f\text{-DIV}}$, α_{MSE} or α_{TV}).

4. Distribution estimation

As we have argued before, any distribution estimator can be expressed as a function of $\check{\boldsymbol{p}}_n$ (and possibly $\boldsymbol{W}$) to the K-dimensional probability simplex. Henceforth, we shall only consider estimators that are *projectors* of $\check{\boldsymbol{p}}_n$ onto the probability simplex,[7] namely, estimators that can be cast into the form

$$\hat{\boldsymbol{p}}_n = \text{Proj}_{\mathbb{P}}(\check{\boldsymbol{p}}_n) \tag{10}$$

where $\text{Proj}_{\mathbb{P}}(\cdot)$ stands for some idempotent function satisfying $\text{Proj}_{\mathbb{P}}(\check{\boldsymbol{p}}) = \check{\boldsymbol{p}}$ for any probability vector $\check{\boldsymbol{p}} \in \mathbb{P}$.

Let us denote the topological interior and closure (in $\mathbb{P}$) of a set of distributions $\mathscr{R} \subset \mathbb{P}$ as $\mathscr{R}^\circ$ and $\overline{\mathscr{R}}$, respectively, and define its boundary as $\partial\mathscr{R} = \overline{\mathscr{R}} \setminus \mathscr{R}^\circ$. Henceforth, for a distribution $\boldsymbol{r} \in \mathbb{P}$ and a set $\mathscr{R} \subset \mathbb{P}$, we adopt Csiszár's notation (Csiszár, 1984) for information projection

$$D(\mathscr{R}\|\boldsymbol{r}) \triangleq \inf_{\boldsymbol{r}' \in \mathscr{R}} D(\boldsymbol{r}'\|\boldsymbol{r})$$

where $D(\cdot\|\cdot)$ denotes the Kullback–Leibler divergence.

Lemma 1 *The following inequality holds:*[8]

$$\mathsf{Pr}\big\{\check{\boldsymbol{p}}_n \notin \mathbb{P}\big\} \leq e^{-nD(\partial\mathbb{P}\boldsymbol{W}\|\boldsymbol{p}\boldsymbol{W})}.$$

7. Note that $\check{\boldsymbol{p}}_n$ [cf. (2)] is not guaranteed to be a probability vector. Due to $\boldsymbol{W}$ being row-stochastic, both $\boldsymbol{W}$ and $\boldsymbol{W}^{-1}$ have 1 as an eigenvalue, with associated all-ones eigenvector $\boldsymbol{1}^{\mathrm{T}}$. Thus, the rows of $\boldsymbol{W}^{-1}$ and hence the entries of $\check{\boldsymbol{p}}_n = \boldsymbol{t}(\boldsymbol{y}_n)\boldsymbol{W}^{-1}$ sum to one. However, some entries of $\check{\boldsymbol{p}}_n$ may lie outside the unit interval. Hence the necessity of a projection operation.

8. We omit parentheses in writing $\partial\mathbb{P}\boldsymbol{W}$ because regardless of how we set parentheses, $\partial(\mathbb{P}\boldsymbol{W}) = (\partial\mathbb{P})\boldsymbol{W}$.

Proof See Appendix A. ■

It is easy to show with Lemma 1 that any estimator of the form (10) is consistent [cf. Section 3], because the cases $\check{\boldsymbol{p}} \neq \hat{\boldsymbol{p}}$ are at least exponentially rare, whereas fidelity metrics decay linearly in the sample size (as we shall see). Moreover, imposing the form (10) outrules the possibility of trivial "genie-aided" estimators such as $\hat{\boldsymbol{p}}_n = \boldsymbol{p}$ by construction.

In addition, the fact that $\check{\boldsymbol{p}}_n$ fails to be a probability vector only in exponentially rare cases, as highlighted by Lemma 1, is helpful in that our asymptotic loss metrics (5a)–(5c) do not depend on $\mathrm{Proj}_{\mathbb{P}}$ (as we shall see). Therefore, all estimators of the form (10) can be regarded as asymptotically equivalent.

Next, we will derive the maximum likelihood (ML) and the minimum mean-squared error (MMSE) distribution estimators, and verify that they are two instances of the general form (10).

4.1 ML estimator

Based on a given output sequence $\tilde{\boldsymbol{y}}_n$, the ML estimator of the source distribution is defined as the distribution that maximizes the probability of the event $\boldsymbol{y}_n = \tilde{\boldsymbol{y}}_n$. By a classic argument, one can express the ML estimator as a KL divergence minimizer:

$$\begin{aligned}
\hat{\boldsymbol{p}}_{\mathrm{ML}}(\tilde{\boldsymbol{y}}_n) &= \operatorname*{argmax}_{\boldsymbol{p}' \in \mathbb{P}} \frac{1}{n} \log \mathsf{Pr}\big\{ \boldsymbol{y}_n = \tilde{\boldsymbol{y}}_n \;\big|\; \boldsymbol{y}_n \sim (\boldsymbol{p}'\boldsymbol{W})^{\otimes n} \big\} \\
&= \operatorname*{argmax}_{\boldsymbol{p}' \in \mathbb{P}} \frac{1}{n} \log \prod_{i=1}^{n} \mathsf{Pr}\big\{ y_i = \tilde{y}_i \;\big|\; y_i \sim \boldsymbol{p}'\boldsymbol{W} \big\} \\
&= \operatorname*{argmax}_{\boldsymbol{p}' \in \mathbb{P}} \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}\{\tilde{y}_i = k\} \log\big([\boldsymbol{p}'\boldsymbol{W}]_k\big) \\
&= \operatorname*{argmax}_{\boldsymbol{p}' \in \mathbb{P}} \sum_{k=1}^{K} t_k(\tilde{\boldsymbol{y}}_n) \log\big([\boldsymbol{p}'\boldsymbol{W}]_k\big) \\
&= \operatorname*{argmin}_{\boldsymbol{p}' \in \mathbb{P}} D\big(\boldsymbol{t}(\tilde{\boldsymbol{y}}_n) \big\| \boldsymbol{p}'\boldsymbol{W}\big). \qquad (11)
\end{aligned}$$

This optimization problem is convex, because the KL divergence is a convex functional and the probability simplex $\mathbb{P}$ is a convex set. To see that this is an instance of (10), it suffices to rewrite $\boldsymbol{t}(\tilde{\boldsymbol{y}}_n)$ as $\boldsymbol{t}(\tilde{\boldsymbol{y}}_n)\boldsymbol{W}^{-1}\boldsymbol{W}$ in (11) so that the idempotence property of projection becomes evident.

4.2 MMSE estimator

The MMSE estimator is defined as the probability vector that minimizes the Euclidean distance between the output distribution it induces, and the empirical output distribution:

$$\hat{\boldsymbol{p}}_{\mathrm{MMSE}}(\tilde{\boldsymbol{y}}_n) = \operatorname*{argmin}_{\boldsymbol{p}' \in \mathbb{P}} \big\| \boldsymbol{t}(\tilde{\boldsymbol{y}}_n) - \boldsymbol{p}'\boldsymbol{W} \big\|_2. \qquad (12)$$

This estimator is similar to the ML estimator, except for replacing the KL divergence in (11) by Euclidean distance. Similarly to the ML estimator, the minimization problem in (12) is convex and manifestly an instance of (10).

5. Convergence of estimates

Recall that $\boldsymbol{p}$ is supported on $[K]$, meaning that all its entries p_k are positive. As a consequence, the convergence in probability (3) implies the convergence to zero of all fidelity loss metrics (including f-divergences) as $n \to \infty$, i.e.,

$$\lim_{n\to\infty} \mathscr{L}^{(n)}(\boldsymbol{p}, \boldsymbol{W}) = 0.$$

As a representative figure for the speed of this convergence to zero, we shall compute the leading terms in the respective asymptotic expansions (as $n \to \infty$) of the different loss metrics. Prior to providing analytical expressions for these, we need to introduce some quantities of interest. For a positive integer ρ, let us define

$$\nu_{\rho,k} \triangleq \boldsymbol{p}\boldsymbol{W}\big(\underbrace{\boldsymbol{W}^{-1} \odot \boldsymbol{W}^{-1} \odot \ldots \odot \boldsymbol{W}^{-1}}_{\rho \text{ factors}}\big)\boldsymbol{e}_k^{\mathrm{T}} \tag{13}$$

where '$\odot$' denotes entrywise multiplication. In particular, the matrix

$$\boldsymbol{\Phi}(\boldsymbol{W}) \triangleq \boldsymbol{W}(\boldsymbol{W}^{-1} \odot \boldsymbol{W}^{-1}) \tag{14}$$

involved in the expression of $\nu_{2,k}$ will play a prominent role in the fidelity loss metrics and will be shown to satisfy data-processing inequalities. Finally, the sum of all entries of $\boldsymbol{\Phi}(\boldsymbol{W})$, i.e.,

$$\varphi(\boldsymbol{W}) \triangleq \mathbf{1}\boldsymbol{\Phi}(\boldsymbol{W})\mathbf{1}^{\mathrm{T}} = \sum_{k=1}^{K}\sum_{\ell=1}^{K} \Phi_{k,\ell}(\boldsymbol{W}) \tag{15}$$

will be repeatedly used.

The following theorem gives an asymptotic expansion applicable to a large class of f-divergences, including KL divergence:

Theorem 2 (Expansion of f-divergence loss) *Assume that*

1. *$f(1) = 0$ and $f(x)$ is four times differentiable at $x = 1$,*
2. *$f(0)$ is finite*[9]
3. *$f(x)$ can be expanded as*

$$f(x) = \sum_{\rho=1}^{4} \frac{f^{(\rho)}(1)}{\rho!}(x-1)^{\rho} + O\big(|x-1|^{4+\gamma}\big)$$

for some $\gamma > 0$, where $f^{(\rho)}(x)$ denotes the ρ-th derivative of $f(x)$.

9. The requirement that $f(0)$ be finite ensures that the expected value $\mathsf{E}\big[D_f\big(\hat{\boldsymbol{p}}_n \big\| \boldsymbol{p}\big)\big]$ is finite, for otherwise there would be a positive probability of $D_f\big(\hat{\boldsymbol{p}}_n \big\| \boldsymbol{p}\big)$ being infinite for any n, and thus its expectation would be infinite.

Furthermore, assume that $\boldsymbol{W}$ is full-rank (invertible). Then the following asymptotic expansion holds:

$$\mathscr{L}^{(n)}_{f\text{-DIV}}(\boldsymbol{p},\boldsymbol{W}) = \frac{Af''(1)}{2n} + \left(\frac{Bf^{(3)}(1)}{6} + \frac{Cf^{(4)}(1)}{8}\right)\frac{1}{n^2} + O(n^{-3}) \tag{16}$$

with coefficients A, B and C given by

$$A = -1 + \sum_{k=1}^{K} \frac{\nu_{2,k}}{\nu_{1,k}} \tag{17a}$$

$$B = 2 + \sum_{k=1}^{K} \left(\frac{\nu_{3,k}}{\nu_{1,k}^2} - 3\frac{\nu_{2,k}}{\nu_{3,k}}\right) \tag{17b}$$

$$C = 1 + \sum_{k=1}^{K} \left(\frac{\nu_{2,k}^2}{\nu_{1,k}^3} - 2\frac{\nu_{2,k}}{\nu_{1,k}}\right) \tag{17c}$$

where $\nu_{\rho,k}$ is defined in (13)*. (Note that $\nu_{1,k}$ is simply p_k)*

	$f(x)$	$f^{(1)}(1)$	$f^{(2)}(1)$	$f^{(3)}(1)$	$f^{(4)}(1)$
KL divergence	$x\ln(x)$	1	1	-1	2
Hellinger distance	$(\sqrt{x}-1)^2$	0	$1/2$	$-3/4$	$15/8$
	$1-\sqrt{x}$	$-1/2$	$1/4$	$-3/8$	$15/16$
Pearson χ^2 divergence	$(x-1)^2$	0	2	0	0
	x^2-1	2	2	0	0
Triangular discrimination	$\frac{(x-1)^2}{x+1}$	0	1	$-3/2$	3
TV distance	$\lvert x-1\rvert$	–	–	–	–

Table 1: First four derivatives at 1 of functions $f(x)$ associated to different f-divergences. Some definitions vary across the literature.

If one assumes that f is only twice differentiable and that $f(x) = f^{(1)}(1)(x-1) + \frac{1}{2}f^{(2)}(1)(x-1)^2 + O\big(|x-1|^{2+\gamma}\big)$, one can also prove a simpler version of Theorem 2, namely that $\mathscr{L}^{(n)}_{f\text{-DIV}}(\boldsymbol{p},\boldsymbol{W}) = \frac{Af''(1)}{2n} + O(n^{-2})$.

Theorem 3 (Expansion of MSE loss) *It holds that*[10]

$$\mathscr{L}^{(n)}_{\text{MSE}}(\boldsymbol{p},\boldsymbol{W}) = \frac{1}{n}\sum_{k=1}^{K}\big(\nu_{2,k} - \nu_{1,k}^2\big) + O\big(e^{-nD(\partial\mathbb{P}_{\boldsymbol{W}}\|\boldsymbol{p}\boldsymbol{W})}\big). \tag{18}$$

10. Unlike the differentiable f-divergence metrics which can be Taylor-expanded [cf. (16)] to, in general, an infinity of terms, the MSE metric expansion has only a single $O(n^{-1})$ term. That is, it is exact up to an exponential remainder term which is attributable to the projection operation.

Theorem 4 (Expansion of TV loss) *It holds that*

$$\mathscr{L}_{\mathrm{TV}}^{(n)}(\boldsymbol{p},\boldsymbol{W}) = \sqrt{\frac{2}{\pi n}}\sum_{k=1}^{K}\sqrt{\nu_{2,k}-\nu_{1,k}^2} + o\Big(\frac{1}{\sqrt{n}}\Big). \tag{19}$$

The proofs of Theorems 2, 3 and 4 are given in Appendices B, D and E, respectively. Note that the proof of Theorem 2 essentially relies on a Taylor expansion, which in principle could be carried further to produce more asymptotic terms, provided that the function f is differentiable enough times around zero. By applying these Theorems on the definition of the normalized first-order terms (5), while calling to mind that $\nu_{1,k} = p_k$ and that $\nu_{2,k} = \boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W})\boldsymbol{e}_k^{\mathrm{T}}$, we obtain the explicit expressions

$$\alpha_{f\text{-DIV}}(\boldsymbol{p},\boldsymbol{W}) = \frac{\boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W})\boldsymbol{p}^{-\mathrm{T}} - 1}{K-1} \tag{20a}$$

$$\alpha_{\mathrm{MSE}}(\boldsymbol{p},\boldsymbol{W}) = \frac{\boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W})\mathbf{1}^{\mathrm{T}} - \|\boldsymbol{p}\|_2^2}{1-\|\boldsymbol{p}\|_2^2} \tag{20b}$$

$$\alpha_{\mathrm{TV}}(\boldsymbol{p},\boldsymbol{W}) = \left(\frac{\big\langle \mathbf{1}, \sqrt{\boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W}) - \boldsymbol{p}\odot\boldsymbol{p}}\big\rangle}{\langle \mathbf{1}, \sqrt{\boldsymbol{p}-\boldsymbol{p}\odot\boldsymbol{p}}\rangle}\right)^2 \tag{20c}$$

where in (20c), square roots on vectors are applied entrywise. Notice how $\boldsymbol{\Phi}(\boldsymbol{W})$ plays a central role, in that it fully captures the dependency in $\boldsymbol{W}$ of all three loss metrics. To better appreciate the significance and behavior of these metrics, a few remarks are in order.

Remark 1 *Interestingly, for the uniform source $\boldsymbol{p} = \mathbf{1}/K$, the f-divergence and MSE metrics happen to coincide, regardless of the choice of mechanism $\boldsymbol{W} \in \mathscr{W}$:*

$$\alpha_{f\text{-DIV}}\big(\tfrac{\mathbf{1}}{K},\boldsymbol{W}\big) = \alpha_{\mathrm{MSE}}\big(\tfrac{\mathbf{1}}{K},\boldsymbol{W}\big) = \frac{\varphi(\boldsymbol{W})-1}{K-1}. \tag{21}$$

By contrast, the TV metric is generally smaller, which can be shown by Jensen's inequality:

$$\alpha_{\mathrm{TV}}\big(\tfrac{\mathbf{1}}{K},\boldsymbol{W}\big) = \left(\frac{\sum_k \sqrt{K\sum_\ell \Phi_{k,\ell}(\boldsymbol{W}) - 1}}{K\sqrt{K-1}}\right)^2 \leq \frac{\varphi(\boldsymbol{W})-1}{K-1}. \tag{22}$$

As we shall see in Lemma 8 of Section 7, this inequality becomes tight when the mechanism is circulant, thus making all three metrics match in such case.

Remark 2 *In the noiseless case, $\boldsymbol{W}$ and its inverse $\boldsymbol{W}^{-1}$ are permutation matrices ($\boldsymbol{\Pi}$ and $\boldsymbol{\Pi}^{\mathrm{T}}$, respectively), whose entries are equal to 0 or 1. Entrywise squaring leaves them unchanged, hence $\boldsymbol{\Phi}(\boldsymbol{\Pi}) = \boldsymbol{\Pi}(\boldsymbol{\Pi}^{\mathrm{T}} \odot \boldsymbol{\Pi}^{\mathrm{T}}) = \boldsymbol{\Pi}\boldsymbol{\Pi}^{\mathrm{T}} = \mathbf{I}$ and we see that the metrics* (20a)–(20c) *all become equal to one, which is consistent with our expectation based on the definitions* (5a)–(5c)*. Furthermore we can verify that for $\boldsymbol{W}$ a permutation, specializing Theorem 2 to the KL divergence recovers the asymptotic expansion* (1) *by evaluation of* (16)–(17c)*.*

Remark 3 *As we shall see in the next section [cf.* (28)*], it holds that $[\boldsymbol{\Phi}(\boldsymbol{W})]_{k,\ell} \geq \delta_{k,\ell}$ (where $\delta_{k,\ell}$ stands for the Dirac delta), thanks to which it becomes manifest that the metrics* (20a)–(20c) *are larger or equal to one.*

Remark 4 *Concerning the dependency on $\boldsymbol{p}$, we notice from inspecting* (20a)–(20c) *that $\alpha_{\mathrm{MSE}}(\boldsymbol{p},\boldsymbol{W})$ and $\alpha_{\mathrm{TV}}(\boldsymbol{p},\boldsymbol{W})$ become large if $\boldsymbol{p}$ tends to a canonical base vector (i.e., when $\|\boldsymbol{p}\|_2^2 \to 1$ or equivalently $\langle \mathbf{1}, \sqrt{\boldsymbol{p}-\boldsymbol{p}\odot\boldsymbol{p}}\,\rangle \to 1$) whereas for the divergence metric $\alpha_{f\text{-DIV}}(\boldsymbol{p},\boldsymbol{W})$, it already suffices to have at least one low-probability symbol ($p_k \to 0$ for some $k \in [K]$) for $\alpha_{f\text{-DIV}}(\boldsymbol{p},\boldsymbol{W})$ to become large. Loosely speaking, low-probability symbols are more "penalizing" for the f-divergence loss metric than they are for the MSE or TV loss metrics.*

To illustrate Theorems 2, 3 and 4, we provide numerical evaluations of $\mathscr{L}^{(n)}_{f\text{-DIV}}$, $\mathscr{L}^{(n)}_{\mathrm{MSE}}$ and $\mathscr{L}^{(n)}_{\mathrm{TV}}$ for both the ML and MSE estimators (11) and (12), respectively, and compare them against the non-privatized performance. For the ϵ-private mechanism, we choose the matrix

$$\boldsymbol{W}_{\epsilon,\star} = \frac{1}{e^\epsilon + K - 1}\begin{bmatrix} e^\epsilon & 1 & \dots & 1 \\ 1 & e^\epsilon & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \dots & 1 & e^\epsilon \end{bmatrix} \tag{23}$$

which we shall call the *step mechanism*, and for which we will prove optimality results further on. With this choice of mechanism, the ML and MMSE estimators (11) and (12) are expressible by waterfilling-type closed forms

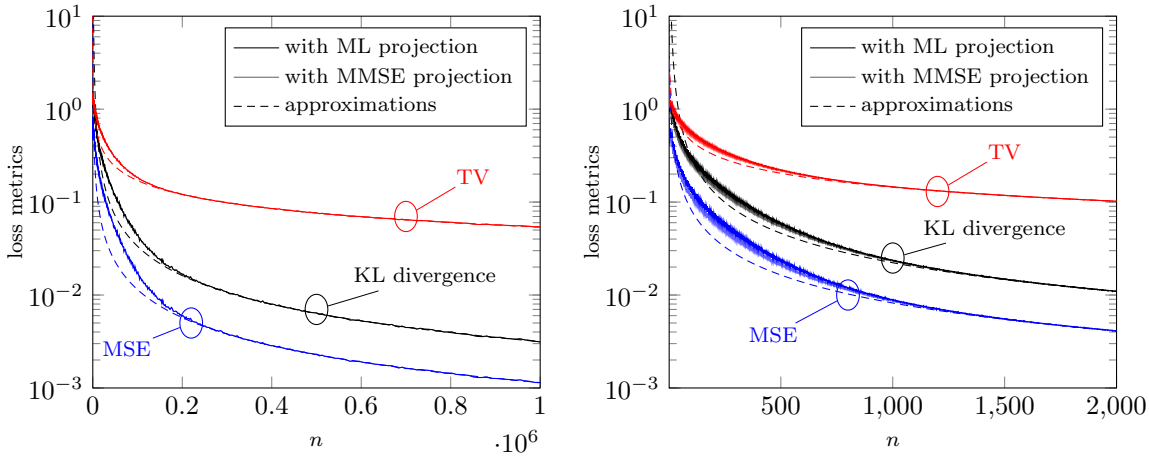
$$\begin{aligned}\hat{p}_{\mathrm{ML},k} &= \frac{1}{e^\epsilon - 1}\max\big\{0, (e^\epsilon + K - 1)t_k + \eta_{\mathrm{ML}}\big\} \\ \hat{p}_{\mathrm{MMSE},k} &= \frac{1}{e^\epsilon - 1}\max\big\{0, \eta_{\mathrm{MMSE}} t_k - 1\big\}\end{aligned}$$

where the scalars η_{ML} and η_{MMSE} are chosen such that the sum constraints $\sum_k \hat{p}_{\mathrm{ML},k} = \sum_k \hat{p}_{\mathrm{MMSE},k} = 1$ are met.

Evaluating the quantities $\nu_{\rho,k}$ defined in (13) for the step mechanism (23) yields

$$\begin{aligned}\nu_{1,k} &= p_k \\ \nu_{2,k} &= \frac{1}{(e^\epsilon-1)^2}\big((e^\epsilon-1)(e^\epsilon+K-3)p_k + e^\epsilon + K - 2\big) \\ \nu_{3,k} &= \frac{1}{(e^\epsilon-1)^3(e^\epsilon+K-1)}\big((e^\epsilon-1)((e^\epsilon+K-2)^3+1)p_k \\ &\qquad + (e^\epsilon+K-1)(e^\epsilon+K-2)(e^\epsilon+K-3)\big).\end{aligned}$$

Figure 2 shows how the fidelity loss metrics decay with n. The exact values of $\mathscr{L}^{(n)}_{f\text{-DIV}}$ (for KL divergence $f(x) = x\ln(x)$), $\mathscr{L}^{(n)}_{\mathrm{MSE}}$ and $\mathscr{L}^{(n)}_{\mathrm{TV}}$ are plotted against the second-order approximation (16) and the first-order approximations (18) and (19), respectively. The same plot is exhibited twice, once for $\epsilon = 10^{-1}$ (Figure 2a) and once for $\epsilon = 1$ (Figure 2b). Though it is not the focus of this article, it is worth mentioning that the small fluctuations visible on Figure 2a for low values of n are traceable to how empirical distributions are better approximations of the limiting distribution for certain values of n than for others, an effect related to Diophantine approximations.



(a) $\epsilon = 0.1$ (b) $\epsilon = 1$

Figure 2: Unnormalized loss metrics $\mathscr{L}^{(n)}(\boldsymbol{p}, \boldsymbol{W}_{\epsilon,\star})$ (for f-divergence, MSE and TV metrics) plotted as functions of the sample size n. By way of example, we have chosen $K = 4$, $\epsilon = 1$, $\boldsymbol{p} = [0.5\ 0.25\ 0.125\ 0.125]$ and the step mechanism $\boldsymbol{W}_{\epsilon,\star}$. The dashed curves show the approximations via truncated expansion, obtained from omitting the big-O or little-o remainder terms in (16), (18), (19), respectively. For each loss metric, one solid curve (opaque) shows the loss metric obtained with ML projection, whereas the other curve (semi-opaque) shows the loss metric obtained with MMSE projection. Note that in Figure 2a, these two curves are nearly indistinguishable.

6. Data processing theorems

Intuitively, it is clear that any estimator of the source distribution $\boldsymbol{p}$ based on the privatized observations $\boldsymbol{y}_n$ will perform worse, in terms of metrics (4a)–(4c), than the empirical distribution estimator $\boldsymbol{t}(\boldsymbol{x}_n)$ based on a clear view of the source symbol vector $\boldsymbol{x}_n$. This idea is formalized by the data-processing theorems stated below.

Theorem 5 (General form) *Consider n copies of the setup as depicted in Figure 6. That is, assume that $\boldsymbol{y}_n \sim (\boldsymbol{pW})^{\otimes n}$ and $\boldsymbol{y}'_n \sim (\boldsymbol{pWW'})^{\otimes n}$ are obtained from passing the samples $\boldsymbol{x}_n$ through n copies of the channels $\boldsymbol{W'}$ and $\boldsymbol{WW'}$, respectively. In addition, assume that*

1. *for f-divergence metrics, $f(x)$ is strictly convex*[11] *at $x = 1$;*

2. *$\boldsymbol{W'}$ is not a permutation.*

11. Note that strict local convexity is satisfied by all f-divergences of interest which also satisfy the conditions of Theorem 2. It is also satisfied by TV distance, in the sense that for $f(x) = |x - 1|$, we have $f(\lambda(1-\epsilon) + (1-\lambda)(1+\epsilon')) < \lambda f(1-\epsilon) + (1-\lambda) f(1+\epsilon')$ for all $\epsilon, \epsilon' > 0$.

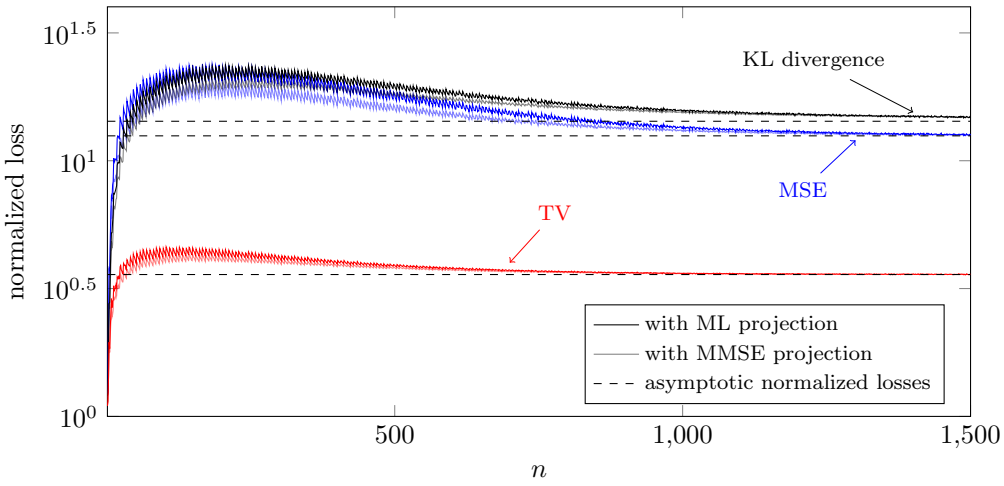


Figure 3: Normalized loss metrics $\mathscr{L}^{(n)}(\boldsymbol{p}, \boldsymbol{W}_{\epsilon,\star})/\mathscr{L}^{(n)}(\boldsymbol{p}, \mathbf{I})$ plotted as functions of the sample size n, with same parameters as in Figure 2. Note that curves corresponding to ML projection are black, whereas curves corresponding to MMSE projection are gray. For better visualization in regions where curves overlap, we have chosen dotted curves for the MSE metric. The dashed curves show the limits as $n \to \infty$, which correspond to the quantities (20a)–(20c).

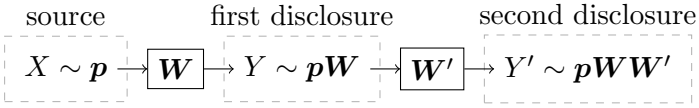


Figure 4: Data processing theorems.

Then, for any estimator of the form (10)*, for any source distribution* $\boldsymbol{p}$ *supported on* $[K]$ *and for sufficiently large* n*, we have*

$$\mathscr{L}^{(n)}(\boldsymbol{p}, \boldsymbol{W}\boldsymbol{W}') > \mathscr{L}^{(n)}(\boldsymbol{p}, \boldsymbol{W}).$$

Proof See Appendix F. The proof essentially exploits convexity, which all loss metrics have in common. Note, however, that the proof is slightly difficilitated by our rigorous treatment of the exponentially decaying remainder terms that arise due to the normalization of our pmf estimators $\hat{\boldsymbol{p}}_n$. ■

Unsurprisingly, this data-processing relationship carries over directly to the respective leading terms in the asymptotic expansions of the three loss metrics, as stated in the following corollary.

Corollary 6 (Data-processing inequality for α) *For any source distribution $\boldsymbol{p}$ and channels $\boldsymbol{W}$ and $\boldsymbol{W}'$, it holds that*

$$\alpha_{f\text{-DIV}}(\boldsymbol{p}, \boldsymbol{W}\boldsymbol{W}') \geq \alpha_{f\text{-DIV}}(\boldsymbol{p}, \boldsymbol{W}) \tag{26a}$$
$$\alpha_{\text{MSE}}(\boldsymbol{p}, \boldsymbol{W}\boldsymbol{W}') \geq \alpha_{\text{MSE}}(\boldsymbol{p}, \boldsymbol{W}) \tag{26b}$$
$$\alpha_{\text{TV}}(\boldsymbol{p}, \boldsymbol{W}\boldsymbol{W}') \geq \alpha_{\text{TV}}(\boldsymbol{p}, \boldsymbol{W}) \tag{26c}$$

with equality if and only if $\boldsymbol{W}'$ is a permutation.

Recall that $\alpha(\boldsymbol{p}, \mathbf{I}) = 1$. By setting $\boldsymbol{W}$ to the identity in (26a)–(26c), we recover the already known fact that $\alpha(\boldsymbol{p}, \boldsymbol{W}') \geq 1$ holds for any mechanism $\boldsymbol{W}' \in \mathscr{W}$.

Proof Suppose that there exists a source distribution $\boldsymbol{p}$ and a pair of channels $(\boldsymbol{W}, \boldsymbol{W}')$ such that $\alpha(\boldsymbol{p}, \boldsymbol{W}\boldsymbol{W}') < \alpha(\boldsymbol{p}, \boldsymbol{W})$ (where α stands for either metric). Then, by Definitions (5a)–(5c), there exist arbitrarily large sample sizes n such that $\mathscr{L}^{(n)}(\boldsymbol{p}, \boldsymbol{W}\boldsymbol{W}') < \mathscr{L}^{(n)}(\boldsymbol{p}, \boldsymbol{W})$. This would contradict Theorem 5. ■

Theorem 7 (Data-processing inequality for $\boldsymbol{\Phi}$) *For any square full-rank channels $\boldsymbol{W}$ and $\boldsymbol{W}'$ of same size, it holds that*[12]

$$\boldsymbol{\Phi}(\boldsymbol{W}\boldsymbol{W}') \geq \boldsymbol{\Phi}(\boldsymbol{W}). \tag{27}$$

with equality[13] *if and only if $\boldsymbol{W}'$ is a permutation.*

Proof See Appendix G. ■

Note that the above-referenced proof of Theorem 7 in Appendix G hinges on Corollary 6 (of Theorem 5). On the other hand, this Corollary 6 could clearly also be recovered as a corollary, in fact, of Theorem 7, since $\alpha_{f\text{-DIV}}$, α_{MSE} and α_{TV} are all monotone functions of the entries of $\boldsymbol{\Phi}$ [cf. (20a)–(20c)]. This means that Theorem 7 and Corollary 6 (to be precise, any one of the three inequalities in Corollary 6) are in fact equivalent statements. A self-contained proof of Theorem 7 (which we do not provide) would consist, by contrast, in proving (27) based on the assumption of row-stochasticity of $\boldsymbol{W}$ and $\boldsymbol{W}'$ alone.

It is instructive to particularize (27) by setting $\boldsymbol{W} = \mathbf{I}$ and using the fact that $\boldsymbol{\Phi}(\mathbf{I}) = \mathbf{I}$. We obtain $\boldsymbol{\Phi}(\boldsymbol{W}') \geq \mathbf{I}$ (which holds for any row-stochastic $\boldsymbol{W}'$), or equivalently,

$$\Phi_{k,\ell}(\boldsymbol{W}') \geq \begin{cases} 1 & \text{if } k = \ell \\ 0 & \text{if } k \neq \ell. \end{cases} \tag{28}$$

This inequality allows to immediately grasp why the normalized metrics $\alpha_{f\text{-DIV}}$, α_{MSE} and α_{TV}, as written out in (20a)–(20c), are quantities larger or equal to one (cf. Remark 3).

12. For matrix-valued $\boldsymbol{A}$ and $\boldsymbol{B}$, an inequality like $\boldsymbol{A} \geq \boldsymbol{B}$ should be read entrywise. Hence (27) denotes an array of $K \times K$ simultaneously holding inequalities.
13. Equality means that all K^2 inequalities are satisfied with equality.

7. Upper bounds on the privacy–fidelity trade-off

Consider the class of circulant mechanisms, which we shall denote as the set $\mathscr{W}_\circ$, and which contains all invertible matrices of the form

$$\boldsymbol{W} = \begin{bmatrix} w_1 & \dots & \dots & w_K \\ w_K & w_1 & \dots & w_{K-1} \\ \vdots & \ddots & \ddots & \vdots \\ w_2 & \dots & \dots & w_1 \end{bmatrix}. \tag{29}$$

Note that circulant mechanisms are fully described by their first row $\boldsymbol{w} = [w_1, \dots, w_K] \in \mathbb{P}$. If $w_k / w_{k'} \leq e^\epsilon$ for all (k, k'), then this matrix constitutes an ϵ-private mechanism and thus yields an upper (achievable) bound on the privacy–fidelity trade-off curve. We choose this class of mechanisms for producing upper bounds due to them

1. appearing as a natural choice, given that all columns have the same composition and thus the same maximum ratio $\max_{k,k'} w_k/w_{k'}$, thereby satisfying an intuitive notion (though not presently backed by a theorem) that for an optimal mechanism, the maximum intra-column ratio should be equal on all columns.

2. yielding simple expressions for the relevant fidelity loss metrics, with the added benefit of making all three normalized metrics $\alpha_{f\text{-DIV}}$, α_{MSE} and $\alpha_{f\text{TV}}$ match exactly (so long as the source is uniform), as we shall see in Lemma 8 stated below.

While a universal characterization of the privacy–fidelity trade-off in the context of local differential privacy seems elusive due to the fact that the trade-off critically depends on the specific choice of fidelity metric (f-divergence, MSE, TV, etc.), the following lemma, however, legitimizes $\varphi(\boldsymbol{W})$ as a useful quantity in that it "universally" captures the privacy-preserving properties of the mechanism $\boldsymbol{W}$, and thus may be subjected to optimization.

Lemma 8 *For a uniform source distribution $\boldsymbol{p} = \mathbf{1}/K$ and a circulant mechanism $\boldsymbol{W} \in \mathscr{W}_\circ$, the three normalized fidelity metrics are equal. That is, for every $\boldsymbol{W} \in \mathscr{W}_\circ$,*

$$\alpha_{f\text{-DIV}}\big(\tfrac{\mathbf{1}}{K}, \boldsymbol{W}\big) = \alpha_{\text{MSE}}\big(\tfrac{\mathbf{1}}{K}, \boldsymbol{W}\big) = \alpha_{\text{TV}}\big(\tfrac{\mathbf{1}}{K}, \boldsymbol{W}\big) = \frac{\varphi(\boldsymbol{W}) - 1}{K - 1}.$$

Proof The first equality was already mentioned previously (see (21)) and holds more generally for *any* (not necessarily circulant) mechanism $\boldsymbol{W}$. The second equality (for the TV metric) can be verified by direct evaluation. For this purpose, it suffices to inspect (20c) while realizing that for a circulant $\boldsymbol{W}$, we have $K\mathbf{1}\boldsymbol{\Phi} = \varphi(\boldsymbol{W})\mathbf{1}$, meaning that $\varphi(\boldsymbol{W})$ becomes the eigenvalue of $\boldsymbol{\Phi}(\boldsymbol{W})$ (up to a factor K) associated to the all-ones eigenvector. Alternatively, we can leverage the fact that all rows and columns of a circulant matrix have the same composition to infer that the Jensen inequality (22) becomes tight. The reader is also referred to detailed derivations for circulant mechanisms in the proof of Theorem 9, in Appendix H. ■

The following theorem determines the optimal mechanism within the class of circulant mechanisms, in the sense that it minimizes $\varphi(\boldsymbol{W})$.

Theorem 9 *The step mechanism $\boldsymbol{W}_{\epsilon,\star}$ as defined in* (23) *is optimal (up to row and column permutations) among all circulant ϵ-private mechanisms in terms of minimizing the quantity $\varphi(\boldsymbol{W})$, i.e.,*

$$\boldsymbol{W}_{\epsilon,\star} = \operatorname*{argmin}_{\boldsymbol{W}\in\mathscr{W}_\epsilon\cap\mathscr{W}_\circ} \varphi(\boldsymbol{W}). \tag{30}$$

Proof See Appendix H. The proof makes use of Fourier analysis and some well-known connections between the eigenvalues and the entries of circulant matrices. ■

By plugging the minimizer $\boldsymbol{W}_{\epsilon,\star}$ of the above problem (30) into the expressions (14)–(15), we obtain

$$\Phi_{k,\ell}(\boldsymbol{W}_{\epsilon,\star}) = \frac{1}{(e^\epsilon-1)^2}\Big[\big(e^\epsilon(e^\epsilon+K-2)+1-e^\epsilon\big)\delta_{k,\ell} + (e^\epsilon+K-2)(1-\delta_{k,\ell})\Big] \tag{31a}$$

$$\varphi(\boldsymbol{W}_{\epsilon,\star}) = \frac{K}{(e^\epsilon-1)^2}[(e^\epsilon+K-1)(e^\epsilon+K-2)+1-e^\epsilon]. \tag{31b}$$

The optimality property of the step mechanism elicited by Theorem 9 gives additional support to its being widely regarded as a natural choice of RR mechanism in any context. Some publications even use the term "randomized response" to refer to the step mechanism itself.

7.1 Upper bounds for the feasibility problem

Particularizing the mechanism $\boldsymbol{W}$ to being the step mechanism $\boldsymbol{W}_{\epsilon,\star}$, we can derive upper bounds on the fundamental privacy–fidelity trade-off curves $\alpha^\star(\epsilon;\boldsymbol{p})$ (for the feasibility problem), simply by inserting (31a) or (31b) into (8), in combination with loss metric expressions (20a)–(20c). With a uniform source, for example, using Lemma 8 we obtain, for any of the three fidelity loss metrics,

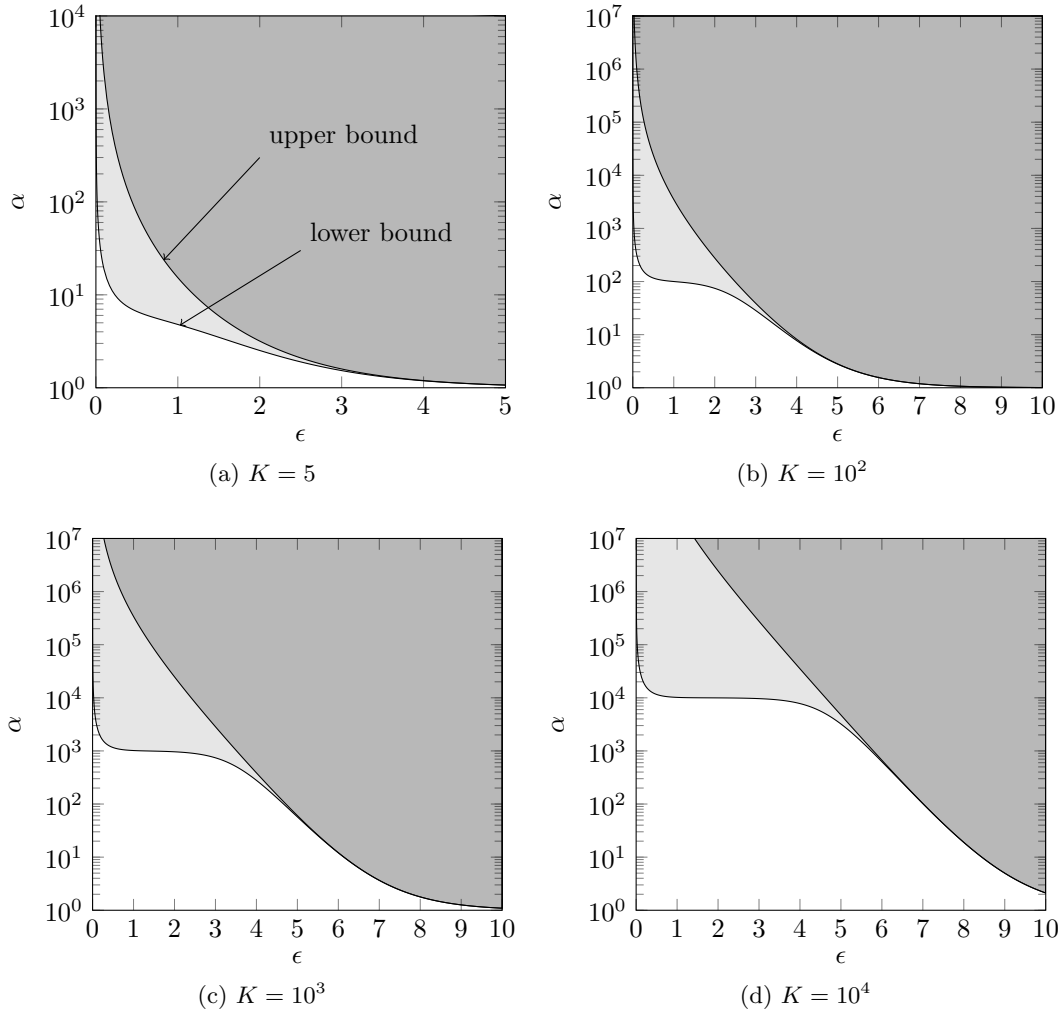
$$\alpha^\star\big(\epsilon;\tfrac{\mathbf{1}}{K}\big) \le \frac{\varphi(\boldsymbol{W}_{\epsilon,\star})-1}{K-1} \tag{32}$$

with $\varphi(\boldsymbol{W}_{\epsilon,\star})$ as given in (31b). The latter quantity is plotted as a function of ϵ in Figure 5 for different values of K, along with a corresponding lower bound derived in Section 8.

Furthermore, we believe that $\boldsymbol{W}_{\epsilon,\star}$ is the minimizer of $\varphi(\boldsymbol{W})$ not only within the class of circulant mechanisms, but over *all* ϵ-private mechanisms. That is, we conjecture that $\boldsymbol{W}_{\epsilon,\star} = \operatorname{argmin}_{\boldsymbol{W}\in\mathscr{W}_\epsilon}\varphi(\boldsymbol{W})$. As long as this conjecture remains unproven, we rely on complementing the upper bounds on privacy–fidelity tradeoffs yielded by the step mechanism, with the (generally non-matching) lower bounds developed in the next Section.

Note that (31a) can be equivalently written in matrix form as

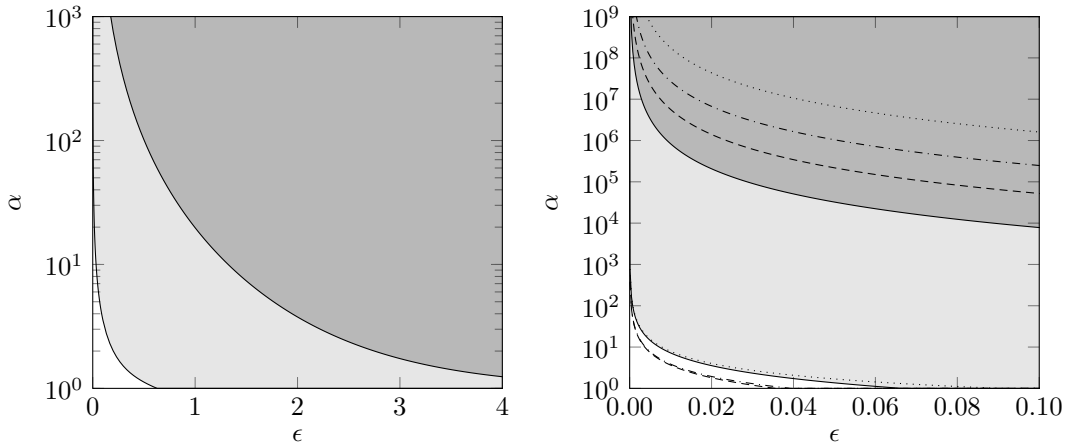
$$\boldsymbol{\Phi}(\boldsymbol{W}_{\epsilon,\star}) = \frac{1}{(e^\epsilon-1)^2}\Big[(e^\epsilon-1)(e^\epsilon+K-3)\mathbf{I} + (e^\epsilon+K-2)\mathbf{1}^{\mathrm{T}}\mathbf{1}\Big]. \tag{33}$$



(a) $K = 5$ (b) $K = 10^2$

(c) $K = 10^3$ (d) $K = 10^4$

Figure 5: Inner and outer bounds on the fundamental (ϵ, α)-trade-off for a uniform source $\boldsymbol{p} = \mathbf{1}/K$. The feasible set $\mathscr{F}(\mathbf{1}/K)$ contains the dark shaded region and is contained in the union of both shaded regions. In other terms, the lower boundary of $\mathscr{F}(\mathbf{1}/K)$, described by $\alpha^\star(\epsilon; \mathbf{1}/K)$, lies in the light shaded region. The upper bounds are based on (32) in combination with (31b). The lower bounds are derived in Section 8 further below.



(a) Zipf's law with $K = 5$

(b) An age pyramid with different resolutions ($K = 5, 10, 20, 50$)

Figure 6: Inner and outer bounds on the fundamental (ϵ, α)-trade-off for non-uniform sources: in (a) for Zipf's law $\boldsymbol{p} = \frac{60}{137}[1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}]$; in (b) for the age pyramid of the Brazilian population, drawn from the IPUMS data sets (Minnesota Population Center, 2019), similar to the data set used by Wang et al. (2019). For the dotted curves, the ages were grouped in $K = 50$ bins corresponding to age ranges 0–1, 2–3, ..., 99+ years; for the dashdotted curves, in $K = 20$ age bins 0–4, 5–9, ..., 95+ years; for the dashed curves, in $K = 10$ age bins 0–9, 10–19, ..., 90+ years; for the solid curves, in $K = 5$ age bins 0–19, 20–39, ..., 80+ years. The shaded region is painted for the latter case, with $K = 5$.

Plugging the latter into (20a)–(20c), we obtain the expressions

$$\alpha_{f\text{-DIV}}(\boldsymbol{p}, \boldsymbol{W}_{\epsilon,\star}) = \frac{\frac{1}{(e^\epsilon-1)^2}\big[(e^\epsilon-1)(e^\epsilon+K-3)K + (e^\epsilon+K-2)\big] - 1}{K-1}$$

$$\alpha_{\text{MSE}}(\boldsymbol{p}, \boldsymbol{W}_{\epsilon,\star}) = \frac{\frac{1}{(e^\epsilon-1)^2}\big[(e^\epsilon-1)(e^\epsilon+K-3) + (e^\epsilon+K-2)K\big] - \|\boldsymbol{p}\|_2^2}{1-\|\boldsymbol{p}\|_2^2}$$

$$\alpha_{\text{TV}}(\boldsymbol{p}, \boldsymbol{W}_{\epsilon,\star}) = \left(\frac{\sum_{k=1}^K \sqrt{\frac{1}{(e^\epsilon-1)^2}\big[(e^\epsilon-1)(e^\epsilon+K-3)p_k + (e^\epsilon+K-2)\big] - p_k^2}}{\sum_{k=1}^K \sqrt{p_k(1-p_k)}}\right)^2$$

which serve as upper bounds, respectively, on $\alpha^\star_{f\text{-DIV}}(\epsilon;\boldsymbol{p})$, $\alpha^\star_{\text{MSE}}(\epsilon;\boldsymbol{p})$ and $\alpha^\star_{\text{TV}}(\epsilon;\boldsymbol{p})$.

7.2 Upper bounds for the minimax problem

Recall that the fundamental minimax trade-off curve for a continuous set $\mathscr{P} \subseteq \mathbb{P}$ is given by [cf. (9)]

$$\alpha^\star(\epsilon;\mathscr{P}) \triangleq \min_{\boldsymbol{W}\in\mathscr{W}_\epsilon} \sup_{\boldsymbol{p}\in\mathscr{P}} \alpha(\boldsymbol{p},\boldsymbol{W}). \tag{35}$$

In the context of the minimax problem formulation, we shall focus on sets $\mathscr{P}$ that are symmetric in such way that all symbol probabilities are larger or equal to some $p_0 > 0$. That is, we set[14]

$$\mathscr{P} = \Big\{(p_1,\dots,p_K) \in [p_0,1]^K \colon \textstyle\sum_{k=1}^K p_k = 1\Big\}.$$

Since this set is closed, we can replace the supremum in (35) with a maximum.

Much like for the feasibility problem in Section 7.1, specializing $\boldsymbol{W}$ to the step mechanism $\boldsymbol{W}_{\epsilon,\star}$ allows us to derive upper bounds on the fundamental privacy–fidelity trade-off curves $\alpha^\star(\epsilon;\mathscr{P})$ for the minimax problem. We present these derivations one by one, in the following three Subsections 7.2.1 through 7.2.3.

7.2.1 f-divergence metric

In (35), let us set the f-divergence metric $\alpha_{f\text{-DIV}}$ and start by focusing on the inner supremization over $\boldsymbol{p} \in \mathscr{P}$. We need the following lemma.

Lemma 10 *Let $\boldsymbol{\Pi}_{\{i,j\}}$ denote the transposition of i and j, that is, the elementary permutation matrix that swaps the position of the i-th and j-th entries. Then the function*

$$\begin{aligned}[0,1] &\to \mathbb{R}_+\\ \lambda &\mapsto \alpha_{f\text{-DIV}}(\lambda\boldsymbol{p} + (1-\lambda)\boldsymbol{p}\boldsymbol{\Pi}_{\{i,j\}}, \boldsymbol{W})\end{aligned}$$

is convex.

Proof The proof is deferred to Appendix I. ∎

Using Lemma 10, we can infer that for any fixed $\boldsymbol{W}$, the maximizing source distribution

$$\boldsymbol{p}^\star(\boldsymbol{W}) = \operatorname*{argmax}_{\boldsymbol{p}\in\mathscr{P}} \alpha_{f\text{-DIV}}(\boldsymbol{p},\boldsymbol{W})$$

14. Note that p_0 must be smaller than $1/K$ for $\mathscr{P}$ to be non-empty.

has at most one entry different from p_0. In fact, if it had two entries $p_i^\star$ and $p_j^\star$ both distinct from (larger than) p_0, then one could argue with Lemma 10 that replacing the (i,j)-th entry pair of $\boldsymbol{p}^\star$ either by $(p_j^\star + p_i^\star - p_0, p_0)$ or by $(p_0, p_j^\star + p_i^\star - p_0)$ would yield a larger value of $\alpha_{f\text{-DIV}}(\boldsymbol{p}^\star, \boldsymbol{W})$, thus leading to a contradiction. We conclude that $\boldsymbol{p}^\star$ must have $K-1$ entries equal to p_0 and one entry equal to $1-(K-1)p_0$. Suppose that this entry is denoted by index k', then

$$\max_{\boldsymbol{p}\in\mathscr{P}} \boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W})\boldsymbol{p}^{-\mathrm{T}} = \Phi_{k',k'}(\boldsymbol{W}) + \sum_{k\neq k'}\sum_{\ell\neq k'} \Phi_{k,\ell}(\boldsymbol{W}) + \frac{p_0}{1-(K-1)p_0}\sum_{k\neq k'}\Phi_{k,k'}(\boldsymbol{W}) + \frac{1-(K-1)p_0}{p_0}\sum_{\ell\neq k'}\Phi_{k',\ell}(\boldsymbol{W}). \tag{36}$$

We now evaluate the latter for the step mechanism $\boldsymbol{W}_{\epsilon,\star}$. Using (33), the terms involving (sums of) entries of $\boldsymbol{\Phi}(\boldsymbol{W})$ can be expressed as

$$\begin{aligned}
\Phi_{k',k'}(\boldsymbol{W}_{\epsilon,\star}) &= \frac{1}{(e^\epsilon-1)^2}\big(e^\epsilon(e^\epsilon+K-2)-(e^\epsilon-1)\big)\\
\sum_{k\neq k'}\sum_{\ell\neq k'}\Phi_{k,\ell}(\boldsymbol{W}_{\epsilon,\star}) &= \frac{K-1}{(e^\epsilon-1)^2}\big((e^\epsilon+K-1)(e^\epsilon+K-2)-(e^\epsilon-1)\big)\\
\sum_{k\neq k'}\Phi_{k,k'}(\boldsymbol{W}_{\epsilon,\star}) &= \frac{K-1}{(e^\epsilon-1)^2}(e^\epsilon+K-2)\\
\sum_{\ell\neq k'}\Phi_{k',\ell}(\boldsymbol{W}_{\epsilon,\star}) &= \frac{K-1}{(e^\epsilon-1)^2}(e^\epsilon+K-2)
\end{aligned}$$

and upon inserting these expressions back into (36) and (20a), we end up with an upper bound on the f-divergence minimax fundamental privacy–fidelity tradeoff (9) given by

$$\alpha^\star_{f\text{-DIV}}(\epsilon;\mathscr{P}) \leq \frac{1}{K-1}\left(\frac{e^\epsilon+K-2}{(e^\epsilon-1)^2}\left[e^\epsilon+(K-1)(e^\epsilon+K-1) + \left(\frac{p_0}{1-(K-1)p_0}+\frac{1-(K-1)p_0}{p_0}\right)(K-1)\right] - \frac{K}{e^\epsilon-1}-1\right). \tag{38}$$

7.2.2 MSE metric

The MSE metric can be expressed as [cf. (20b)]

$$\alpha_{\mathrm{MSE}}(\boldsymbol{p},\boldsymbol{W}) = \tilde{\alpha}(\boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W})\mathbf{1}^{\mathrm{T}}, \|\boldsymbol{p}\|_2^2)$$

where $\tilde{\alpha}(u,v) \triangleq \frac{u-v}{1-v}$. On the domain $(u,v)\in[1,+\infty)\times[0,1)$, the function $\tilde{\alpha}$ is marginally non-decreasing in u (for fixed v) and in v (for fixed u). Therefore, one obtains an upper bound on $\max_{\boldsymbol{p}\in\mathscr{P}}\alpha_{\mathrm{MSE}}(\boldsymbol{p},\boldsymbol{W})$ by separately maximizing both arguments, namely

$$\max_{\boldsymbol{p}\in\mathscr{P}}\alpha_{\mathrm{MSE}}(\boldsymbol{p},\boldsymbol{W}) \leq \tilde{\alpha}\left(\max_{\boldsymbol{p}\in\mathscr{P}}\boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W})\mathbf{1}^{\mathrm{T}}, \max_{\boldsymbol{p}\in\mathscr{P}}\|\boldsymbol{p}\|_2^2\right). \tag{39}$$

For any fixed $\boldsymbol{W}$, the maximizing $\boldsymbol{p}$ in the first argument of the function $\tilde{\alpha}(\cdot,\cdot)$ on the right-hand side of (39) is

$$\begin{aligned}\boldsymbol{p}^\star(\boldsymbol{W}) &= \operatorname*{argmax}_{\boldsymbol{p}\in\mathscr{P}} \boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W})\mathbf{1}^{\mathrm{T}} \\ &= (1-(K-1)p_0)\boldsymbol{e}_{k^\star(\boldsymbol{W})} + p_0 \sum_{k\neq k^\star(\boldsymbol{W})} \boldsymbol{e}_k \end{aligned} \tag{40}$$

where $\boldsymbol{e}_k = (0,\ldots,0,1,0,\ldots,0)$ denotes the k-th canonical basis vector (with the k-th entry equal to one), and the index $k^\star(\boldsymbol{W})$ corresponds to the column of $\boldsymbol{\Phi}(\boldsymbol{W})$ with maximum column-sum:

$$k^\star(\boldsymbol{W}) = \operatorname*{argmax}_{k\in[K]} \boldsymbol{e}_k\boldsymbol{\Phi}(\boldsymbol{W})\mathbf{1}^{\mathrm{T}} = \operatorname*{argmax}_{k\in[K]} \sum_{k'=1}^{K} \Phi_{k,k'}(\boldsymbol{W}).$$

As to the second argument of $\tilde{\alpha}(\cdot,\cdot)$ in (39), its maximizer is any vector of the form (40) as well, yet with arbitrary $k^\star$. Hence, $\boldsymbol{p}^\star(\boldsymbol{W})$ as defined in (40) is a common maximizer for both arguments of $\tilde{\alpha}(\cdot,\cdot)$, and consequently, the upper bound (39) is tight, i.e.,

$$\max_{\boldsymbol{p}\in\mathscr{P}} \alpha_{\mathrm{MSE}}(\boldsymbol{p},\boldsymbol{W}) = \tilde{\alpha}\big(\boldsymbol{p}^\star(\boldsymbol{W})\boldsymbol{\Phi}(\boldsymbol{W})\mathbf{1}^{\mathrm{T}}, \|\boldsymbol{p}^\star(\boldsymbol{W})\|_2^2\big).$$

Here, the two arguments of the function $\tilde{\alpha}(\cdot,\cdot)$ can be evaluated in closed form, respectively, as

$$\|\boldsymbol{p}^\star(\boldsymbol{W})\|_2^2 = 1 - p_0(K-1)(2-Kp_0) \tag{41a}$$

$$\boldsymbol{p}^\star(\boldsymbol{W})\boldsymbol{\Phi}(\boldsymbol{W})\mathbf{1}^{\mathrm{T}} = p_0\varphi(\boldsymbol{W}) + (1-Kp_0)\sum_{\ell=1}^{K} \Phi_{k^\star(\boldsymbol{W}),\ell}(\boldsymbol{W}). \tag{41b}$$

Evaluating $\boldsymbol{\Phi}(\boldsymbol{W})$ [cf. (31a)] for the step mechanism $\boldsymbol{W}_{\epsilon,\star}$, we obtain for the summation term in (41b)

$$\sum_{\ell=1}^{K} \Phi_{k,\ell}(\boldsymbol{W}_{\epsilon,\star}) = \frac{(e^\epsilon+K-1)(e^\epsilon+K-2)}{(e^\epsilon-1)^2} - \frac{1}{e^\epsilon-1}$$

for any $k=1,\ldots,K$, based on which we can evaluate (41b) for the step mechanism $\boldsymbol{W}_{\epsilon,\star}$, namely

$$\boldsymbol{p}^\star(\boldsymbol{W}_{\epsilon,\star})\boldsymbol{\Phi}(\boldsymbol{W}_{\epsilon,\star})\mathbf{1}^{\mathrm{T}} = \frac{1}{(e^\epsilon-1)^2}\Big[(e^\epsilon+K-1)(e^\epsilon+K-2) - (e^\epsilon-1)\Big].$$

With this expression, we can compute the upper bound as

$$\begin{aligned}&\alpha^\star_{\mathrm{MSE}}(\epsilon;\mathscr{P}) \\ &\quad\le \tilde{\alpha}\bigg(\max_{\boldsymbol{p}\in\mathscr{P}} \boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W}_{\epsilon,\star})\mathbf{1}^{\mathrm{T}}, \max_{\boldsymbol{p}\in\mathscr{P}}\|\boldsymbol{p}\|_2^2\bigg) \\ &\quad= \frac{\frac{1}{(e^\epsilon-1)^2}\Big[(e^\epsilon+K-1)(e^\epsilon+K-2)-(e^\epsilon-1)\Big]-1}{p_0(K-1)(2-Kp_0)} + 1. \end{aligned} \tag{42}$$

7.2.3 TV metric

For the TV metric, we can derive an upper bound as follows:

$$\begin{aligned}
\alpha^\star_{\mathrm{TV}}(\epsilon;\mathscr{P}) &\leq \sup_{\boldsymbol{p}\in\mathscr{P}} \alpha_{\mathrm{TV}}(\boldsymbol{p},\boldsymbol{W}_{\epsilon,\star}) \\
&= \sup_{\boldsymbol{p}\in\mathscr{P}} \left(\frac{\langle \mathbf{1}, \sqrt{\boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W}_{\epsilon,\star}) - \boldsymbol{p}\odot\boldsymbol{p}}\rangle}{\langle \mathbf{1}, \sqrt{\boldsymbol{p}-\boldsymbol{p}\odot\boldsymbol{p}}\rangle} \right)^2 \\
&\overset{\text{(a)}}{\leq} \sup_{\boldsymbol{p}\in\mathscr{P}} \left(\frac{\sqrt{\varphi(\boldsymbol{W}_{\epsilon,\star}) - K\|\boldsymbol{p}\|_2^2}}{\langle \mathbf{1}, \sqrt{\boldsymbol{p}-\boldsymbol{p}\odot\boldsymbol{p}}\rangle} \right)^2 \\
&\overset{\text{(b)}}{\leq} \frac{\varphi(\boldsymbol{W}_{\epsilon,\star}) - K\min_{\boldsymbol{p}\in\mathscr{P}}\|\boldsymbol{p}\|_2^2}{\min_{\boldsymbol{p}\in\mathscr{P}}\langle \mathbf{1}, \sqrt{\boldsymbol{p}-\boldsymbol{p}\odot\boldsymbol{p}}\rangle^2} \\
&\overset{\text{(c)}}{=} \frac{\varphi(\boldsymbol{W}_{\epsilon,\star}) - 1}{\left((K-1)\sqrt{p_0(1-p_0)} + \sqrt{(1-(K-1)p_0)(K-1)p_0}\right)^2}.
\end{aligned} \tag{43}$$

Here, Step (a) relies on Jensen's inequality $\langle \mathbf{1}_K, \sqrt{\boldsymbol{x}}\rangle \leq \sqrt{K\langle \mathbf{1}, \boldsymbol{x}\rangle}$ and makes use of the fact that $\langle \mathbf{1}, \boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W}_{\epsilon,\star})\rangle = \varphi(\boldsymbol{W}_{\epsilon,\star})/K$; Step (b) consists in upper-bounding the maximum of a fraction by the ratio between the numerator's maximum and the denominator's minimum. For Step (c), the numerator is readily evaluated by noticing that the minimum of the convex symmetric function $\boldsymbol{p}\mapsto\|\boldsymbol{p}\|_2^2$ over the convex set $\mathscr{P}$ is achieved by the uniform distribution $\boldsymbol{p}=\mathbf{1}/K$, hence $\min_{\boldsymbol{p}\in\mathscr{P}}\|\boldsymbol{p}\|_2^2 = 1/K$. The denominator, in turn, is the minimum of a symmetric concave function $\boldsymbol{p}\mapsto\langle \mathbf{1}, \sqrt{\boldsymbol{p}-\boldsymbol{p}\odot\boldsymbol{p}}\rangle = \sum_{k=1}^K \sqrt{p_k(1-p_k)}$ over the symmetric simplex set $\mathscr{P}$, thus its minimizer is a vertex point of $\mathscr{P}$. The vertices of $\mathscr{P}$ are vectors with $K-1$ entries equal to p_0 and one entry equal to $1-(K-1)p_0$. Therefore, the denominator is (up to squaring) equal to

$$\min_{\boldsymbol{p}\in\mathscr{P}} \sum_{k=1}^K \sqrt{p_k(1-p_k)} = (K-1)\sqrt{p_0(1-p_0)} + \sqrt{(1-(K-1)p_0)(K-1)p_0}.$$

8. Lower bounds on the privacy–fidelity trade-off

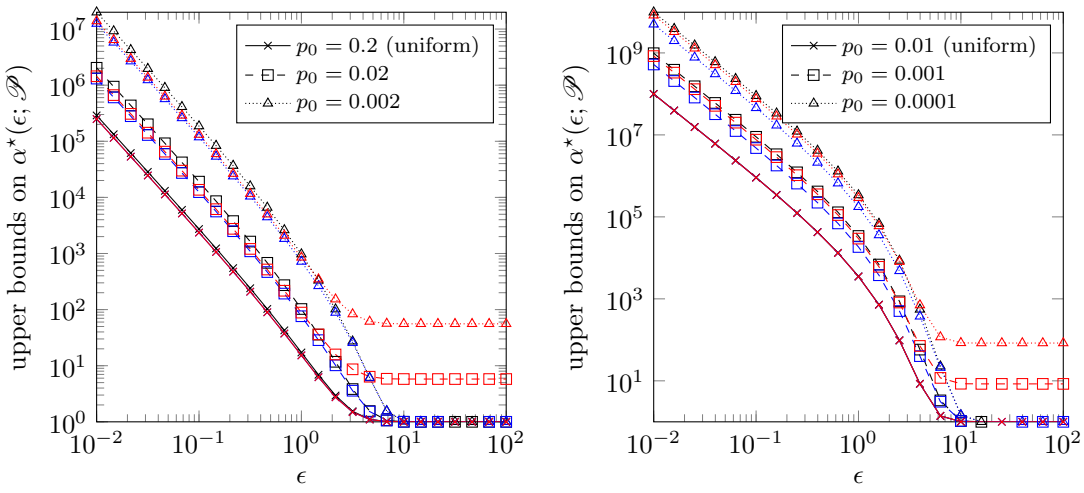
The lower bounds presented in this section rely on the following key lemma.

Lemma 11 *It holds that*

$$\min_{\boldsymbol{W}\in\mathscr{W}_\epsilon} \varphi(\boldsymbol{W}) \geq \frac{K}{1-e^{-4\epsilon}} \frac{(e^\epsilon + K - 1)^2}{e^{2\epsilon}+K-1} \triangleq \varphi_{\mathsf{LB}}(\epsilon;K). \tag{44}$$

Proof The proof is deferred to Appendix K. ■

Note that for ϵ fixed and K tending to infinity, the lower bound on the right-hand side of (44) behaves as $O(K^2)$, whereas the upper bound $\varphi(\boldsymbol{W}_{\epsilon,\star})$ as given by the right-hand side of (31b) behaves as $O(K^3)$. This gap is clearly visible as a difference in slope, in the limit of large K, between lower bounds (slope of 2 in logarithmic units) and upper bounds (slope of 3 in logarithmic units) in Figure 8. Closing this gap by sharpening either bound (upper or lower, or both) is an open problem.



(a) $K = 5$ (b) $K = 100$

(c) $K = 10^{10}$ (d) $K = 10^{20}$

Figure 7: Upper bounds (38), (42), (43) on the optimal privacy–fidelity minimax tradeoff curve $\alpha^\star(\epsilon;\mathscr{P})$ for different values of K and p_0. Curves corresponding to the KL divergence, MSE and TV metric are in black, blue and red, respectively.

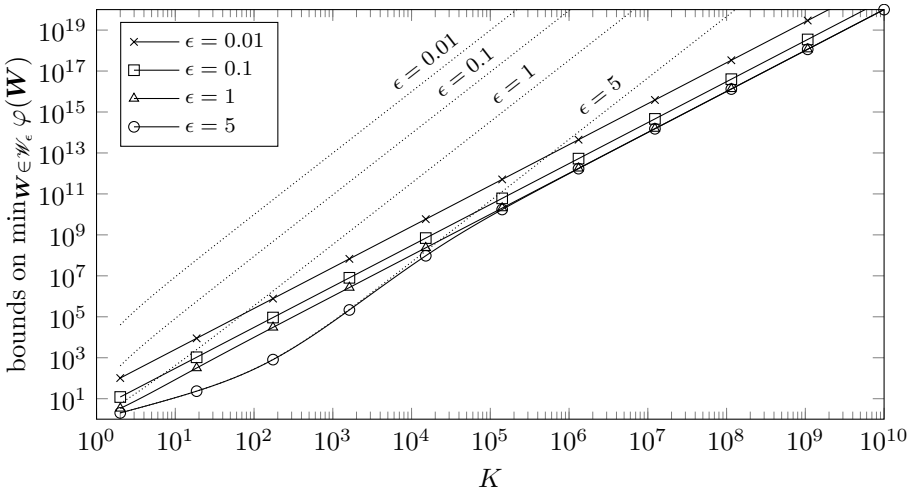


Figure 8: Upper bound $\varphi(\boldsymbol{W}_{\epsilon,\star})$ and lower bound $\varphi_{\mathsf{LB}}(\epsilon; K)$ on the quantity $\min_{\boldsymbol{W}\in\mathscr{W}_\epsilon} \varphi(\boldsymbol{W})$ as a function of K for different values of ϵ. Notice that, much like the upper bounds, the lower bounds each tend to a linear asymptote (although upper and lower bound asymptotes differ in slope). This is also true for $\epsilon = 5$ and $\epsilon = 10$, whose curves first follow their matching upper bound very closely, before eventually approaching their respective asymptote.

In the next subsections, we will present lower bounds for the feasibility problem and an instance of the minimax problem. In each case, we will need to address the three metrics (5a)–(5c) separately.

8.1 Lower bounds for the feasibility problem

In order to leverage Lemma 11 for lower-bounding the fundamental trade-off curve $\alpha^\star(\epsilon; \boldsymbol{p})$, we need to derive lower bounds on $\alpha^\star(\epsilon; \boldsymbol{p})$ that depend on $\boldsymbol{W}$ only via $\varphi(\boldsymbol{W})$. For the f-divergence and MSE metric, such bounds can be obtained via $p_{\min}\mathbf{1} \le \boldsymbol{p} \le p_{\max}\mathbf{1}$ and exploiting the fact that the metrics $\alpha_{f\text{-DIV}}$ and α_{MSE} are not smaller than one:

$$\alpha^\star_{f\text{-DIV}}(\epsilon; \boldsymbol{p}) \ge \min_{\boldsymbol{W}\in\mathscr{W}_\epsilon} \frac{\max\big\{K, \frac{p_{\min}}{p_{\max}}\varphi(\boldsymbol{W})\big\} - 1}{K-1} \tag{45a}$$

$$\alpha^\star_{\text{MSE}}(\epsilon; \boldsymbol{p}) \ge \min_{\boldsymbol{W}\in\mathscr{W}_\epsilon} \frac{\max\big\{1, p_{\min}\varphi(\boldsymbol{W})\big\} - \|\boldsymbol{p}\|_2^2}{1-\|\boldsymbol{p}\|_2^2} \tag{45b}$$

$$\alpha^\star_{\text{TV}}(\epsilon; \boldsymbol{p}) \ge \min_{\boldsymbol{W}\in\mathscr{W}_\epsilon} \left(\frac{\sqrt{p_{k_0}(1-p_{k_0}) + (\varphi(\boldsymbol{W})p_{\min} - 1)^+} + \sum_{k\in[K]\setminus\{k_0\}} \sqrt{p_k(1-p_k)}}{\sum_{k\in[K]} \sqrt{p_k(1-p_k)}} \right)^2 \tag{45c}$$

where $k_0 = \operatorname{argmin}_{k\in[K]} \big|p_k - \frac{1}{2}\big|$ and where $(\cdot)^+$ denotes $\max\{\cdot, 0\}$. While (45a) and (45b) are straightforward, deriving (45c) can be done by lower-bounding the numerator of (20c)

as follows:

$$\begin{aligned}\left\langle \mathbf{1}, \sqrt{\boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W}) - \boldsymbol{p}\odot\boldsymbol{p}}\right\rangle &\geq \left\langle \mathbf{1}, \sqrt{\max\{\boldsymbol{p}, p_{\min}\mathbf{1}\boldsymbol{\Phi}(\boldsymbol{W})\} - \boldsymbol{p}\odot\boldsymbol{p}}\right\rangle \\ &\overset{\text{(a)}}{\geq} \min_{\substack{\boldsymbol{\psi}\geq\mathbf{1}:\\ \langle\mathbf{1},\boldsymbol{\psi}\rangle=\varphi(\boldsymbol{W})}} \left\langle \mathbf{1}, \sqrt{\max\{\boldsymbol{p}, p_{\min}\boldsymbol{\psi}\} - \boldsymbol{p}\odot\boldsymbol{p}}\right\rangle \\ &= \begin{cases} \sqrt{\boldsymbol{p} - \boldsymbol{p}\odot\boldsymbol{p}} & \text{if } \varphi(\boldsymbol{W}) \leq \frac{1}{p_{\min}} \\ \min\limits_{\substack{\bar{\psi}_k\geq 0:\\ \sum_k \bar{\psi}_k = \varphi(\boldsymbol{W}) - \frac{1}{p_{\min}}}} \sum\limits_{k=1}^{K} \sqrt{p_k(1-p_k) + p_{\min}\bar{\psi}_k} & \text{if } \varphi(\boldsymbol{W}) > \frac{1}{p_{\min}}. \end{cases}\end{aligned}$$

For bounding step (a), we make use of the fact that $\mathbf{1}\boldsymbol{\Phi}(\boldsymbol{W}) \geq \mathbf{1}$, which is a consequence of $\boldsymbol{\Phi}(\boldsymbol{W}) \geq \mathbf{I}$. Note that (45a) and (45b) have the merit of being tight when $\boldsymbol{p} = \mathbf{1}/K$, whereas (45c) is generally not tight even for the uniform source distribution.

Now, by replacing $\varphi(\boldsymbol{W})$ on the right-hand sides of (45a)–(45c) with the lower bound $\varphi_{\mathsf{LB}}(\epsilon; K)$ from Lemma 11, we eventually obtain lower bounds on the fundamental trade-off curves $\alpha^\star(\epsilon; \boldsymbol{p})$ that only depend on K, ϵ and $\boldsymbol{p}$, but not on the mechanism $\boldsymbol{W}$. Namely,

$$\begin{aligned}\alpha^\star_{f\text{-DIV}}(\epsilon; \boldsymbol{p}) &\geq \frac{\max\left\{K, \frac{p_{\min}}{p_{\max}}\varphi_{\mathsf{LB}}(\epsilon; K)\right\} - 1}{K-1} \\ \alpha^\star_{\text{MSE}}(\epsilon; \boldsymbol{p}) &\geq \frac{\max\left\{1, p_{\min}\varphi_{\mathsf{LB}}(\epsilon; K)\right\} - \|\boldsymbol{p}\|_2^2}{1 - \|\boldsymbol{p}\|_2^2} \\ \alpha^\star_{\text{TV}}(\epsilon; \boldsymbol{p}) &\geq \left(\frac{\sqrt{p_{k_0}(1-p_{k_0}) + (\varphi_{\mathsf{LB}}(\epsilon; K)p_{\min} - 1)^+} + \sum_{k\neq k_0}\sqrt{p_k(1-p_k)}}{\sum_{k=1}^K \sqrt{p_k(1-p_k)}}\right)^2.\end{aligned}$$

Although the so-obtained bounds may be loose—especially when $\frac{p_{\min}}{p_{\max}} \ll 1$ or $p_{\min} \ll \frac{1}{K}$, that is, when $\boldsymbol{p}$ is highly non-uniform—they nonetheless highlight the relevance of the quantity $\varphi(\boldsymbol{W})$ as a proxy for characterizing the privacy–fidelity trade-off. This corroborates the relevance of lower bounds on $\varphi(\boldsymbol{W})$ such as the one given by Lemma 11.

8.2 Lower bounds for the minimax problem

In the next subsections, we will compute lower bounds on $\alpha^\star(\epsilon; \mathscr{P})$ for each of the three loss metrics. First off, note that a simple, robust[15] lower bound on $\alpha^\star(\epsilon; \mathscr{P})$ is obtained by setting $\boldsymbol{p} = \mathbf{1}/K$, i.e.,

$$\alpha^\star(\epsilon; \mathscr{P}) \geq \min_{\boldsymbol{W}\in\mathscr{W}_\epsilon} \alpha(\mathbf{1}/K, \boldsymbol{W}). \tag{47}$$

where the quantity $\alpha(\mathbf{1}/K, \boldsymbol{W})$ evaluates as in Remark 1 [cf. (21) and (22)]. It can be considered a robust bound in the sense that it does not depend on the value of p_0. We will leverage this for the f-divergence and TV metric (Subsections 8.2.1 and 8.2.3). By contrast, for the MSE metric (Subsection 8.2.2) we will derive a sharper lower bound that does depend on p_0.

15. It is robust in the sense that it does not depend on the parameter p_0.

8.2.1 f-divergence metric

Starting off with the simple lower bound (47) and using Lemma 11, a lower bound on $\alpha^\star_{f\text{-DIV}}(\epsilon;\mathscr{P})$ is given by

$$\alpha^\star_{f\text{-DIV}}(\epsilon;\mathscr{P}) \geq \frac{\varphi_{\mathsf{LB}}(\epsilon;K)-1}{K-1}. \tag{48}$$

A plot of the lower bound (48) is provided in Figure 9 further below.

8.2.2 MSE metric

The latter expression still depends on $\boldsymbol{W}$, so we further lower-bound it as follows:

$$\begin{aligned}
\boldsymbol{p}^\star(\boldsymbol{W})\boldsymbol{\Phi}(\boldsymbol{W})\mathbf{1}^{\mathrm{T}} &\geq \min_{\boldsymbol{W}\in\mathscr{W}_\epsilon} \boldsymbol{p}^\star(\boldsymbol{W})\boldsymbol{\Phi}(\boldsymbol{W})\mathbf{1}^{\mathrm{T}} \\
&= \min_{\boldsymbol{W}\in\mathscr{W}_\epsilon} \Big\{ p_0\varphi(\boldsymbol{W}) + (1-Kp_0)\sum_{\ell\in[K]} \Phi_{k^\star(\boldsymbol{W}),\ell}(\boldsymbol{W}) \Big\} \\
&\overset{\text{(a)}}{\geq} p_0 \min_{\boldsymbol{W}\in\mathscr{W}_\epsilon} \varphi(\boldsymbol{W}) + (1-Kp_0) \min_{\boldsymbol{W}\in\mathscr{W}_\epsilon} \sum_{\ell\in[K]} \Phi_{k^\star(\boldsymbol{W}),\ell}(\boldsymbol{W}) \\
&\overset{\text{(b)}}{\geq} \frac{1}{K} \min_{\boldsymbol{W}\in\mathscr{W}_\epsilon} \varphi(\boldsymbol{W}) \\
&\overset{\text{(c)}}{\geq} \frac{\varphi_{\mathsf{LB}}(\epsilon;K)}{K}.
\end{aligned} \tag{49}$$

Here, inequality (a) results from splitting the minimum of a sum into the sum of two minima; step (b) follows from lower-bounding the maximum column-sum of $\boldsymbol{\Phi}(\boldsymbol{W})$ by the average column-sum; step (c) is the application of Lemma 11. Combining (39), (40) and (49), we obtain

$$\alpha^\star_{\mathrm{MSE}}(\epsilon;\mathscr{P}) \geq \frac{\frac{\varphi_{\mathsf{LB}}(\epsilon;K)}{K} - 1 + p_0(K-1)(2-Kp_0)}{p_0(K-1)(2-Kp_0)}. \tag{50}$$

A plot of the lower bound (50) is provided in Figure 9 further below.

8.2.3 TV metric

For the TV metric, we will derive two alternative bounds: one based on the subadditivity of the square root function, the other based on Jensen's inequality and the concavity of the root function.

Bound 1 Since the square root is subadditive, we have for any non-negative vector $\boldsymbol{x}$ that $\langle \mathbf{1}, \sqrt{\boldsymbol{x}}\rangle \geq \sqrt{\langle \mathbf{1}, \boldsymbol{x}\rangle}$, which can be applied on the numerator of the TV metric expression to obtain the lower bound

$$\begin{aligned}
\alpha_{\mathrm{TV}}(\boldsymbol{p},\boldsymbol{W}) &= \frac{\big\langle \mathbf{1}, \sqrt{\boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W}) - \boldsymbol{p}\odot\boldsymbol{p}}\big\rangle^2}{\big\langle \mathbf{1}, \sqrt{\boldsymbol{p}-\boldsymbol{p}\odot\boldsymbol{p}}\big\rangle^2} \\
&\geq \frac{\langle \mathbf{1}, \boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W})\rangle - \|\boldsymbol{p}\|_2^2}{\big\langle \mathbf{1}, \sqrt{\boldsymbol{p}-\boldsymbol{p}\odot\boldsymbol{p}}\big\rangle^2}.
\end{aligned}$$

Upon inserting the uniform source distribution $\boldsymbol{p} = \mathbf{1}/K$, we get

$$\alpha_{\mathrm{TV}}(\mathbf{1}/K, \boldsymbol{W}) \geq \frac{1}{K} \frac{\varphi(\boldsymbol{W}) - 1}{K - 1}.$$

So finally,

$$\alpha^{\star}_{\mathrm{TV}}(\epsilon; \mathscr{P}) \geq \frac{1}{K} \frac{\varphi_{\mathsf{LB}}(\epsilon; K) - 1}{K - 1}. \tag{51}$$

Since $\varphi_{\mathsf{LB}}(\epsilon; K)$ can be arbitrarily close to K (as $\epsilon \to \infty$), the bound (51) can be improved by noticing that $\alpha^{\star}_{\mathrm{TV}}(\epsilon; \mathscr{P})$, by the data processing inequality, is never smaller than one.

Bound 2 We start with the basic lower bound (47) in combination with (22), i.e.,

$$\begin{aligned} \alpha^{\star}_{\mathrm{TV}}(\epsilon; \mathscr{P}) &\geq \min_{\boldsymbol{W} \in \mathscr{W}_\epsilon} \alpha_{\mathrm{TV}}\big(\tfrac{\mathbf{1}}{K}, \boldsymbol{W}\big) \\ &= \min_{\boldsymbol{W} \in \mathscr{W}_\epsilon} \left(\frac{\sum_k \sqrt{K \sum_\ell \Phi_{k,\ell}(\boldsymbol{W}) - 1}}{K\sqrt{K-1}} \right)^2. \end{aligned} \tag{52}$$

The next aim is to lower-bound the right-hand side of (52) so as to obtain a bound that only depends on $\boldsymbol{W}$ via $\varphi(\boldsymbol{W})$. For this purpose, observe that the function

$$g\colon \quad [1/K, +\infty) \to \mathbb{R}_+, \quad x \mapsto \sqrt{Kx - 1}$$

is increasing and concave. By the non-negativity of $\boldsymbol{\Phi}(\boldsymbol{W})$ and the data-processing inequality in matrix form [Theorem 7 particularized to (28)] we have $\mathbf{0} \leq \boldsymbol{\Phi} \leq \mathbf{I}$, whence we infer that the sum $\sum_\ell \Phi_{k,\ell}(\boldsymbol{W})$ is comprised between 1 and $\varphi(\boldsymbol{W})$. On the other hand, the restriction of g to the interval $[1, \varphi(\boldsymbol{W})]$ is lower-bounded by the corresponding secant:

$$\begin{aligned} g(x) &\geq g(1) + \frac{g(\varphi(\boldsymbol{W})) - g(1)}{\varphi(\boldsymbol{W}) - 1}(x - 1) \\ &\geq g(1) + g'(\varphi(\boldsymbol{W}))(x - 1) \\ &= g(1) + \frac{K}{2\sqrt{K\varphi(\boldsymbol{W}) - 1}}(x - 1), \quad (\text{for } 1 \leq x \leq \varphi(\boldsymbol{W})). \end{aligned}$$

(Here, in the last inequality we use the concavity of g to obtain a weaker but simpler lower bound on the secant slope.) Applying this lower bound on the right-hand side of (52), we obtain

$$\begin{aligned} \alpha^{\star}_{\mathrm{TV}}(\epsilon; \mathscr{P}) &\geq \min_{\boldsymbol{W} \in \mathscr{W}_\epsilon} \left(\frac{g(1) + \frac{1}{2\sqrt{K\varphi(\boldsymbol{W}) - 1}}(\varphi(\boldsymbol{W}) - K)}{\sqrt{K-1}} \right)^2 \\ &\geq \frac{1}{K-1} \left(\sqrt{K-1} + \frac{\varphi_{\mathsf{LB}}(\epsilon; K) - K}{2\sqrt{K\varphi_{\mathsf{LB}}(\epsilon; K) - 1}} \right)^2. \end{aligned} \tag{53}$$

The last bounding step can be made owing to the monotonicity of $\mathbb{R}_+ \mapsto \mathbb{R}_+, \varphi \mapsto (\varphi - K)/\sqrt{K\varphi - 1}$, which holds for all $K \geq 2$.

Combining Bound 1 and Bound 2 Notice that, unlike the first lower bound (51), this second bound (53) is always lower-bounded by unity. Our final lower bound on the fundamental minimax tradeoff curve for the TV metric is obtained by combining (51) and (53):

$$\alpha^{\star}_{\mathrm{TV}}(\epsilon;\mathscr{P}) \geq \frac{1}{K-1}\max\left\{\frac{\varphi_{\mathsf{LB}}(\epsilon;K)-1}{K}, \left(\sqrt{K-1}+\frac{\varphi_{\mathsf{LB}}(\epsilon;K)-K}{2\sqrt{K\varphi_{\mathsf{LB}}(\epsilon;K)-1}}\right)^2\right\}. \tag{54}$$

Plots of the above lower bounds (48), (50), (51) and (53) are provided in Figure 9.

9. Conclusion

We have proposed a framework for the study of privacy–fidelity trade-off problems in the context of randomized response mechanisms, in which the privatization channel is supposed to facilitate the estimation of the unknown source distribution while obfuscating the source realizations. A privacy metric based on the concept of local differential privacy, and fidelity loss metrics based on f-divergence, MSE and TV distance have been proposed as figures of merit. We have identified the quantities $\boldsymbol{\Phi}(\boldsymbol{W})$ and $\varphi(\boldsymbol{W})$, which capture the essence of the dependency of fidelity loss metrics on the random mechanism $\boldsymbol{W}$, and studied some of its properties, including data-processing inequalities. Finally, we have derived inner and outer bounds to some specific instances of the fundamental privacy–fidelity trade-off curve, all of which depend on the random mechanism via $\boldsymbol{\Phi}(\boldsymbol{W})$ or $\varphi(\boldsymbol{W})$.

For a better understanding of the fundamental privacy–fidelity trade-off problems, it would be desirable to tighten the gap between inner and outer bounds much further. There is some indication that the lower bounds are loose, so that the step mechanism $\boldsymbol{W}_{\epsilon,\star}$ stands as an optimality candidate among all ϵ-private mechanisms (in terms of minimizing $\varphi(\boldsymbol{W})$). A proof or counterexample to this claim is left as an open problem. Other interesting research directions include, for instance, broad channels ($L \geq K$) (which encompass the RAPPOR mechanism and its generalization by Ye and Barg (2018)), sources and/or privatization channels with memory, interactive mechanisms and batch processing, extensions to other types of statistical tests or queries (beyond distribution estimation), and to fidelity loss metrics based on tail probabilities rather than expectations.

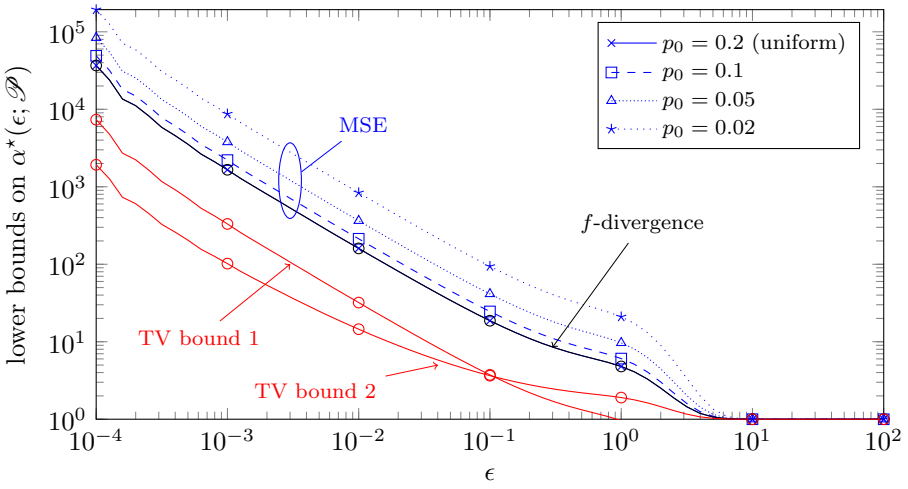
Acknowledgments

This work has been supported by the Catalan Government under grant 2017 SGR 1479, by the Spanish Ministery of Economy and Competitiveness through project RTI2018-099722-B-I00 (ARISTIDES), and in part by the European ERC Starting Grant 259530-ComCom.

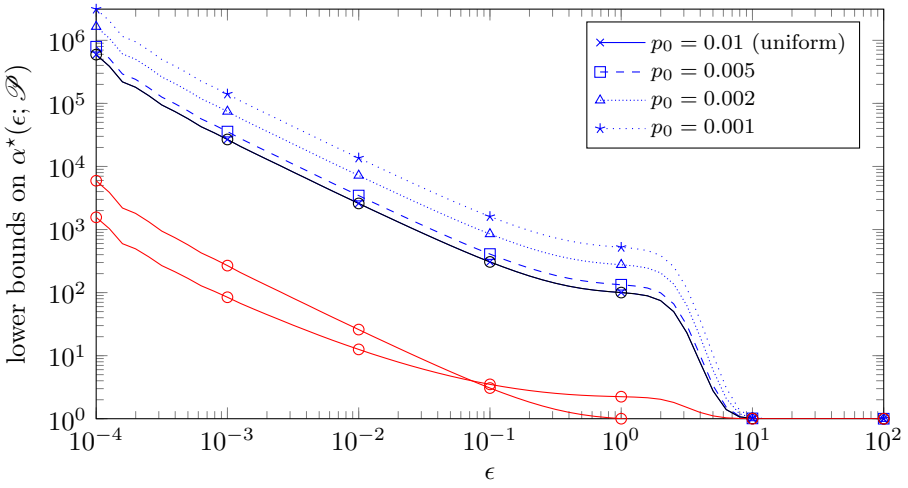
Appendix A. Proof of Lemma 1

By Csiszár's concentration inequality (Csiszár, 1984, Theorem 1),

$$\begin{aligned}\mathsf{Pr}\big\{\boldsymbol{t}(\boldsymbol{y}_n)\boldsymbol{W}^{-1} \notin \mathbb{P}\big\} &= \mathsf{Pr}\big\{\boldsymbol{t}(\boldsymbol{y}_n) \notin \mathbb{P}\boldsymbol{W}\big\} \\ &= \mathsf{Pr}\big\{\boldsymbol{t}(\boldsymbol{y}_n) \in \mathbb{P} \setminus \mathbb{P}\boldsymbol{W}\big\} \\ &\leq e^{-nD(\mathbb{P}\setminus\mathbb{P}\boldsymbol{W}\|\boldsymbol{p}\boldsymbol{W})}.\end{aligned}$$



(a) $K = 5$



(b) $K = 100$

Figure 9: Lower bounds (48) [f-divergence metric, in black], (50) [MSE metric, in blue] as well as (51) and (53) [TV metric, in red, which can be combined to (54)] on the optimal privacy–fidelity minimax tradeoff curve $\sup_{\boldsymbol{p}\in\mathscr{P}} \alpha^\star(\boldsymbol{p}, \boldsymbol{W}_{\epsilon,\star})$ for different values of K and p_0.

Since $\boldsymbol{pW}$ has only positive entries, we can express $D(\mathbb{P} \setminus \mathbb{P}\boldsymbol{W}\|\boldsymbol{pW})$ as a minimum over the topological closure of $\mathbb{P} \setminus \mathbb{P}\boldsymbol{W}$ (instead of an infimum):

$$D(\mathbb{P} \setminus \mathbb{P}\boldsymbol{W}\|\boldsymbol{pW}) = \min_{\boldsymbol{r}' \in \overline{\mathbb{P} \setminus \mathbb{P}\boldsymbol{W}}} D(\boldsymbol{r}'\|\boldsymbol{pW}).$$

Furthermore, since $\boldsymbol{pW}$ belongs to the compact set $\mathbb{P}\boldsymbol{W}$, the minimizing distribution is located on its boundary $\partial\mathbb{P}\boldsymbol{W}$. In fact, for any distribution $\boldsymbol{r}' \in \overline{\mathbb{P} \setminus \mathbb{P}\boldsymbol{W}}$, there exists a distribution $\boldsymbol{r}''$ on the intersection between the boundary $\partial\mathbb{P}\boldsymbol{W}$ and the segment connecting $\boldsymbol{r}'$ and $\boldsymbol{pW}$, such that $D(\boldsymbol{r}''\|\boldsymbol{pW}) \leq D(\boldsymbol{r}'\|\boldsymbol{pW})$. Hence,

$$\begin{aligned} D(\mathbb{P} \setminus \mathbb{P}\boldsymbol{W}\|\boldsymbol{pW}) &= D(\partial\mathbb{P}\boldsymbol{W}\|\boldsymbol{pW}) \\ &= \min_{\boldsymbol{r} \in \partial\mathbb{P}} D(\boldsymbol{rW}\|\boldsymbol{pW}). \end{aligned} \tag{55}$$

In addition, observing that

$$\partial\mathbb{P} = \big\{\boldsymbol{p}' \in \mathbb{P}\colon p'_i = 0 \text{ for some } i\big\}$$

it becomes evident from the last line of (55) that $D(\partial\mathbb{P}\boldsymbol{W}\|\boldsymbol{pW}) > 0$, since $\boldsymbol{rW} \neq \boldsymbol{pW}$ for all $\boldsymbol{r}$ in the compact set $\partial\mathbb{P}$.

Appendix B. Proof of Theorem 2

Let us define the random variables [cf. (10)]

$$\begin{aligned} \hat{\boldsymbol{q}}_n &= [\hat{q}_{n,1}, \ldots, \hat{q}_{n,K}] = \boldsymbol{t}(\boldsymbol{y}_n) \\ \check{\boldsymbol{p}}_n &= [\check{p}_{n,1}, \ldots, \check{p}_{n,K}] = \boldsymbol{t}(\boldsymbol{y}_n)\boldsymbol{W}^{-1} \\ \hat{\boldsymbol{p}}_n &= [\hat{p}_{n,1}, \ldots, \hat{p}_{n,K}] = \mathrm{Proj}_{\mathbb{P}}(\boldsymbol{t}(\boldsymbol{y}_n)\boldsymbol{W}^{-1}). \end{aligned}$$

The histogram $n\hat{\boldsymbol{q}}_n$ is a K-variate random variable which follows the multinomial distribution with support set

$$\mathscr{S}_n \triangleq \big\{(n_1, \ldots, n_K) \in \mathbb{N}^K\colon n_1 + \ldots + n_K = n\big\}$$

(where $\mathbb{N}$ are the non-negative integers) and with a conditional probability mass function given by

$$\mathsf{Pr}\big\{n\hat{\boldsymbol{q}}_n = (n_1, \ldots, n_K)\big\} = \begin{cases} \frac{n!}{n_1!\ldots n_K!} q_1^{n_1} \ldots q_K^{n_K} & \text{for } (n_1, \ldots, n_K) \in \mathscr{S}_n \\ 0 & \text{for } (n_1, \ldots, n_K) \notin \mathscr{S}_n. \end{cases}$$

By assumption,

$$f(x) = \sum_{\rho=1}^{4} \frac{f^{(\rho)}(1)}{\rho!}(x-1)^\rho + \varrho\big(|x-1|^{4+\gamma}\big) \tag{56}$$

where the remainder function $\varrho(x)$ satisfies that $\limsup_{x\to 0} |\varrho(x)|/x$ is finite. By combining the definition of f-divergence (5a) with the above Taylor expansion, we obtain

$$\mathsf{E}\big[D_f\big(\hat{\boldsymbol{p}}_n\big\|\boldsymbol{p}\big)\big] = \sum_{k=1}^{K} p_k \left(\sum_{\rho=1}^{4} \frac{f^{(\rho)}(1)}{\rho!} \frac{\hat{\mu}_{n,k}^{(\rho)}}{n^\rho p_k^\rho} + R_{n,k} \right)$$

where

$$\hat{\mu}_{n,k}^{(\rho)} \triangleq n^{\rho}\,\mathsf{E}\big[\big(\hat{p}_{n,k}-p_k\big)^{\rho}\big]$$
$$R_{n,k} \triangleq \mathsf{E}\bigg[\varrho\Big(\Big|\tfrac{\hat{p}_{n,k}}{p_k}-1\Big|^{4+\gamma}\Big)\bigg].$$

Let $\tilde{\boldsymbol{w}}_k = [\tilde{W}_{1,k},\dots,\tilde{W}_{K,k}]$ denote the k-th column of $\boldsymbol{W}^{-1}$ (transposed into a row vector) and let us define the ρ-th central moment of $n\langle\tilde{\boldsymbol{w}}_k,\hat{\boldsymbol{q}}_n\rangle$ as

$$\begin{aligned}\check{\mu}_{n,k}^{(\rho)} &\triangleq n^{\rho}\,\mathsf{E}\big[\big\langle\tilde{\boldsymbol{w}}_k,\hat{\boldsymbol{q}}_n-\boldsymbol{q}\big\rangle^{\rho}\big]\\ &= n^{\rho}\,\mathsf{E}\big[\big(\check{p}_{n,k}-p_k\big)^{\rho}\big].\end{aligned}\tag{57}$$

Lemma 12 *The difference between $\hat{\mu}_{n,k}^{(\rho)}$ and $\check{\mu}_{n,k}^{(\rho)}$ decays at least exponentially in n, in that we can upper-bound its absolute value as follows:*

$$\Big|\hat{\mu}_{n,k}^{(\rho)}-\check{\mu}_{n,k}^{(\rho)}\Big| \le Cn^{\rho}e^{-nD(\partial\mathbb{P}_{\boldsymbol{W}}\|\boldsymbol{p}_{\boldsymbol{W}})}$$

for some positive constant $C>0$.[16]

Proof See Appendix C. ∎

Lemma 12 is a consequence of $\hat{\boldsymbol{p}}_n$ and $\check{\boldsymbol{p}}_n$ being equal except in exponentially rare cases. This allows us to exchange $\hat{\mu}_{n,k}^{(\rho)}$ for the easier-to-analyze $\check{\mu}_{n,k}^{(\rho)}$ in the study of asymptotic expansions. Next, we will turn our attention to evaluating the ρ-th moment $\check{\mu}_{n,k}^{(\rho)}$.

The multivariate moment-generating function of the multinomially distributed $n(\hat{\boldsymbol{q}}_n-\boldsymbol{q})$ being

$$\mathscr{M}(\boldsymbol{\lambda}) = \mathsf{E}\Big[e^{n\langle\boldsymbol{\lambda},\hat{\boldsymbol{q}}_n-\boldsymbol{q}\rangle}\Big] = \left(\sum_{\ell=1}^{K} q_\ell e^{\lambda_\ell}\right)^{n} e^{-n\langle\boldsymbol{\lambda},\boldsymbol{q}\rangle}\tag{58}$$

with $\boldsymbol{\lambda}\in\mathbb{R}^K$, it is immediate to obtain the moment-generating function of a weighted sum of the variates of $n(\hat{\boldsymbol{q}}_n-\boldsymbol{q})$. For a weight vector $\tilde{\boldsymbol{w}}_k$, it suffices to replace $\boldsymbol{\lambda}$ with $\lambda\tilde{\boldsymbol{w}}_k$ in (58) to obtain the moment-generating function of $n\langle\tilde{\boldsymbol{w}}_k,\hat{\boldsymbol{q}}_n-\boldsymbol{q}\rangle = n(\check{p}_{n,k}-p_k)$, i.e.,

$$\begin{aligned}\mathscr{M}_k(\lambda) &\triangleq \mathscr{M}(\lambda\tilde{\boldsymbol{w}}_k)\\ &= \left(\sum_{\ell=1}^{K} q_\ell e^{\lambda\tilde{W}_{\ell,k}}\right)^{n} e^{-\lambda n\langle\tilde{\boldsymbol{w}}_k,\boldsymbol{q}\rangle}\end{aligned}\tag{59}$$

with scalar argument $\lambda\in\mathbb{R}$. The ρ-th central moment of $n(\check{p}_{n,k}-p_k)$ can now be expressed as the ρ-th derivative at $\lambda=0$ of the corresponding moment-generating function:

$$\check{\mu}_{n,k}^{(\rho)} = \mathscr{M}_k^{(\rho)}(0).\tag{60}$$

16. For an explicit bound on C, see (76).

The ρ-th derivative of $\mathscr{M}_k$ can be expressed as

$$\begin{aligned}\mathscr{M}_k^{(\rho)}(\lambda) &= \sum_{i=0}^{\rho} \binom{\rho}{i} f_k^{(i)}(\lambda) g_k^{(\rho-i)}(\lambda) \\ &= f_k(\lambda) g_k^{(\rho)}(\lambda) + \sum_{i=1}^{\rho} \binom{\rho}{i} f_k^{(i)}(\lambda) g_k^{(\rho-i)}(\lambda) \end{aligned} \tag{61}$$

where the functions f_k and g are defined as

$$\begin{aligned} f_k(\lambda) &\triangleq \left(\sum_{\ell=1}^{K} q_\ell e^{\lambda \tilde{W}_{\ell,k}} \right)^n \\ g_k(\lambda) &\triangleq e^{-\lambda n \langle \tilde{\boldsymbol{w}}_k, \boldsymbol{q} \rangle}. \end{aligned}$$

We now seek to derive an explicit expression for $\mathscr{M}_k^{(\rho)}(\lambda)$. The derivative $g_k^{(\rho-i)}(\lambda)$ can be easily evaluated as

$$g_k^{(\rho-i)}(\lambda) = (-n\langle \tilde{\boldsymbol{w}}_k, \boldsymbol{q} \rangle)^{\rho-i} g_k(\lambda).$$

As to the derivative $f_k^{(i)}$, it can be evaluated for $i \geq 1$ using Faà di Bruno's formula for derivatives of concatenated functions:

$$\begin{aligned} f_k^{(i)}(\lambda) &= \frac{\mathrm{d}^i}{\mathrm{d}\lambda^i}(u \circ v_k)(\lambda) \\ &= \widetilde{\sum} \frac{i!}{m_1! \dots m_i!} u^{(m_1+\ldots+m_i)}(v_k(\lambda)) \prod_{j=1}^{i} \left(\frac{v_k^{(j)}(\lambda)}{j!} \right)^{m_j} \end{aligned}$$

where the summation $\widetilde{\sum}$, denoted with a tilde, is over all tuples of non-negative $(m_1, \dots, m_i) \in \mathbb{N}^i$ satisfying $1 \cdot m_1 + \ldots + i \cdot m_i = i$. The functions u and v_k are respectively defined as

$$\begin{aligned} u(\lambda) &= \lambda^n \\ v_k(\lambda) &= \sum_{\ell=1}^{K} q_\ell e^{\lambda \tilde{W}_{\ell,k}} \end{aligned}$$

and have respective derivatives

$$\begin{aligned} u^{(j)}(\lambda) &= \begin{cases} \frac{n!}{(n-j)!} \lambda^{n-j} & \text{if } j \leq n \\ 0 & \text{if } j > n \end{cases} \\ v_k^{(j)}(\lambda) &= \sum_{\ell=1}^{K} q_\ell \tilde{W}_{\ell,k}^j e^{\lambda \tilde{W}_{\ell,k}}. \end{aligned}$$

Noticing that $v_k(0) = 1$ and assuming that n is sufficiently large so as to ensure that $m_1 + \ldots + m_i \leq n$ for all values of the sum $m_1 + \ldots + m_i$ taken by summation indices of $\widetilde{\sum}$, we can evaluate the derivative $f_k^{(i)}(0)$ as follows:

$$f_k^{(i)}(0) = \widetilde{\sum} \frac{i!}{m_1! \dots m_i!} \frac{n!}{(n - m_1 - \ldots - m_i)!} \prod_{j=1}^{i} \frac{\left\langle \tilde{\boldsymbol{w}}_k^{\odot j}, \boldsymbol{q} \right\rangle^{m_j}}{j!^{m_j}} \tag{62}$$

where the superscript notation $\tilde{\boldsymbol{w}}_k^{\odot j} \triangleq (\tilde{W}_{1,k}^j, \dots, \tilde{W}_{K,k}^j)$ denotes entrywise exponentiation. Note that the inner product $\langle \tilde{\boldsymbol{w}}_k^{\odot j}, \boldsymbol{q} \rangle$ appearing in the last expression is in fact nothing else than [cf. (13)]

$$\begin{aligned}\langle \tilde{\boldsymbol{w}}_k^{\odot j}, \boldsymbol{q} \rangle &= \boldsymbol{p}\boldsymbol{W}\big(\underbrace{\boldsymbol{W}^{-1} \odot \boldsymbol{W}^{-1} \odot \ldots \odot \boldsymbol{W}^{-1}}_{j \text{ factors}}\big)\boldsymbol{e}_k^{\mathrm{T}} \\ &= \nu_{j,k}.\end{aligned}$$

Combining (62) with (61) and noticing that $f_k(0) = g_k(0) = 1$, we obtain

$$\begin{aligned}\mathscr{M}_k^{(\rho)}(0) &= (-n\nu_{1,k})^\rho + \sum_{i=1}^{\rho} \binom{\rho}{i} f_k^{(i)}(0)(-n\nu_{1,k})^{\rho-i} \\ &= (-n\nu_{1,k})^\rho + \sum_{i=1}^{\rho} \binom{\rho}{i} \widetilde{\sum} \prod_{j=1}^{i} \frac{\nu_{j,k}^{m_j}}{j!^{m_j}} \frac{i!n!(-n\nu_{1,k})^{\rho-i}}{m_1! \ldots m_i!(n - m_1 - \ldots - m_i)!}. \end{aligned} \tag{63}$$

The second, third and fourth moments of the centered variable $n(\check{p}_{n,k} - p_k)$ can be evaluated from (60) and (63) as

$$\check{\mu}_{n,k}^{(2)} = n\big(\nu_{2,k} - \nu_{1,k}^2\big) \tag{64a}$$

$$\check{\mu}_{n,k}^{(3)} = n\big(2\nu_{1,k}^3 - 3\nu_{1,k}\nu_{2,k} + \nu_{3,k}\big) \tag{64b}$$

$$\check{\mu}_{n,k}^{(4)} = n\big((3n-6)\nu_{1,k}^4 + 3(n-1)\nu_{2,k}^2 + (12-6n)\nu_{1,k}^2\nu_{2,k} - 4\nu_{1,k}\nu_{3,k} + 3\nu_{4,k}\big). \tag{64c}$$

These explicit evaluations may now be inserted in the Taylor expansion (56). In the last part, what remains to be proven is that the remainder term of said Taylor expansion satisfies

$$\lim_{n\to\infty} n^2 R_{n,k} = 0. \tag{65}$$

For this purpose we upper-bound the absolute value of $R_{n,k}$ as follows:

$$\begin{aligned}|R_{n,k}| \leq{}& \mathsf{E}\left[\left|\varrho\Big(\left|\tfrac{\hat{p}_{n,k}}{p_k} - 1\right|^{4+\gamma}\Big)\right| \,\middle|\, \left|\tfrac{\hat{p}_{n,k}}{p_k} - 1\right| < \delta\right] \\ &+ \mathsf{E}\left[\left|\varrho\Big(\left|\tfrac{\hat{p}_{n,k}}{p_k} - 1\right|^{4+\gamma}\Big)\right| \,\middle|\, \left|\tfrac{\hat{p}_{n,k}}{p_k} - 1\right| \geq \delta\right] \mathsf{Pr}\Big\{\left|\tfrac{\hat{p}_{n,k}}{p_k} - 1\right| \geq \delta\Big\}. \end{aligned} \tag{66}$$

This bound results from the convexity of the absolute value (Jensen's inequality), from expanding the expectation using the law of total expectation, and upper-bounding $\mathsf{Pr}\{|\hat{p}_{n,k}/p_k - 1| < \delta\}$ by one. The remaining three terms on the right-hand side of (66) are upper-bounded in the following.

For the first term, recall that $\varrho(x) = O(x)$ in the vicinity of $x = 0$, hence there exists a value $\omega > 0$ such that for any sufficiently small $\delta > 0$, we have $|\varrho(\delta)| \leq \omega\delta$. Consequently, there exists a $\delta_0 > 0$ such that, so long as $\delta \in (0, \delta_0)$,

$$\mathsf{E}\left[\left|\varrho\Big(\left|\tfrac{\hat{p}_{n,k}}{p_k} - 1\right|^{4+\gamma}\Big)\right| \,\middle|\, \left|\tfrac{\hat{p}_{n,k}}{p_k} - 1\right| < \delta\right] \leq \omega\delta^{4+\gamma}. \tag{67}$$

For the second term in (66), recall that $f(0)$ is bounded and $f(x)$ is well-defined for $x > 0$, and as a consequence, $\varrho(x)$ must be well-defined for $x \geq 0$. In particular, it follows that

$\varrho(|x-1|^{4+\gamma})$ is bounded on the closed interval $x \in [0, 1/p_k]$. Therefore, the second term in (66) is upper-bounded by a constant

$$\mathsf{E}\Big[\Big|\varrho\Big(\Big|\tfrac{\hat{p}_{n,k}}{p_k}-1\Big|^{4+\gamma}\Big)\Big|\,\Big|\,\Big|\tfrac{\hat{p}_{n,k}}{p_k}-1\Big| \geq \delta\Big] \leq \max_{x\in[0,1/p_k]}\big\{\big|\varrho\big(|x-1|^{4+\gamma}\big)\big|\big\} \triangleq M \tag{68}$$

where M will serve subsequently as an abbreviative notation. For the third term in (66), let us expand it using the law of total expectation:

$$\begin{aligned}
\mathsf{Pr}\Big\{\Big|\tfrac{\hat{p}_{n,k}}{p_k}-1\Big| \geq \delta\Big\} &= \mathsf{Pr}\Big\{\Big|\tfrac{\check{p}_{n,k}}{p_k}-1\Big| \geq \delta\Big|\check{\boldsymbol{p}}_n \in \mathbb{P}\Big\}\,\mathsf{Pr}\{\check{\boldsymbol{p}}_n \in \mathbb{P}\} \\
&\quad + \mathsf{Pr}\Big\{\Big|\tfrac{\hat{p}_{n,k}}{p_k}-1\Big| \geq \delta\Big|\check{\boldsymbol{p}}_n \notin \mathbb{P}\Big\}\,\mathsf{Pr}\{\check{\boldsymbol{p}}_n \notin \mathbb{P}\} \\
&= \mathsf{Pr}\Big\{\Big|\tfrac{\check{p}_{n,k}}{p_k}-1\Big| \geq \delta\Big\} + \Big[\mathsf{Pr}\Big\{\Big|\tfrac{\hat{p}_{n,k}}{p_k}-1\Big| \geq \delta\Big|\check{\boldsymbol{p}}_n \notin \mathbb{P}\Big\} \\
&\quad - \mathsf{Pr}\Big\{\Big|\tfrac{\check{p}_{n,k}}{p_k}-1\Big| \geq \delta\Big|\check{\boldsymbol{p}}_n \notin \mathbb{P}\Big\}\Big]\,\mathsf{Pr}\{\check{\boldsymbol{p}}_n \notin \mathbb{P}\} \\
&\leq \mathsf{Pr}\Big\{\Big|\tfrac{\check{p}_{n,k}}{p_k}-1\Big| \geq \delta\Big\} + \mathsf{Pr}\{\check{\boldsymbol{p}}_n \notin \mathbb{P}\} \\
&\leq \mathsf{Pr}\Big\{\Big|\tfrac{\check{p}_{n,k}}{p_k}-1\Big| \geq \delta\Big\} + e^{-nD(\partial\mathbb{P}\boldsymbol{W}\|\boldsymbol{p}\boldsymbol{W})}.
\end{aligned} \tag{69}$$

Here, in the first equality, we have exploited the fact conditioned on $\check{\boldsymbol{p}}_n \in \mathbb{P}$, it holds that $\check{\boldsymbol{p}}_n = \hat{\boldsymbol{p}}_n$ (notice $\check{p}_{n,k}$ in the first probability term on the right-hand side). In the last bounding step, we have used Lemma 1.

To tightly bound the deviation probability

$$\mathsf{Pr}\Big\{\Big|\tfrac{\check{p}_{n,k}}{p_k}-1\Big| \geq \delta\Big\} = \mathsf{Pr}\Big\{\check{p}_{n,k} \notin \big(p_k(1-\delta), p_k(1+\delta)\big)\Big\}$$

remaining on the right-hand side of (69), recall that $\check{p}_{n,k}$ is equal to the inner product $\langle\tilde{\boldsymbol{w}}_k, \hat{\boldsymbol{q}}_n\rangle$. The entries of the type vector $\hat{\boldsymbol{q}}_n$ are jointly multinomially distributed, and their joint cumulant-generating function is given by

$$\boldsymbol{\lambda} \mapsto \log \mathsf{E}\Big[e^{\langle\boldsymbol{\lambda},\hat{\boldsymbol{q}}_n\rangle}\Big] = n\log\left(\sum_{\ell=1}^{K} q_\ell e^{\lambda_\ell/n}\right)$$

with argument $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_K) \in \mathbb{R}^K$, and therefore the cumulant-generating function of $\check{p}_{n,k} = \langle\tilde{\boldsymbol{w}}_k, \hat{\boldsymbol{q}}_n\rangle$ is given by

$$\mathscr{K}_{n,k}(\lambda) \triangleq n\log\left(\sum_{\ell=1}^{K} q_\ell e^{\lambda\tilde{W}_{\ell,k}/n}\right).$$

Since the limit

$$\mathscr{K}_k(\lambda) \triangleq \lim_{n\to\infty}\frac{1}{n}\mathscr{K}_{n,k}(n\lambda) = \log\left(\sum_{\ell=1}^{K} q_\ell e^{\lambda\tilde{W}_{\ell,k}}\right)$$

is a well-defined strictly convex function of $\lambda \in \mathbb{R}$, we can define its Legendre-Fenchel transform as

$$\mathscr{K}_k^\star(x) \triangleq \sup_{\lambda\in\mathbb{R}}\{\lambda x - \mathscr{K}_k(\lambda)\}$$

which is a non-negative, quasi-convex function vanishing at $x = \mathscr{K}'(0) = \langle \tilde{\boldsymbol{w}}_{k'}, \boldsymbol{q} \rangle = p_k \in (0,1)$, and apply the Gärtner–Ellis Theorem (Dembo and Zeitouni, 1998, Theorem 2.3.6) to obtain an upper bound on the large-deviation exponent:

$$\begin{aligned}\limsup_{n\to\infty} \frac{1}{n} \log \mathsf{Pr}\Big\{\Big|\tfrac{\check{p}_{n,k}}{p_k} - 1\Big| \geq \delta\Big\} &\leq - \inf_{x \in (-\infty, p_k(1-\delta)] \cup [p_k(1+\delta), \infty)} \mathscr{K}_k^\star(x) \\ &= -\min\big\{\mathscr{K}_k^\star(p_k(1-\delta)), \mathscr{K}_k^\star(p_k(1+\delta))\big\} \\ &\leq -\min\big\{\lambda p_k(1-\delta) - \mathscr{K}_k(\lambda), \lambda p_k(1+\delta) - \mathscr{K}_k(\lambda)\big\} \\ &= \mathscr{K}_k(\lambda) - p_k\big(\lambda + \delta|\lambda|\big)\end{aligned}$$

for an arbitrary $\lambda \in \mathbb{R}$. Since we are free to choose λ, we may as well restrict λ to being non-negative, thus rendering the absolute value $|\lambda|$ equal to λ. This gives us

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathsf{Pr}\Big\{\Big|\tfrac{\check{p}_{n,k}}{p_k} - 1\Big| \geq \delta\Big\} \leq \mathscr{K}_k(\lambda) - p_k(1+\delta)\lambda \qquad \text{(for any } \lambda \geq 0\text{)}. \tag{70}$$

Given that $\mathscr{K}_k(\lambda)$ is infinitely differentiable, we can replace it in (70) with its second-order Taylor expansion around the origin

$$\begin{aligned}\mathscr{K}_k(\lambda) &= \mathscr{K}_k'(0)\lambda + \frac{\mathscr{K}_k''(0)}{2}\lambda^2 + O(\lambda^3) \\ &= p_k\lambda + \frac{1}{2}\Big(-p_k^2 + \textstyle\sum_{\ell=1}^K q_\ell \tilde{W}_{\ell,k}^2\Big)\lambda^2 + O(\lambda^3)\end{aligned}$$

and, noting that $\mathscr{K}_k''(0) > 0$ (due to strict convexity of $\mathscr{K}_k$), we can further proceed with upper-bounding the large-deviation exponent as follows:

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathsf{Pr}\Big\{\Big|\tfrac{\check{p}_{n,k}}{p_k} - 1\Big| \geq \delta\Big\} \leq \frac{\mathscr{K}_k''(0)}{2}\lambda^2 - p_k\delta\lambda + O(\lambda^3)$$

for arbitrary $\lambda \geq 0$. Setting $\lambda = p_k\delta/\mathscr{K}_k''(0)$ this bound becomes

$$\limsup_{n\to\infty} \frac{1}{n} \log \mathsf{Pr}\Big\{\Big|\tfrac{\check{p}_{n,k}}{p_k} - 1\Big| \geq \delta\Big\} \leq -\frac{p_k^2\delta^2}{2\mathscr{K}_k''(0)} + O\bigg(\frac{p_k^3\delta^3}{\mathscr{K}_k''(0)^3}\bigg)$$

The latter implies that there exist constants $C > 0$ and $\delta_0' > 0$ such that for every $\delta \in (0, \delta_0')$,

$$\mathsf{Pr}\Big\{\Big|\tfrac{\check{p}_{n,k}}{p_k} - 1\Big| \geq \delta\Big\} \leq C\exp\bigg(-\frac{np_k^2\delta^2}{2\mathscr{K}_k''(0)}\bigg). \tag{71}$$

Upon inserting inequalities (67), (68) and (71) into (66), we finally obtain that, for any $\delta \in (0, \min\{\delta_0, \delta_0'\})$,

$$|R_{n,k}| \leq \omega\delta^{4+\gamma} + MC\exp\bigg(-\frac{np_k^2\delta^2}{2\mathscr{K}_k''(0)}\bigg).$$

Multiplying either side with n^2, and substituting δ with $n^{-\frac{1}{2}+\frac{\gamma}{4(4+\gamma)}}$, it holds for all n sufficiently large so as to satisfy $n^{-\frac{1}{2}+\frac{\gamma}{4(4+\gamma)}} < \min\{\delta_0, \delta_0'\}$, we have

$$n^2|R_{n,k}| \leq \omega n^{-\frac{\gamma}{4}} + n^2MC\exp\bigg(-\frac{p_k^2 n^{\frac{\gamma}{2(4+\gamma)}}}{2\mathscr{K}_k''(0)}\bigg).$$

Taking the limit as $n \to \infty$, we conclude (65). This finishes the proof.

Appendix C. Proof of Lemma 12

Using successively the Jensen inequality and the triangle inequality, we obtain the upper bound

$$\begin{aligned}\left|\hat{\mu}_{n,k}^{(\rho)} - \check{\mu}_{n,k}^{(\rho)}\right| &\leq n^\rho \,\mathsf{E}\big[\big|\big(\hat{p}_{n,k} - p_k\big)^\rho - \big(\check{p}_{n,k} - p_k\big)^\rho\big|\big] \\ &= n^\rho \,\mathsf{E}\left[\left|\sum_{r=1}^{\rho}\binom{\rho}{r}\big(\check{p}_{n,k} - p_k\big)^{\rho-r}\big(\hat{p}_{n,k} - \check{p}_{n,k}\big)^r\right|\right] \\ &\leq n^\rho \,\mathsf{E}\left[\sum_{r=1}^{\rho}\binom{\rho}{r}\left|\check{p}_{n,k} - p_k\right|^{\rho-r}\left|\hat{p}_{n,k} - \check{p}_{n,k}\right|^r\right].\end{aligned}$$

Next, by the Cauchy–Schwarz inequality,

$$\begin{aligned}\left|\check{p}_{n,k} - p_k\right| &= \left|\big\langle \tilde{\boldsymbol{w}}_k, \hat{\boldsymbol{q}}_n - \boldsymbol{q}\big\rangle\right| \\ &\leq \big\|\tilde{\boldsymbol{w}}_k\big\|_2 \big\|\hat{\boldsymbol{q}}_n - \boldsymbol{q}\big\|_2 \\ &\leq \sqrt{K}\big\|\tilde{\boldsymbol{w}}_k\big\|_2\end{aligned}$$

(where the last inequality is due to $\hat{\boldsymbol{q}}_n$ and $\boldsymbol{q}$ being probability vectors) so we obtain

$$\left|\hat{\mu}_{n,k}^{(\rho)} - \check{\mu}_{n,k}^{(\rho)}\right| \leq n^\rho \sum_{r=1}^{\rho}\binom{\rho}{r}\Big(\sqrt{K}\big\|\tilde{\boldsymbol{w}}_k\big\|_2\Big)^{\rho-r}\,\mathsf{E}\Big[\big|\hat{p}_{n,k} - \check{p}_{n,k}\big|^r\Big]$$

wherein the remaining expectation may be expressed as

$$\mathsf{E}\Big[\Big|\hat{p}_{n,k} - \check{p}_{n,k}\Big|^r\Big] = \mathsf{E}\Big[\Big|\hat{p}_{n,k} - \check{p}_{n,k}\Big|^r \Big|\hat{p}_{n,k} \neq \check{p}_{n,k}\Big]\,\mathsf{Pr}\{\hat{p}_{n,k} \neq \check{p}_{n,k}\}. \tag{72}$$

To upper-bound it, notice that the expectation on the right-hand side of (72) can be upper-bounded by means of

$$\begin{aligned}\Big|\hat{p}_{n,k} - \check{p}_{n,k}\Big| &= \Big|\hat{p}_{n,k} - \big\langle\hat{\boldsymbol{q}}_n, \tilde{\boldsymbol{w}}_k\big\rangle\Big| \\ &\leq \big|\hat{p}_{n,k}\big| + \big|\big\langle\hat{\boldsymbol{q}}_n, \tilde{\boldsymbol{w}}_k\big\rangle\big| \\ &\leq 1 + \big\|\hat{\boldsymbol{q}}_n\big\|_2\big\|\tilde{\boldsymbol{w}}_k\big\|_2 \\ &\leq 1 + \big\|\tilde{\boldsymbol{w}}_k\big\|_2\end{aligned} \tag{73}$$

to yield

$$\left|\hat{\mu}_{n,k}^{(\rho)} - \mu_{n,k}^{(\rho)}\right| \leq n^\rho \sum_{r=1}^{\rho}\binom{\rho}{r}\Big(\sqrt{K}\big\|\tilde{\boldsymbol{w}}_k\big\|_2\Big)^{\rho-r}\big(1 + \big\|\tilde{\boldsymbol{w}}_k\big\|_2\big)^r\,\mathsf{Pr}\{\hat{p}_{n,k} \neq \check{p}_{n,k}\}. \tag{74}$$

Finally, to upper-bound the probability $\mathsf{Pr}\big\{\hat{p}_{n,k} \neq \check{p}_{n,k}\big\}$, notice that by definition of the projector $\mathrm{Proj}_{\mathbb{P}}(\cdot)$, for the event $\hat{p}_{n,k} \neq \check{p}_{n,k}$ to occur, it is necessary that $\check{\boldsymbol{p}}_n$ be no probability vector. However, the fact that the rows of $\boldsymbol{W}^{-1}$ sum to one ensures that the entries of $\check{\boldsymbol{p}}_n$ will sum to one as well. Therefore $\check{\boldsymbol{p}}_n$ fails to be a probability vector whenever some of its entries lie outside of the unit interval. Hence,

$$\begin{aligned}\mathsf{Pr}\{\hat{p}_{n,k} \neq \check{p}_{n,k}\} &\leq \mathsf{Pr}\{\hat{\boldsymbol{p}}_n \neq \check{\boldsymbol{p}}_n\} \\ &\leq e^{-nD(\partial\mathbb{P}\boldsymbol{W}\|\boldsymbol{p}\boldsymbol{W})}\end{aligned} \tag{75}$$

by Lemma 1. Combining (74) and (75), we can conclude that

$$\left|\hat{\mu}_{n,k}^{(\rho)} - \check{\mu}_{n,k}^{(\rho)}\right| \leq C n^{\rho} e^{-nD(\partial\mathbb{P}_{\boldsymbol{W}}\|\boldsymbol{p}\boldsymbol{W})}.$$

for some positive constant $C > 0$ which can be bounded for instance as

$$\begin{aligned} C &\leq \sum_{r=1}^{\rho} \binom{\rho}{r} \left(\sqrt{K}\left\|\tilde{\boldsymbol{w}}_k\right\|_2\right)^{\rho-r} \left(1 + \left\|\tilde{\boldsymbol{w}}_k\right\|_2\right)^{r} \\ &\leq \left(1 + \sqrt{K}\left(1 + \left\|\tilde{\boldsymbol{w}}_k\right\|_2\right)\right)^{\rho}. \end{aligned} \tag{76}$$

This finishes the proof.

Appendix D. Proof of Theorem 3

The MSE loss metric can be expressed as

$$\begin{aligned} \mathscr{L}_{\mathrm{MSE}}^{(n)}(\boldsymbol{p}, \boldsymbol{W}) &= \mathsf{E}\left[\|\hat{\boldsymbol{p}}_n - \boldsymbol{p}\|_2^2\right] \\ &= \mathsf{E}\left[\|\hat{\boldsymbol{p}}_n - \check{\boldsymbol{p}}_n + \check{\boldsymbol{p}}_n - \boldsymbol{p}\|_2^2\right] \\ &= \sum_{k\in[K]} \frac{\check{\mu}_{n,k}^{(2)}}{n^2} + \mathsf{E}\left[\|\hat{\boldsymbol{p}}_n - \check{\boldsymbol{p}}_n\|_2^2\right] + 2\,\mathsf{E}\left[\langle \hat{\boldsymbol{p}}_n - \check{\boldsymbol{p}}_n, \check{\boldsymbol{p}}_n - \boldsymbol{p}\rangle\right]. \end{aligned}$$

We shall prove that, in the last line, the first term [cf. (57)]

$$\frac{1}{n^2}\sum_{k\in[K]} \check{\mu}_{n,k}^{(2)} = \frac{1}{n}\sum_{k\in[K]} (\nu_{2,k} - \nu_{1,k}^2)$$

is the dominant term (as $n \to \infty$), whereas the two other terms decay exponentially in n. Recall from the proof of Lemma 12 that [cf. (73)]

$$\left|\hat{p}_{n,k} - \check{p}_{n,k}\right| \leq 1 + \|\tilde{\boldsymbol{w}}_k\|_2 \tag{77}$$

where $\tilde{\boldsymbol{w}}_k$ is the k-th column of $\boldsymbol{W}^{-1}$. Following the exact same steps as in (73) except for replacing $\hat{p}_{n,k}$ with p_k, one can show

$$\left|p_k - \check{p}_{n,k}\right| \leq 1 + \|\tilde{\boldsymbol{w}}_k\|_2. \tag{78}$$

Hence, using (77),

$$\begin{aligned} \mathsf{E}\left[\|\hat{\boldsymbol{p}}_n - \check{\boldsymbol{p}}_n\|_2^2\right] &= \mathsf{E}\left[\|\hat{\boldsymbol{p}}_n - \check{\boldsymbol{p}}_n\|_2^2 \middle| \hat{\boldsymbol{p}}_n \neq \check{\boldsymbol{p}}_n\right] \mathsf{Pr}\{\hat{\boldsymbol{p}}_n \neq \check{\boldsymbol{p}}_n\} \\ &\leq \sum_{k\in[K]} \left(1 + \|\tilde{\boldsymbol{w}}_k\|_2\right)^2 \mathsf{Pr}\{\hat{\boldsymbol{p}}_n \neq \check{\boldsymbol{p}}_n\}. \end{aligned}$$

Likewise, using (77) and (78),

$$\begin{aligned} \left|\mathsf{E}\left[\langle \hat{\boldsymbol{p}}_n - \check{\boldsymbol{p}}_n, \check{\boldsymbol{p}}_n - \boldsymbol{p}\rangle\right]\right| &= \left|\mathsf{E}\left[\langle \hat{\boldsymbol{p}}_n - \check{\boldsymbol{p}}_n, \check{\boldsymbol{p}}_n - \boldsymbol{p}\rangle \middle| \hat{\boldsymbol{p}}_n \neq \check{\boldsymbol{p}}_n\right]\right| \cdot \mathsf{Pr}\{\hat{\boldsymbol{p}}_n \neq \check{\boldsymbol{p}}_n\} \\ &\leq \sum_{k\in[K]} \mathsf{E}\left[\left|\hat{p}_{n,k} - \check{p}_{n,k}\right| \cdot \left|\check{p}_{n,k} - p_k\right| \middle| \hat{\boldsymbol{p}}_n \neq \check{\boldsymbol{p}}_n\right] \mathsf{Pr}\{\hat{\boldsymbol{p}}_n \neq \check{\boldsymbol{p}}_n\} \\ &\leq \sum_{k\in[K]} \left(1 + \|\tilde{\boldsymbol{w}}_k\|_2\right)^2 \mathsf{Pr}\{\hat{\boldsymbol{p}}_n \neq \check{\boldsymbol{p}}_n\}. \end{aligned}$$

Invoking Lemma 1, we can conclude

$$\mathscr{L}^{(n)}_{\mathrm{MSE}}(\boldsymbol{p},\boldsymbol{W}) = \frac{1}{n}\sum_{k\in[K]}(\nu_{2,k}-\nu_{1,k}^2) + O\big(e^{-nD(\partial\mathbb{P}\boldsymbol{W}\|\boldsymbol{p}\boldsymbol{W})}\big).$$

Appendix E. Proof of Theorem 4

The TV loss metric is close to $\mathsf{E}\big[\|\check{\boldsymbol{p}}_n-\boldsymbol{p}\|_1\big]$, up to an error term which can be bounded by means of Jensen's inequality and the triangle inequality as

$$\begin{aligned}\Big|\mathscr{L}^{(n)}_{\mathrm{TV}}(\boldsymbol{p},\boldsymbol{W}) - \mathsf{E}\big[\|\check{\boldsymbol{p}}_n-\boldsymbol{p}\|_1\big]\Big| &= \Big|\mathsf{E}\big[\|\hat{\boldsymbol{p}}_n-\boldsymbol{p}\|_1\big] - \mathsf{E}\big[\|\check{\boldsymbol{p}}_n-\boldsymbol{p}\|_1\big]\Big| \\ &\le \mathsf{E}\Big[\big|\|\hat{\boldsymbol{p}}_n-\boldsymbol{p}\|_1 - \|\check{\boldsymbol{p}}_n-\boldsymbol{p}\|_1\big|\Big] \\ &\le \mathsf{E}\big[\|\hat{\boldsymbol{p}}_n-\check{\boldsymbol{p}}_n\|_1\big] \\ &= \mathsf{E}\big[\|\hat{\boldsymbol{p}}_n-\check{\boldsymbol{p}}_n\|_1 \;\big|\; \hat{\boldsymbol{p}}_n\neq\check{\boldsymbol{p}}_n\big]\big|\cdot\mathsf{Pr}\{\hat{\boldsymbol{p}}_n\neq\check{\boldsymbol{p}}_n\}.\end{aligned}$$

Using the bound $\big|\hat{p}_{n,k}-\check{p}_{n,k}\big|\le 1+\|\tilde{\boldsymbol{w}}_k\|_2$ from the proof of Lemma 12 [cf. (73)], we obtain

$$\Big|\mathscr{L}^{(n)}_{\mathrm{TV}}(\boldsymbol{p},\boldsymbol{W}) - \mathsf{E}\big[\|\check{\boldsymbol{p}}_n-\boldsymbol{p}\|_1\big]\Big| \le \Big(K+\textstyle\sum_{k\in[K]}\|\tilde{\boldsymbol{w}}_k\|_2\Big)\,\mathsf{Pr}\{\hat{\boldsymbol{p}}_n\neq\check{\boldsymbol{p}}_n\}.$$

Finally, by Lemma 1, we conclude that

$$\mathscr{L}^{(n)}_{\mathrm{TV}}(\boldsymbol{p},\boldsymbol{W}) = \mathsf{E}\big[\|\check{\boldsymbol{p}}_n-\boldsymbol{p}\|_1\big] + O\big(e^{-nD(\partial\mathbb{P}\boldsymbol{W}\|\boldsymbol{p}\boldsymbol{W})}\big)$$

meaning that we can focus on expanding $\mathsf{E}\big[\|\check{\boldsymbol{p}}_n-\boldsymbol{p}\|_1\big]$ instead of $\mathscr{L}^{(n)}_{\mathrm{TV}}(\boldsymbol{p},\boldsymbol{W})$, since both quantities are exponentially close (hence asymptotically equivalent). Next, let us define the centered unit-variance random variables

$$Z_{n,k} \triangleq \frac{\check{p}_{n,k}-p_k}{\sqrt{\mathrm{var}(\check{p}_{n,k})}} = \frac{\sqrt{n}(\check{p}_{n,k}-p_k)}{\sqrt{\nu_{2,k}-\nu_{1,k}^2}}$$

where the second equality holds because [cf. (57)]

$$n^2\,\mathrm{var}(\check{p}_{n,k}) = \check{\mu}^{(2)}_{n,k} = n\big(\nu_{2,k}-\nu_{1,k}^2\big).$$

Recall that the moment-generating function of $n(\check{p}_{n,k}-p_k)$ was introduced in (59), where it is denoted as $\mathscr{M}_k$. Hence, from (59) we infer that the moment-generating function of $Z_{n,k}$ is

$$\begin{aligned}\mathsf{E}\big[e^{\lambda Z_{n,k}}\big] &= \mathscr{M}_k\left(\frac{\lambda}{\sqrt{n(\nu_{2,k}-\nu_{1,k}^2)}}\right) \\ &= \left(\sum_{\ell=1}^{K} q_\ell \exp\left(\frac{\lambda\big(\tilde{W}_{\ell,k}-\langle\tilde{\boldsymbol{w}}_k,\boldsymbol{q}\rangle\big)}{\sqrt{n(\nu_{2,k}-\nu_{1,k}^2)}}\right)\right)^n \\ &= \left(1-\frac{\lambda^2}{2n}+o(\lambda^2)\right)^n \qquad (79)\end{aligned}$$

where the last equality follows from Taylor-expanding the exponential function to second order and from noticing that the first and second order terms of said expansion simplify due to

$$\sum_{\ell=1}^{K} q_\ell \tilde{W}_{\ell,k} - \langle \tilde{\boldsymbol{w}}_k, \boldsymbol{q} \rangle = 0$$
$$\sum_{\ell=1}^{K} q_\ell \left(\frac{\tilde{W}_{\ell,k} - \langle \tilde{\boldsymbol{w}}_k, \boldsymbol{q} \rangle}{\sqrt{\nu_{2,k} - \nu_{1,k}^2}} \right)^2 = 1.$$

For the latter equality, recall that $\langle \tilde{\boldsymbol{w}}_k, \boldsymbol{q} \rangle = \sum_\ell q_\ell \tilde{W}_{\ell,k} = p_k = \nu_{1,k}$ and that [cf. (64a)]

$$\sum_{\ell=1}^{K} q_\ell \Big(\tilde{W}_{\ell,k} - \langle \tilde{\boldsymbol{w}}_k, \boldsymbol{q} \rangle \Big)^2 = \nu_{2,k} - \nu_{1,k}^2 = \frac{\check{\mu}_{n,k}^{(2)}}{n}.$$

Since $\lim_{n\to\infty} \big(1 + \frac{x}{n} + o(n^{-1})\big)^n = e^x$, we have that for any $\lambda \in \mathbb{R}$, the limit of the moment-generating function of $Z_{n,k}$ as given in (79) is

$$\lim_{n\to\infty} \mathsf{E}\big[e^{\lambda Z_{n,k}}\big] = e^{-\lambda^2/2}$$

which is the moment-generating function of the normal distribution. By Lévy's continuity theorem for moment-generating functions (Billingsley, 1995, Problem 30.4), the pointwise convergence of the moment-generating functions of $Z_{n,k}$ to that of the normal distribution as $n \to \infty$ implies that $Z_{n,k}$ converges in law to a normal random variable $Z \sim \mathscr{N}(0,1)$. Therefore,

$$\begin{aligned}
\lim_{n\to\infty} \sqrt{n} \mathscr{L}_{\mathrm{TV}}^{(n)}(\boldsymbol{p}, \boldsymbol{W}) &= \lim_{n\to\infty} \sqrt{n} \sum_{k=1}^{K} \mathsf{E}\big[|\check{p}_{n,k} - p_k|\big] \\
&= \sum_{k=1}^{K} \sqrt{\nu_{2,k} - \nu_{1,k}^2} \lim_{n\to\infty} \mathsf{E}\big[|Z_{n,k}|\big] \\
&= \sum_{k=1}^{K} \sqrt{\nu_{2,k} - \nu_{1,k}^2}\, \mathsf{E}\big[|Z|\big] \\
&= \sum_{k=1}^{K} \sqrt{\nu_{2,k} - \nu_{1,k}^2} \sqrt{\frac{2}{\pi}}.
\end{aligned}$$

This concludes the proof.

Appendix F. Proof of Theorem 5

Let $\mathscr{D}(\cdot,\cdot)$ denote a distance metric between probability vectors, which shall stand either for the f-divergence $D_f(\cdot\|\cdot)$ or for the MSE metric $\|\cdot - \cdot\|_2^2$. These metrics are jointly convex and continuous in both arguments. In particular, they are convex in the first argument, i.e.,

$$\mathscr{D}(\lambda \boldsymbol{p} + (1-\lambda)\boldsymbol{p}', \boldsymbol{p}'') \le \lambda \mathscr{D}(\boldsymbol{p}, \boldsymbol{p}'') + (1-\lambda)\mathscr{D}(\boldsymbol{p}', \boldsymbol{p}'').$$

Denoting the probability simplex as $\mathbb{P}$, let us define

$$\bar{\mathscr{D}}(\boldsymbol{p}) \triangleq \sup_{\boldsymbol{p}' \in \mathbb{P}} \mathscr{D}(\boldsymbol{p}', \boldsymbol{p})$$

which is always finite in the case of the MSE metric, and also finite in the case of the f-divergence metric as long as $\lim_{x \downarrow 0} x f(x) < +\infty$, which holds due to the assumption that $f(0)$ is finite [cf. Theorem 2].

As usual, assume that $\boldsymbol{x}_n \sim \boldsymbol{p}^{\otimes n}$ is a vector of n i.i.d. source samples, and let $\boldsymbol{y}_n \sim (\boldsymbol{p}\boldsymbol{W})^{\otimes n}$ be the corresponding outputs from n copies of the channel $\boldsymbol{W}$. In addition, assume that $\boldsymbol{y}'_n$ are the outputs from passing $\boldsymbol{y}_n$ through n copies of another channel $\boldsymbol{W}'$. Let us define the estimates

$$\begin{aligned}\check{\boldsymbol{p}}'_n &\triangleq \boldsymbol{t}(\boldsymbol{y}'_n)(\boldsymbol{W}\boldsymbol{W}')^{-1} \\ \hat{\boldsymbol{p}}'_n &\triangleq \operatorname{Proj}_{\mathbb{P}}\big(\boldsymbol{t}(\boldsymbol{y}'_n)(\boldsymbol{W}\boldsymbol{W}')^{-1}\big)\end{aligned}$$

based on the degraded observation $\boldsymbol{y}'_n$. In addition, let us define the estimates of the output distribution of the first channel $\boldsymbol{W}$ as

$$\begin{aligned}\check{\boldsymbol{q}}_n &\triangleq \boldsymbol{t}(\boldsymbol{y}'_n)(\boldsymbol{W}')^{-1} \\ \hat{\boldsymbol{q}}_n &\triangleq \operatorname{Proj}_{\mathbb{P}}\big(\boldsymbol{t}(\boldsymbol{y}'_n)(\boldsymbol{W}')^{-1}\big).\end{aligned}$$

Using the law of total probability,

$$\begin{aligned}\mathscr{L}^{(n)}(\boldsymbol{p}, \boldsymbol{W}\boldsymbol{W}') &= \mathsf{E}\big[\mathscr{D}\big(\hat{\boldsymbol{p}}'_n, \boldsymbol{p}\big)\big] \\ &\geq \mathsf{E}\big[\mathscr{D}\big(\check{\boldsymbol{p}}'_n, \boldsymbol{p}\big)\big|\big(\check{\boldsymbol{q}}_n, \check{\boldsymbol{p}}'_n\big) \in \mathbb{P}^2\big]\, \mathsf{Pr}\big\{\big(\check{\boldsymbol{q}}_n, \check{\boldsymbol{p}}'_n\big) \in \mathbb{P}^2\big\} \qquad (80)\end{aligned}$$

wherein we have exploited the fact that the event $\check{\boldsymbol{p}}'_n \in \mathbb{P}$ implies $\check{\boldsymbol{p}}'_n = \hat{\boldsymbol{p}}'_n$ by idempotence of the projection. We can bound the probability factor on the right-hand side of (80) as

$$\begin{aligned}\mathsf{Pr}\big\{\big(\check{\boldsymbol{q}}_n, \check{\boldsymbol{p}}'_n\big) \in \mathbb{P}^2\big\} &= \mathsf{Pr}\{\check{\boldsymbol{p}}'_n \in \mathbb{P}\} - \mathsf{Pr}\{\check{\boldsymbol{p}}'_n \in \mathbb{P} \;\big|\; \check{\boldsymbol{q}}_n \notin \mathbb{P}\}\, \mathsf{Pr}\{\check{\boldsymbol{p}}_n \notin \mathbb{P}\} \\ &\geq \mathsf{Pr}\{\check{\boldsymbol{p}}'_n \in \mathbb{P}\} - \mathsf{Pr}\{\check{\boldsymbol{q}}_n \notin \mathbb{P}\} \\ &\geq 1 - e^{-n\mathscr{D}(\partial\mathbb{P}\boldsymbol{W}\boldsymbol{W}'\|\boldsymbol{p}\boldsymbol{W}\boldsymbol{W}')} - e^{-n\mathscr{D}(\partial\mathbb{P}\boldsymbol{W}'\|\boldsymbol{p}\boldsymbol{W}\boldsymbol{W}')} \\ &\geq 1 - 2e^{-nM} \qquad (81)\end{aligned}$$

by twice applying Lemma 1. Here, M denotes the constant where

$$M = \min\big\{\mathscr{D}(\partial\mathbb{P}\boldsymbol{W}\boldsymbol{W}'\|\boldsymbol{p}\boldsymbol{W}\boldsymbol{W}'), \mathscr{D}(\partial\mathbb{P}\boldsymbol{W}'\|\boldsymbol{p}\boldsymbol{W}\boldsymbol{W}')\big\}.$$

In order to bound the other factor in (80), let us first define $\boldsymbol{\tau}_n$ and $\boldsymbol{\tau}'_n$ as being the respective types $\boldsymbol{t}(\boldsymbol{y}_n)$ and $\boldsymbol{t}(\boldsymbol{y}'_n)$, conditionally on the event

$$\big\{\big(\check{\boldsymbol{q}}_n, \check{\boldsymbol{p}}'_n\big) \in \mathbb{P}^2\big\} = \{\boldsymbol{t}(\boldsymbol{y}_n)\boldsymbol{W}^{-1} \in \mathbb{P}, \boldsymbol{t}(\boldsymbol{y}'_n)(\boldsymbol{W}\boldsymbol{W}')^{-1} \in \mathbb{P}\}.$$

The pair $(\boldsymbol{\tau}_n, \boldsymbol{\tau}'_n)$ thus has a probability mass function

$$\begin{aligned}\mathsf{Pr}\{(\boldsymbol{\tau}_n, \boldsymbol{\tau}'_n) = (\tilde{\boldsymbol{\tau}}_n, \tilde{\boldsymbol{\tau}}'_n)\} &= \mathsf{Pr}\{(\boldsymbol{t}(\boldsymbol{y}_n), \boldsymbol{t}(\boldsymbol{y}'_n)) = (\tilde{\boldsymbol{\tau}}_n, \tilde{\boldsymbol{\tau}}'_n)\} \\ &\quad \times \frac{\mathbb{1}\big\{\tilde{\boldsymbol{\tau}}'_n(\boldsymbol{W}')^{-1} \in \mathbb{P}, \tilde{\boldsymbol{\tau}}'_n(\boldsymbol{W}\boldsymbol{W}')^{-1} \in \mathbb{P}\big\}}{\mathsf{Pr}\big\{\boldsymbol{t}(\boldsymbol{y}'_n)(\boldsymbol{W}')^{-1} \in \mathbb{P}, \boldsymbol{t}(\boldsymbol{y}'_n)(\boldsymbol{W}\boldsymbol{W}')^{-1} \in \mathbb{P}\big\}}\end{aligned}$$

where $\mathbb{1}\{\cdot\}$ stands for the indicator function. Hence, the first factor on the right-hand side of (80) can be lower-bounded as follows:

$$\begin{aligned}\mathsf{E}\big[\mathscr{D}\big(\check{\boldsymbol{p}}'_n,\boldsymbol{p}\big)\big|\big(\check{\boldsymbol{q}}_n,\check{\boldsymbol{p}}'_n\big)\in\mathbb{P}^2\big] &= \mathsf{E}\big[\mathscr{D}\big(\boldsymbol{\tau}'_n(\boldsymbol{W}\boldsymbol{W}')^{-1},\boldsymbol{p}\big)\big]\\ &= \mathsf{E}\big[\mathsf{E}\big[\mathscr{D}\big(\boldsymbol{\tau}'_n(\boldsymbol{W}\boldsymbol{W}')^{-1},\boldsymbol{p}\big)\big|\boldsymbol{\tau}_n\big]\big]\\ &> \mathsf{E}\big[\mathscr{D}\big(\mathsf{E}[\boldsymbol{\tau}'_n|\boldsymbol{\tau}_n](\boldsymbol{W}\boldsymbol{W}')^{-1},\boldsymbol{p}\big)\big]. \end{aligned}\tag{82}$$

The bounding step is due to the convexity of $\mathscr{D}(\cdot,\cdot)$ in the first argument (Jensen's inequality). The inequality is strict because the metrics $\mathscr{D}(\cdot,\cdot)$ are strictly convex in the first argument by assumption, and because $\boldsymbol{\tau}'_n$ conditioned on any $\boldsymbol{\tau}_n$ is non-deterministic.

Next, we show that $\mathsf{E}[\boldsymbol{\tau}'_n|\boldsymbol{\tau}_n](\boldsymbol{W}')^{-1}$ is exponentially close to $\boldsymbol{\tau}_n$ as $n\to\infty$. By the law of total probability,

$$\begin{aligned}\tilde{\boldsymbol{\tau}}_n\boldsymbol{W}' &= \mathsf{E}\big[\boldsymbol{t}(\boldsymbol{y}'_n)\big|\boldsymbol{t}(\boldsymbol{y}_n)=\tilde{\boldsymbol{\tau}}_n\big]\\ &= \mathsf{E}\big[\boldsymbol{t}(\boldsymbol{y}'_n)\big|\boldsymbol{t}(\boldsymbol{y}_n)=\tilde{\boldsymbol{\tau}}_n,\check{\boldsymbol{q}}_n\in\mathbb{P}\big]\,\mathsf{Pr}\big\{\check{\boldsymbol{q}}_n\in\mathbb{P}\big\}+\mathsf{E}[\boldsymbol{t}(\boldsymbol{y}'_n)|\boldsymbol{t}(\boldsymbol{y}_n)=\tilde{\boldsymbol{\tau}}_n,\check{\boldsymbol{q}}_n\notin\mathbb{P}]\,\mathsf{Pr}\{\check{\boldsymbol{q}}_n\notin\mathbb{P}\}.\end{aligned}$$

Note that the first expectation in the last line is equal to

$$\mathsf{E}\big[\boldsymbol{t}(\boldsymbol{y}'_n)\big|\boldsymbol{t}(\boldsymbol{y}_n)=\tilde{\boldsymbol{\tau}}_n,\check{\boldsymbol{q}}_n\in\mathbb{P}\big]=\mathsf{E}[\boldsymbol{\tau}'_n|\boldsymbol{\tau}_n=\tilde{\boldsymbol{\tau}}_n]$$

by definition of $(\boldsymbol{\tau}_n,\boldsymbol{\tau}'_n)$. Since by Lemma (1), we have

$$\begin{aligned}1-\mathsf{Pr}\{\check{\boldsymbol{q}}_n\in\mathbb{P}\} &= \mathsf{Pr}\{\check{\boldsymbol{q}}_n\notin\mathbb{P}\}\\ &\le e^{-n\mathscr{D}(\partial\mathbb{P}\boldsymbol{W}'\|\boldsymbol{p}\boldsymbol{W}\boldsymbol{W}')}\end{aligned}$$

it is straightforward to bound the difference between $\tilde{\boldsymbol{\tau}}_n\boldsymbol{W}'$ and $\mathsf{E}[\boldsymbol{\tau}'_n|\boldsymbol{\tau}_n=\tilde{\boldsymbol{\tau}}_n]$ entrywise from below and from above as

$$\begin{aligned}\tilde{\boldsymbol{\tau}}_n\boldsymbol{W}'-\mathsf{E}[\boldsymbol{\tau}'_n|\boldsymbol{\tau}_n=\tilde{\boldsymbol{\tau}}_n] &\ge -\mathsf{E}[\boldsymbol{\tau}'_n|\boldsymbol{\tau}_n=\tilde{\boldsymbol{\tau}}_n]e^{-n\mathscr{D}(\partial\mathbb{P}\boldsymbol{W}'\|\boldsymbol{p}\boldsymbol{W}\boldsymbol{W}')}\\ &\ge -\mathbf{1}e^{-n\mathscr{D}(\partial\mathbb{P}\boldsymbol{W}'\|\boldsymbol{p}\boldsymbol{W}\boldsymbol{W}')} &\text{(83a)}\\ \tilde{\boldsymbol{\tau}}_n\boldsymbol{W}'-\mathsf{E}[\boldsymbol{\tau}'_n|\boldsymbol{\tau}_n=\tilde{\boldsymbol{\tau}}_n] &\le \mathsf{E}[\boldsymbol{t}(\boldsymbol{y}'_n)|\boldsymbol{t}(\boldsymbol{y}_n)=\tilde{\boldsymbol{\tau}}_n,\check{\boldsymbol{q}}_n\notin\mathbb{P}]e^{-n\mathscr{D}(\partial\mathbb{P}\boldsymbol{W}'\|\boldsymbol{p}\boldsymbol{W}\boldsymbol{W}')}\\ &\le \mathbf{1}e^{-n\mathscr{D}(\partial\mathbb{P}\boldsymbol{W}'\|\boldsymbol{p}\boldsymbol{W}\boldsymbol{W}')}. &\text{(83b)}\end{aligned}$$

The inequalities in (83a)–(83b) hold entrywise in the sense that they stand for K lines of simultaneously holding inequalities. Combining (83a) and (83b), we can bound the Euclidean distance

$$\big\|\tilde{\boldsymbol{\tau}}_n\boldsymbol{W}'-\mathsf{E}[\boldsymbol{\tau}'_n|\boldsymbol{\tau}_n=\tilde{\boldsymbol{\tau}}_n]\big\|_2\le\sqrt{K}e^{-n\mathscr{D}(\partial\mathbb{P}\boldsymbol{W}'\|\boldsymbol{p}\boldsymbol{W}\boldsymbol{W}')}.$$

Taking into account that the metric $\mathscr{D}(\cdot,\cdot)$ is continuous in the first argument, the mapping

$$\boldsymbol{\tau}\mapsto\mathscr{D}(\boldsymbol{\tau}(\boldsymbol{W}\boldsymbol{W}')^{-1}\|\boldsymbol{p})\quad(\boldsymbol{\tau}\in\mathbb{P})$$

is continuous. Being defined on a compact set (the probability simplex), this mapping is also uniformly continuous. Hence, for any $\epsilon>0$ there exists an N_ϵ such that for all $n\ge N_\epsilon$,

$$\mathsf{E}\big[\mathscr{D}\big(\mathsf{E}[\boldsymbol{\tau}'_n|\boldsymbol{\tau}_n](\boldsymbol{W}\boldsymbol{W}')^{-1},\boldsymbol{p}\big)\big]\ge\mathsf{E}\big[\mathscr{D}\big(\boldsymbol{\tau}_n\boldsymbol{W}^{-1},\boldsymbol{p}\big)\big]-\epsilon.\tag{84}$$

On the other hand,

$$\begin{aligned}
\mathscr{L}^{(n)}(\boldsymbol{p},\boldsymbol{W}) &= \mathsf{E}\big[\mathscr{D}\big(\hat{\boldsymbol{p}}_n,\boldsymbol{p}\big)\big] \\
&= \mathsf{E}\big[\mathscr{D}\big(\check{\boldsymbol{p}}_n,\boldsymbol{p}\big)\big|\big(\check{\boldsymbol{p}}_n,\check{\boldsymbol{p}}'_n\big)\in\mathbb{P}^2\big]\,\mathsf{Pr}\big\{\big(\check{\boldsymbol{p}}_n,\check{\boldsymbol{p}}'_n\big)\in\mathbb{P}^2\big\} \\
&\qquad + \mathsf{E}\big[\mathscr{D}\big(\hat{\boldsymbol{p}}_n,\boldsymbol{p}\big)\big|\big(\check{\boldsymbol{p}}_n,\check{\boldsymbol{p}}'_n\big)\notin\mathbb{P}^2\big]\,\mathsf{Pr}\big\{\big(\check{\boldsymbol{p}}_n,\check{\boldsymbol{p}}'_n\big)\notin\mathbb{P}^2\big\} \\
&\le \mathsf{E}\big[\mathscr{D}\big(\check{\boldsymbol{p}}_n,\boldsymbol{p}\big)\big|\big(\check{\boldsymbol{p}}_n,\check{\boldsymbol{p}}'_n\big)\in\mathbb{P}^2\big] \\
&\qquad + \bar{\mathscr{D}}(\boldsymbol{p})\big(e^{-n\mathscr{D}(\partial\mathbb{P}\boldsymbol{W}\|\boldsymbol{p}\boldsymbol{W})}+e^{-n\mathscr{D}(\partial\mathbb{P}\boldsymbol{W}\boldsymbol{W}'\|\boldsymbol{p}\boldsymbol{W}\boldsymbol{W}')}\big) \\
&\le \mathsf{E}\big[\mathscr{D}\big(\check{\boldsymbol{p}}_n,\boldsymbol{p}\big)\big|\big(\check{\boldsymbol{p}}_n,\check{\boldsymbol{p}}'_n\big)\in\mathbb{P}^2\big] \\
&\qquad + 2\bar{\mathscr{D}}(\boldsymbol{p})e^{-n\mathscr{D}(\partial\mathbb{P}\boldsymbol{W}\boldsymbol{W}'\|\boldsymbol{p}\boldsymbol{W}\boldsymbol{W}')} \\
&= \mathsf{E}\big[\mathscr{D}\big(\boldsymbol{\tau}_n\boldsymbol{W}^{-1},\boldsymbol{p}\big)\big] + 2\bar{\mathscr{D}}(\boldsymbol{p})e^{-n\mathscr{D}(\partial\mathbb{P}\boldsymbol{W}\boldsymbol{W}'\|\boldsymbol{p}\boldsymbol{W}\boldsymbol{W}')} \\
&\le \mathsf{E}\big[\mathscr{D}\big(\boldsymbol{\tau}_n\boldsymbol{W}^{-1},\boldsymbol{p}\big)\big] + 2\bar{\mathscr{D}}(\boldsymbol{p})e^{-nM}
\end{aligned} \tag{85}$$

where the first bounding step follows from the union of events, and from twice applying Lemma 1, while in the second bounding step we have used the data processing inequality $\mathscr{D}(\cdot\boldsymbol{W}\boldsymbol{W}'\|\boldsymbol{p}\boldsymbol{W}\boldsymbol{W}') \le \mathscr{D}(\cdot\boldsymbol{W}\|\boldsymbol{p}\boldsymbol{W})$. Combining all inequalities (80), (81), (82), (84) and (85), we obtain

$$\mathscr{L}^{(n)}(\boldsymbol{p},\boldsymbol{W}\boldsymbol{W}') > \big(\mathscr{L}^{(n)}(\boldsymbol{p},\boldsymbol{W}) - 2\bar{\mathscr{D}}(\boldsymbol{p})e^{-nM} - \epsilon\big)\big(1-2e^{-nM}\big)$$

for arbitrarily small $\epsilon > 0$ and sufficiently large n. It follows that $\mathscr{L}^{(n)}(\boldsymbol{p},\boldsymbol{W}\boldsymbol{W}') > \mathscr{L}^{(n)}(\boldsymbol{p},\boldsymbol{W})$ for sufficiently large n, which concludes the proof.

Appendix G. Proof of Theorem 7

Using the identity (20a), the data-processing inequality (26a) can be written as

$$\sum_{(k,k')\in[K]^2}\frac{p_k}{p_{k'}}\big(\Phi_{k,k'}(\boldsymbol{W}\boldsymbol{W}') - \Phi_{k,k'}(\boldsymbol{W})\big) \ge 0. \tag{86}$$

Since this inequality holds for any probability vector $\boldsymbol{p}$, it holds in particular for the probability vector taking value ε^2 at the i-th coordinate, value $1-(K-2)\varepsilon-\varepsilon^2$ at the j-th coordinate, and value ε on all other coordinates, where $\varepsilon>0$ is assumed to be sufficiently small to ensure $1-(K-2)\varepsilon-\varepsilon^2\ge 0$. Expanding the sum on the left-hand side of (86) for this probability vector, while abbreviating the matrix $\boldsymbol{\Phi}(\boldsymbol{W}\boldsymbol{W}')-\boldsymbol{\Phi}(\boldsymbol{W})$ as $\boldsymbol{\Delta\Phi}$ for notational concision, we obtain the inequality

$$\begin{aligned}
&\frac{\varepsilon^2}{1-(K-2)\varepsilon-\varepsilon^2}\Delta\Phi_{i,j} + \frac{1-(K-2)\varepsilon-\varepsilon^2}{\varepsilon^2}\Delta\Phi_{j,i} \\
&\quad + \varepsilon\sum_{k'\in[K]\setminus\{i,j\}}\Delta\Phi_{i,k'} + \frac{1}{\varepsilon}\sum_{k\in[K]\setminus\{i,j\}}\Delta\Phi_{k,i} \\
&\quad + \frac{\varepsilon}{1-(K-2)\varepsilon-\varepsilon^2}\sum_{k\in[K]\setminus\{i,j\}}\Delta\Phi_{k,j} \\
&\quad + \frac{1-(K-2)\varepsilon-\varepsilon^2}{\varepsilon}\sum_{k'\in[K]\setminus\{i,j\}}\Delta\Phi_{j,k'} \\
&\quad + \sum_{(k,k')\in([K]\setminus\{i,j\})^2\cup\{i,i\}\cup\{j,j\}}\Delta\Phi_{k,k'} \ge 0.
\end{aligned}$$

Multiplying both sides of this inequality by ε^2 and taking the limit as $\varepsilon \to 0$, we obtain the desired result $\Delta\Phi_{j,i} \geq 0$ which concludes the proof.

Appendix H. Proof of Theorem 9

Let $\boldsymbol{w} = [w_1, \ldots, w_K]$ denote the first row of a circulant mechanism, as defined in (29), which we assume to be full-rank. Owing to its circulant structure, $\boldsymbol{W}$ has an eigendecomposition $\boldsymbol{W} = \boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}^\dagger$ with a symmetric Fourier eigenbasis $[\boldsymbol{F}]_{i,j} = \frac{1}{\sqrt{K}}\xi_K^{(i-1)(j-1)}$ where $\xi_K = e^{-2\pi\jmath/K}$, and a diagonal matrix of eigenvalues $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$. The vector $\boldsymbol{\lambda} = [\lambda_1, \ldots, \lambda_K]$ of eigenvalues is the discrete Fourier transform of $\boldsymbol{w}$, i.e.,

$$\boldsymbol{\lambda} = \boldsymbol{w}\sqrt{K}\boldsymbol{F}.$$

Note that the row-stochasticity of $\boldsymbol{W}$ implies $\lambda_1 = 1$. Furthermore, since $\boldsymbol{W}$ is real-valued, the remaining eigenvalues $(\lambda_2, \ldots, \lambda_K)$ are skew-symmetric, in the sense that $\lambda_k^* = \lambda_{K-k+2}$. Let $\boldsymbol{\lambda} * \boldsymbol{\lambda}$ denote the cyclic self-convolution of $\boldsymbol{\lambda}$, i.e.,

$$[\boldsymbol{\lambda} * \boldsymbol{\lambda}]_k = \sum_{k'=1}^{K} \lambda_{k'} \bar{\lambda}_{k+1-k'}$$

where $\bar{\lambda}$ denotes the K-periodic continuation of $(\lambda_1, \ldots, \lambda_K)$. Then, by the (self-)convolution theorem of the discrete Fourier transform, for any transform pair $(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{\lambda}})$, i.e., $\tilde{\boldsymbol{\lambda}} = \tilde{\boldsymbol{w}}\sqrt{K}\boldsymbol{F}$, it holds that

$$\frac{1}{K}\tilde{\boldsymbol{\lambda}} * \tilde{\boldsymbol{\lambda}} = (\tilde{\boldsymbol{w}} \odot \tilde{\boldsymbol{w}})\sqrt{K}\boldsymbol{F}. \tag{87}$$

Hence, for circulant $\boldsymbol{W} = \boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}^\dagger$, the quantity $\varphi(\boldsymbol{W})$ can be evaluated as follows:

$$\begin{aligned}
\varphi(\boldsymbol{W}) &= \boldsymbol{1}\boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}^\dagger(\boldsymbol{F}\boldsymbol{\Lambda}^{-1}\boldsymbol{F}^\dagger \odot \boldsymbol{F}\boldsymbol{\Lambda}^{-1}\boldsymbol{F}^\dagger)\boldsymbol{1}^{\mathrm{T}} \\
&\overset{\text{(a)}}{=} \frac{1}{K}\boldsymbol{1}\boldsymbol{F}\boldsymbol{\Lambda}\,\mathrm{diag}(\boldsymbol{\lambda}^{-1} * \boldsymbol{\lambda}^{-1})\boldsymbol{F}^\dagger\boldsymbol{1}^{\mathrm{T}} \\
&\overset{\text{(b)}}{=} \boldsymbol{e}_1\boldsymbol{\Lambda}\,\mathrm{diag}(\boldsymbol{\lambda}^{-1} * \boldsymbol{\lambda}^{-1})\boldsymbol{e}_1^{\mathrm{T}} \\
&\overset{\text{(c)}}{=} \boldsymbol{e}_1\,\mathrm{diag}(\boldsymbol{\lambda}^{-1} * \boldsymbol{\lambda}^{-1})\boldsymbol{e}_1^{\mathrm{T}} \\
&= \sum_{k=1}^{K} \frac{1}{\lambda_k \bar{\lambda}_{2-k}} \\
&= 1 + \sum_{k=2}^{K} \frac{1}{\lambda_k \lambda_{K-k+2}} \\
&= 1 + \sum_{k=2}^{K} \frac{1}{|\lambda_k|^2}.
\end{aligned} \tag{88}$$

Here, (a) results from the self-convolution identity (87); step (b) is due to $\boldsymbol{F}^\dagger\boldsymbol{F} = \mathbf{I}$ and $\boldsymbol{1}\boldsymbol{F} = \sqrt{K}\boldsymbol{e}_1$; step (c) follows from $\boldsymbol{e}_1\boldsymbol{\Lambda} = \boldsymbol{e}_1$. By the Plancherel Theorem,

$$1 + \sum_{k=2}^{K} |\lambda_k|^2 = K \sum_{k=1}^{K} w_k^2$$

and by the harmonic-arithmetic-mean inequality, we can lower-bound the quantity (88) as

$$\varphi(\boldsymbol{W}) \geq 1 + \frac{K^2}{-1 + K\sum_{k=1}^K w_k^2}. \tag{89}$$

Note that the denominator in the last expression is positive, hence minimizing the fraction (subject to an ϵ-privacy constraint) amounts to maximizing the sum of squares $\sum_{k=1}^K w_k^2$, which by Appendix J, Lemma 13, is achieved (up to permutation) by a vector

$$\boldsymbol{w}_\star = \frac{1}{e^\epsilon + K - 1}\begin{bmatrix} e^\epsilon & 1 & 1 & \ldots & 1 \end{bmatrix}. \tag{90}$$

It now suffices to show that the harmonic-arithmetic-mean inequality (89) is indeed satisfied with equality for a circulant mechanism generated by $\boldsymbol{w}_\star$. Said inequality is tight for $|\lambda_2| = \ldots = |\lambda_K|$, and the choice (90) yields eigenvalues

$$\begin{aligned} \lambda_k &= \sum_{\ell=1}^K w_\ell \xi_K^{(k-1)(\ell-1)} \\ &= \frac{1}{e^\epsilon + K - 1}\left(e^\epsilon + \sum_{\ell=2}^K \xi_K^{(k-1)(\ell-1)}\right) \\ &= \frac{e^\epsilon - 1}{e^\epsilon + K - 1}, \quad (k = 2, \ldots, K) \end{aligned} \tag{91}$$

since for $k = 2, \ldots, K$, we have

$$\sum_{\ell=1}^K \xi_K^{(k-1)(\ell-1)} = 0.$$

Given that the right-hand side of (91) does not depend on k, we have indeed $|\lambda_2| = \ldots = |\lambda_K|$, which implies that the harmonic-arithmetic-mean inequality is tight, and thus concludes the proof.

Appendix I. Proof of Lemma 10

It suffices to prove that the function $a_{\boldsymbol{W}} \colon \boldsymbol{p} \mapsto \boldsymbol{p}\boldsymbol{\Phi}(\boldsymbol{W})\boldsymbol{p}^{-\mathrm{T}}$ satisfies that the restriction

$$[0,1] \to \mathbb{R}_+, \quad \lambda \mapsto a_{\boldsymbol{W}}(\lambda \boldsymbol{p} + (1-\lambda)\boldsymbol{p}\boldsymbol{\Pi}_{[i,j]})$$

is convex for any $\boldsymbol{p}$, $\{i,j\}$ and $\boldsymbol{W}$. Singling out two arbitrary variables p_i and p_j out of the coordinates of $\boldsymbol{p}$, we can write $a(\boldsymbol{p})$ in the following way:

$$\begin{aligned} a_{\boldsymbol{W}}(\boldsymbol{p}) &= \sum_{(k,\ell)\in[K]^2} \frac{p_k}{p_\ell}\Phi_{k,\ell}(\boldsymbol{W}) \\ &= \frac{p_i}{p_j}\Phi_{i,j}(\boldsymbol{W}) + \frac{p_j}{p_i}\Phi_{j,i}(\boldsymbol{W}) \\ &\quad + \sum_{\ell\in[K]\setminus\{i,j\}} \frac{p_i}{p_\ell}\Phi_{i,\ell}(\boldsymbol{W}) + \sum_{k\in[K]\setminus\{i,j\}} \frac{p_k}{p_j}\Phi_{k,j}(\boldsymbol{W}) \\ &\quad + \sum_{\ell\in[K]\setminus\{i,j\}} \frac{p_j}{p_\ell}\Phi_{j,\ell}(\boldsymbol{W}) + \sum_{k\in[K]\setminus\{i,j\}} \frac{p_k}{p_i}\Phi_{k,i}(\boldsymbol{W}) \\ &\quad + \sum_{(k,\ell)\in([K]\setminus\{i,j\})^2} \frac{p_k}{p_\ell}\Phi_{k,\ell}(\boldsymbol{W}). \end{aligned} \tag{92}$$

Then, $a_{\boldsymbol{W}}(\lambda\boldsymbol{p}+(1-\lambda)\boldsymbol{p}\boldsymbol{\Pi}_{[i,j]})$ can be written out similarly, by replacing all occurrences of p_i and p_j on the right-hand side of (92) with $\lambda p_i+(1-\lambda)p_j$ and $\lambda p_j+(1-\lambda)p_i$, respectively. It is then easy to see that $a_{\boldsymbol{W}}(\lambda\boldsymbol{p}+(1-\lambda)\boldsymbol{p}\boldsymbol{\Pi}_{[i,j]})$ is convex in λ, since all entries of $\boldsymbol{\Phi}(\boldsymbol{W})$ are non-negative. In fact, every summand on the right-hand side of (92) is (weakly or strongly) convex in λ.

Appendix J. Auxiliary lemma

Lemma 13 *The solution to the maximization problem*

$$\boldsymbol{x}_\star = \underset{\substack{\boldsymbol{x}\in\mathbb{R}_+^K:\\ \|\boldsymbol{x}\|_1=1\\ \forall k,k':\, x_k/x_{k'}\le e^\epsilon}}{\operatorname{argmax}} \|\boldsymbol{x}\|_2^2 \tag{93}$$

is given (up to an arbitrary permutation) by

$$\boldsymbol{x}_\star = \frac{1}{e^\epsilon + K - 1}\begin{bmatrix} e^\epsilon & 1 & \dots & 1 \end{bmatrix}.$$

Proof Let us first ascertain that the optimization domain is convex. Consider any two vectors $\boldsymbol{x}^{(1)}$ and $\boldsymbol{x}^{(2)}$ that belong to said domain. Any convex combination $\lambda\boldsymbol{x}^{(1)}+(1-\lambda)\boldsymbol{x}^{(2)}$ will also belong to this domain, because

$$\frac{\lambda x_k^{(1)} + (1-\lambda)x_k^{(2)}}{\lambda x_{k'}^{(1)} + (1-\lambda)x_{k'}^{(2)}} \le \max\left\{\frac{x_k^{(1)}}{x_{k'}^{(1)}}, \frac{x_k^{(2)}}{x_{k'}^{(2)}}\right\} \le e^\epsilon.$$

Since problem (93) is the maximization of a convex objective over a convex optimization domain, we infer that the maximizer lies on the domain boundary, that is, the privacy constraint is satisfied with equality. This means that we can restrict the privacy constraint to retain only those vectors satisfying

$$\max_{k,k'} \frac{x_k}{x_{k'}} = e^\epsilon.$$

Given this equality constraint and the symmetry (permutation invariance) of the objective function and optimization constraints, we can now conveniently parametrize the optimization domain as follows, without loss of generality:

$$\boldsymbol{x} = \frac{\begin{bmatrix} e^{\epsilon} & 1 & e^{\lambda_3 \epsilon} & \dots & e^{\lambda_K \epsilon} \end{bmatrix}}{e^{\epsilon} + 1 + e^{\lambda_3 \epsilon} + \ldots + e^{\lambda_K \epsilon}}. \tag{94}$$

where $\lambda_3, \ldots, \lambda_K \in [0,1]^{K-2}$ are the parameters left to optimize. Let us consider any single one of them, with index $k \in \{3, \ldots, K\}$, and study the maximum of the squared Euclidean norm of (94) as a function of λ_k, which we define as a function[17]

$$\begin{aligned} f(e^{\lambda_k \epsilon}) &\triangleq \frac{e^{2\epsilon} + 1 + e^{2\lambda_3 \epsilon} + \ldots + e^{2\lambda_K \epsilon}}{\left(e^{\epsilon} + 1 + e^{\lambda_3 \epsilon} + \ldots + e^{\lambda_K \epsilon}\right)^2} \\ &= \frac{A + e^{2\lambda_k \epsilon}}{\left(B + e^{\lambda_k \epsilon}\right)^2} \end{aligned}$$

where the constants A and B are defined as

$$\begin{aligned} A &\triangleq e^{2\epsilon} + 1 + \sum_{\substack{k'=3 \\ k' \neq k}}^{K} e^{2\lambda_{k'} \epsilon} \\ B &\triangleq e^{\epsilon} + 1 + \sum_{\substack{k'=3 \\ k' \neq k}}^{K} e^{\lambda_{k'} \epsilon} \end{aligned}$$

and have a ratio A/B upper and lower-bounded as

$$1 \leq \frac{A}{B} \leq e^{\epsilon}.$$

Note that we also have $A \leq B^2$. The function f is differentiable on $\mathbb{R}^+$ and has a derivative

$$f'(x) = \frac{\mathrm{d}}{\mathrm{d}x} \left\{ \frac{A + x^2}{(B + x)^2} \right\} = \frac{2(Bx - A)}{(B + x)^3}$$

which is negative for $x < A/B$ and positive for $x > A/B$. It follows that f is quasi-convex on $\mathbb{R}_+$, and thus in particular on $[1, e^{\epsilon}]$, so its maximum is attained at either one of the boundary points, i.e.,

$$\max_{1 \leq x \leq e^{\epsilon}} f(x) = \max\{f(1), f(e^{\epsilon})\}.$$

It follows that the maximizer $\boldsymbol{x}_\star$ of (93) can be represented in the parametric form (94) in which some number κ of parameters λ_k are set to zero, while the remaining $K - 2 - \kappa$ parameters are set to one. That is, up to a permutation, the optimal $\boldsymbol{x}_\star$ has the form

$$\boldsymbol{x}_\star(\kappa) = \frac{\begin{bmatrix} e^{\epsilon} & e^{\epsilon} & \dots & e^{\epsilon} & 1 & 1 & \dots & 1 \end{bmatrix}}{(K - \kappa - 1)e^{\epsilon} + \kappa + 1}$$

17. The function f depends on the other parameters $\lambda_{k'}$, but we omit this in notation.

where the vector in the numerator contains $K-\kappa-1$ entries equal to e^ϵ and $\kappa+1$ entries equal to one, and the optimal value of $\kappa \in \{0,\dots,K-2\}$ is yet to be determined. Hence, the optimum of (93) is given by

$$\begin{aligned}\max_{\substack{\boldsymbol{x}\in\mathbb{R}_+^K:\\ \|\boldsymbol{x}\|_1=1\\ \forall k,k':\,\frac{x_k}{x_{k'}}\le e^\epsilon}} \|\boldsymbol{x}\|_2^2 &= \max_{\kappa\in\{0,\dots,K-2\}} \|\boldsymbol{x}_\star(\kappa)\|_2^2 \\ &= \max_{\kappa\in\{0,\dots,K-2\}} \frac{\kappa+1+(K-\kappa-1)e^{2\epsilon}}{(\kappa+1+(K-\kappa-1)e^\epsilon)^2} \\ &= \max_{\kappa'\in\{1,\dots,K-1\}} \zeta(\kappa'). \end{aligned} \tag{95}$$

where the function ζ is defined as

$$\zeta(\kappa') = \frac{\kappa'+(K-\kappa')e^{2\epsilon}}{(\kappa'+(K-\kappa')e^\epsilon)^2}.$$

By considering the continuous extension of ζ to the interval $[0,K]$ and studying the sign of its derivative

$$\frac{\mathrm{d}\zeta}{\mathrm{d}\kappa'} = \frac{(e^\epsilon-1)^2((K-\kappa')e^\epsilon-\kappa')}{(\kappa'+(K-\kappa')e^\epsilon)^3}$$

we conclude that ζ is quasi-concave on the interval $[0,K]$ with a maximum located at

$$\frac{e^\epsilon K}{e^\epsilon+1} \in (0,K).$$

Consequently, the maximum (95) is attained at either $\kappa'=1$ or $\kappa'=K-1$, i.e.,

$$\begin{aligned}\max_{\substack{\boldsymbol{x}\in\mathbb{R}_+^K:\\ \|\boldsymbol{x}\|_1=1\\ \forall k,k':\,\frac{x_k}{x_{k'}}\le e^\epsilon}} \|\boldsymbol{x}\|_2^2 &= \max\{\zeta(1),\zeta(K-1)\} \\ &= \max\left\{\frac{1+(K-1)e^{2\epsilon}}{(1+(K-1)e^\epsilon)^2}, \frac{K-1+e^{2\epsilon}}{(K-1+e^\epsilon)^2}\right\}.\end{aligned}$$

To compute this maximum of two fractions, consider the cross-product of numerators and denominators, which can be factorized as follows:

$$\begin{aligned}\big(K-1+e^{2\epsilon}\big)(1+(K-1)e^\epsilon)^2 - \big(1+(K-1)e^{2\epsilon}\big)(K-1+e^\epsilon)^2 \\ = (K-1)(K-2)(e^{2\epsilon}-1)(e^\epsilon-1)^2.\end{aligned}$$

From the right-hand side of the last equality, it appears that this expression is non-negative, hence we conclude that

$$\max_{\substack{\boldsymbol{x}\in\mathbb{R}_+^K:\\ \|\boldsymbol{x}\|_1=1\\ \forall k,k':\,\frac{x_k}{x_{k'}}\le e^\epsilon}} \|\boldsymbol{x}\|_2^2 = \frac{K-1+e^{2\epsilon}}{(K-1+e^\epsilon)^2} = \zeta(K-1)$$

which, up to an arbitrary permutation, is attained by

$$\boldsymbol{x}^\star = \boldsymbol{x}^\star(K-2) = \frac{\begin{bmatrix}e^\epsilon & 1 & 1 & \dots & 1\end{bmatrix}}{e^\epsilon+K-1}.$$

This concludes the proof. ∎

Appendix K. Proof of Lemma 11

Let the columns of $\boldsymbol{W}$ be denoted as (row vectors) $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_K$ and the rows of its inverse $\boldsymbol{W}^{-1}$ be denoted as $\tilde{\boldsymbol{w}}_1, \ldots, \tilde{\boldsymbol{w}}_K$. By definition of the inverse, we have

$$\left\langle \boldsymbol{w}_{k'}, \tilde{\boldsymbol{w}}_k \right\rangle = \begin{cases} 0, & \text{if } k \neq k' \\ 1, & \text{if } k = k'. \end{cases} \tag{96}$$

Let

$$\cos(\boldsymbol{a}, \boldsymbol{b}) = \frac{\langle \boldsymbol{a}, \boldsymbol{b} \rangle}{\|\boldsymbol{a}\|_2 \|\boldsymbol{b}\|_2}$$

denote the cosine of the angle enclosed by vectors $\boldsymbol{a}$ and $\boldsymbol{b}$. For any orthogonal basis $(\boldsymbol{b}_1, \ldots, \boldsymbol{b}_K)$ of $\mathbb{R}^K$ and an arbitrary vector $\boldsymbol{a}$, we have the relationship

$$\sum_{k=1}^{K} \cos(\boldsymbol{a}, \boldsymbol{b}_k)^2 = 1.$$

For any $k \neq k'$, since $\boldsymbol{w}_{k'}$ and $\tilde{\boldsymbol{w}}_k$ are orthogonal by (96), it follows that

$$\cos(\boldsymbol{w}_k, \boldsymbol{w}_{k'})^2 + \cos(\boldsymbol{w}_k, \tilde{\boldsymbol{w}}_k)^2 \leq 1. \tag{97}$$

The first squared cosine in (97) can be lower-bounded as follows: Due to the privacy constraint $\boldsymbol{W} \in \mathscr{W}_\epsilon$ and the row-stochasticity of $\boldsymbol{W}$, the vectors $\boldsymbol{w}_k$ can be represented in parametric form as

$$\boldsymbol{w}_k = \frac{\|\boldsymbol{w}_k\|_2}{\sqrt{\sum_{\ell=1}^{K} e^{2\epsilon\lambda_{\ell,k}}}} \begin{bmatrix} e^{\epsilon\lambda_{1,k}} & \ldots & e^{\epsilon\lambda_{K,k}} \end{bmatrix} \tag{98}$$

where the coefficients $\lambda_{\ell,k}$ belong to the unit interval $[0, 1]$. Any two distinct column vectors $\boldsymbol{w}_k$ and $\boldsymbol{w}_{k'}$ span an angle of cosine at least

$$\begin{aligned} \cos(\boldsymbol{w}_k, \boldsymbol{w}_{k'}) &= \frac{\langle \boldsymbol{w}_k, \boldsymbol{w}_{k'} \rangle}{\|\boldsymbol{w}_k\|_2 \|\boldsymbol{w}_{k'}\|_2} \\ &= \frac{\sum_{\ell=1}^{K} e^{\epsilon(\lambda_{\ell,k} + \lambda_{\ell,k'})}}{\sqrt{\sum_{\ell=1}^{K} e^{2\epsilon\lambda_{\ell,k}} \sum_{\ell'=1}^{K} e^{2\epsilon\lambda_{\ell',k'}}}} \\ &\geq e^{-2\epsilon}. \end{aligned} \tag{99}$$

Combining (97) and (99), we obtain

$$\begin{aligned} \cos\big(\boldsymbol{w}_k, \tilde{\boldsymbol{w}}_k\big)^2 &= \frac{1}{\|\boldsymbol{w}_k\|_2^2 \|\tilde{\boldsymbol{w}}_k\|_2^2} \\ &\leq 1 - e^{-4\epsilon}. \end{aligned}$$

This inequality allows us to establish the following lower bound:

$$\begin{aligned}\sum_{(k,\ell)\in[K]^2}\Phi_{k,\ell}(\boldsymbol{W}) &= \sum_{k=1}^{K}\sum_{\ell=1}^{K}\sum_{m=1}^{K} W_{k,m}\tilde{W}_{m,\ell}^2\\ &= \sum_{m=1}^{K}\sum_{k=1}^{K} W_{k,m}\sum_{\ell=1}^{K}\tilde{W}_{m,\ell}^2\\ &= \sum_{m=1}^{K}\|\boldsymbol{w}_m\|_1\|\tilde{\boldsymbol{w}}_m\|_2^2\\ &\geq \frac{1}{1-e^{-4\epsilon}}\sum_{m=1}^{K}\frac{\|\boldsymbol{w}_m\|_1}{\|\boldsymbol{w}_m\|_2^2}.\end{aligned} \tag{100}$$

As a final step, when minimizing the right-hand side of (100) over privatization channels $\boldsymbol{W}\in\mathscr{W}_\epsilon$, we will show in the following that the minimum is attained for a step-circulant mechanism

$$\frac{1}{e^\epsilon+K-1}\begin{bmatrix} e^\epsilon & 1 & \dots & 1\\ 1 & e^\epsilon & \ddots & \vdots\\ \vdots & \ddots & \ddots & 1\\ 1 & \dots & 1 & e^\epsilon\end{bmatrix}.$$

Hence,

$$\min_{\boldsymbol{W}\in\mathscr{W}_\epsilon}\sum_{m=1}^{K}\frac{\|\boldsymbol{w}_m\|_1}{\|\boldsymbol{w}_m\|_2^2} = K\frac{(e^\epsilon+K-1)^2}{e^{2\epsilon}+K-1} \tag{101}$$

which upon recombining with (100) yields the desired statement of Lemma 11.

The statement (101) will be proven in what follows. To begin with, consider the relaxation of problem (101) that one obtains when minimizing over a superset $\overline{\mathscr{W}}_\epsilon\supset\mathscr{W}_\epsilon$ defined as

$$\overline{\mathscr{W}}_\epsilon \triangleq \Big\{\boldsymbol{W}\in\mathbb{R}_+^{K\times K}\colon \textstyle\sum_{k,\ell}W_{k,\ell}=K \text{ and } \forall i,j,j'\colon W_{i,j,j'}\leq e^\epsilon W_{i,j'}\Big\}.$$

In other words, we relax the row-stochasticity constraint while keeping the sum of all entries equal to K. This relaxed problem can be rewritten as follows:

$$\min_{\boldsymbol{W}\in\overline{\mathscr{W}}_\epsilon}\sum_{m=1}^{K}\frac{\|\boldsymbol{w}_m\|_1}{\|\boldsymbol{w}_m\|_2^2} = \min_{\substack{(N_1,\dots,N_K)\in\mathbb{R}_+^K\colon\\ N_1+\dots+N_K=K}}\sum_{m=1}^{K}\min_{\substack{\boldsymbol{w}_m\in\mathbb{R}_+^K\colon\\ \|\boldsymbol{w}_m\|_2=N_m\\ \forall k,k'\colon W_{k,m}/W_{k',m}\leq e^\epsilon}}\frac{N_m}{\|\boldsymbol{w}_m\|_2^2}. \tag{102}$$

We now need to solve the inner minimization problem (for any given m) on the right-hand side of (102), whose minimizer can be equivalently expressed as the maximizer

$$\boldsymbol{w}_\star = \operatorname*{argmax}_{\substack{\boldsymbol{w}\in\mathbb{R}_+^K\colon\\ \|\boldsymbol{w}\|_1=N_m\\ \forall k,k'\colon w_k/w_{k'}\leq e^\epsilon}}\|\boldsymbol{w}\|_2^2$$

which according to Lemma 13 in Appendix J admits the solution (up to a permutation)

$$\boldsymbol{w}_\star = N_m \frac{\begin{bmatrix} e^\epsilon & 1 & 1 & \dots & 1 \end{bmatrix}}{e^\epsilon + K - 1}.$$

The reciprocal of the squared Euclidean norm of $\boldsymbol{w}_\star$ equals

$$\frac{1}{\|\boldsymbol{w}_\star\|_2^2} = \frac{1}{N_m^2} \frac{(e^\epsilon + K - 1)^2}{e^{2\epsilon} + K - 1}.$$

Hence, the relaxed optimization problem considered further above can be written as

$$\begin{aligned} \min_{\boldsymbol{W} \in \overline{\mathscr{W}}_\epsilon} \sum_{m=1}^K \frac{1}{\|\boldsymbol{w}_m\|_2} &= \min_{\substack{(N_1,\dots,N_K) \in \mathbb{R}_+^K : \\ N_1 + \dots + N_K = K}} \sum_{m=1}^K \frac{1}{N_m} \frac{K - 1 + e^\epsilon}{\sqrt{K - 1 + e^{2\epsilon}}} \\ &= K \frac{K - 1 + e^\epsilon}{\sqrt{K - 1 + e^{2\epsilon}}}. \end{aligned}$$

The latter expression is also the minimum of the original optimization problem (before relaxation) and can be achieved by picking $\boldsymbol{w}_m$ to be the rows of the step-circulant matrix

$$\boldsymbol{W}_{\epsilon,\star} = \frac{1}{e^\epsilon + K - 1} \begin{bmatrix} e^\epsilon & 1 & \dots & 1 \\ 1 & e^\epsilon & \ddots & \vdots \\ \vdots & \ddots & \ddots & 1 \\ 1 & \dots & 1 & e^\epsilon \end{bmatrix}.$$

as defined in (23), which establishes (101) and thus finalizes the proof of Lemma 11.

References

Syuuji Abe. Expected relative entropy between a finite distribution and its empirical distribution. *SUT Journal of Mathematics*, 32(2), 1996.

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Hadamard response: Estimating distributions privately, efficiently, and with little communication. volume 89 of *Proceedings of Machine Learning Research*, pages 1120–1129. PMLR, 16–18 Apr 2019. URL `http://proceedings.mlr.press/v89/acharya19a.html`.

Rakesh Agrawal and Ramakrishnan Srikant. Privacy-preserving data mining. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, pages 439–450, New York, NY, USA, 2000.

Patrick Billingsley. *Probability and Measure*. John Wiley and Sons, third edition, 1995.

Imre Csiszár. Sanov property, generalized I-projection and a conditional limit theorem. *The Annals of Probability*, 12(3):768–793, 1984.

Anindya De. Lower bounds in differential privacy. In Ronald Cramer, editor, *Theory of Cryptography*, pages 321–338, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Applications of mathematics. Springer, 1998.

John C. Duchi, Martin J. Wainwright, and Michael I. Jordan. Local privacy and minimax bounds: Sharp rates for probability estimation. In *Advances in Neural Information Processing Systems (NIPS) 26*, pages 1529–1537. 2013.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Privacy aware learning. *Journal of the ACM*, 61(6):38:1–38:57, December 2014.

Cynthia Dwork. Differential privacy. In *Automata, Languages and Programming*, Lecture Notes in Computer Science, pages 1–12. Springer, Berlin, Heidelberg, July 2006.

Cynthia Dwork. Differential privacy: A survey of results. In *Theory and Applications of Models of Computation*, Lecture Notes in Computer Science, pages 1–19. Springer Berlin Heidelberg, April 2008.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography*, Lecture Notes in Computer Science, pages 265–284. Springer, Berlin, Heidelberg, March 2006.

Úlfar Erlingsson, Vasyl Pihur, and Aleksandra Korolova. RAPPOR: Randomized aggregatable privacy-preserving ordinal response. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 1054–1067, New York, NY, USA, 2014.

Alexandre Evfimievski, Johannes Gehrke, and Ramakrishnan Srikant. Limiting privacy breaches in privacy preserving data mining. In *Proceedings of the Twenty-Second ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, page 211–222, New York, NY, USA, 2003. Association for Computing Machinery.

Marco Gaboardi and Ryan Rogers. Local private hypothesis testing: Chi-square tests. *arXiv:1709.07155 [cs, math, stat]*, September 2017. URL `http://arxiv.org/abs/1709.07155`. arXiv: 1709.07155.

Chong Huang, Peter Kairouz, Xiao Chen, Lalitha Sankar, and Ram Rajagopal. Context-aware generative adversarial privacy. *Entropy*, 19(12):656, December 2017.

Zhanglong Ji, Zachary C. Lipton, and Charles Elkan. Differential privacy and machine learning: a survey and review. December 2014. URL `http://arxiv.org/abs/1412.7584`.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Extremal mechanisms for local differential privacy. In *Advances in Neural Information Processing Systems (NIPS) 27*, pages 2879–2887. 2014.

Peter Kairouz, Sewoong Oh, and Pramod Viswanath. Secure multi-party differential privacy. In *Advances in Neural Information Processing Systems (NIPS) 28*, pages 2008–2016. 2015.

Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48, pages 2436–2444, New York, USA, June 2016a.

Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. *arXiv:1602.07387 [cs, stat]*, February 2016b. URL `http://arxiv.org/abs/1602.07387`.

Shiva Prasad Kasiviswanathan, Homin K. Lee, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. What can we learn privately? *SIAM Journal on Computing*, 40(3):793–826, January 2011.

David Leoni. Non-interactive differential privacy: A survey. In *Proceedings of the First International Workshop on Open Data (WOD)*, pages 40–52, New York, NY, USA, 2012. ACM.

Minnesota Population Center. Integrated Public Use Microdata Series, International: Version 7.2 [dataset]. Minneapolis, MN: IPUMS. `https://doi.org/10.18128/D020.V7.2`, 2019. The underlying data used in this article was provided by the Institute of Geography and Statistics, Brazil.

Ajaykrishnan Nageswaran and Prakash Narayan. Distribution privacy under function recoverability. In *IEEE International Symposium on Information Theory (ISIT)*, pages 890–895, 2020.

Adriano Pastore and Michael Gastpar. Locally differentially-private distribution estimation. In *IEEE International Symposium on Information Theory (ISIT)*, pages 2694–2698, 2016.

Anand D. Sarwate and Lalitha Sankar. A rate-distortion perspective on local differential privacy. In *52nd Annual Allerton Conference on Communication, Control, and Computing*, pages 903–908, September 2014.

Salil Vadhan. *The Complexity of Differential Privacy*, pages 347–450. Springer International Publishing, Cham, 2017.

Martin J. Wainwright, Michael I. Jordan, and John C. Duchi. Privacy aware learning. In *Advances in Neural Information Processing Systems (NIPS) 25*, pages 1430–1438. 2012.

Ning Wang, Xiaokui Xiao, Yin Yang, Jun Zhao, Siu Hui, Hyejin Shin, Junbum Shin, and Ge Yu. Collecting and analyzing multidimensional data with local differential privacy. 04 2019.

Shaowei Wang, Liusheng Huang, Pengzhan Wang, Yiwen Nie, Hongli Xu, Wei Yang, Xiang-Yang Li, and Chunming Qiao. Mutual information optimally local private discrete distribution estimation. *arXiv:1607.08025 [cs, math]*, July 2016. URL `http://arxiv.org/abs/1607.08025`.

Stanley L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.

Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

Min Ye and Alexander Barg. Asymptotically optimal private estimation under mean square loss. *arXiv:1708.00059 [cs, math, stat]*, July 2017. URL `http://arxiv.org/abs/1708.00059`.

Min Ye and Alexander Barg. Optimal schemes for discrete distribution estimation under locally differential privacy. *IEEE Transactions on Information Theory*, 64(8):5662–5676, August 2018.

Min Ye and Alexander Barg. Optimal locally private estimation under ℓ_p loss for $1 \leq p \leq 2$. *Electronic Journal of Statistics*, 13(2):4102–4120, 2019.