Instance-Optimality for Private KL Distribution Estimation

Jiayuan Ye[*]
National University of Singapore

Vitaly Feldman
Apple

Kunal Talwar
Apple

Abstract

We study the fundamental problem of estimating an unknown discrete distribution p over d symbols, given n i.i.d. samples from the distribution. We are interested in minimizing the KL divergence between the true distribution and the algorithm's estimate. We first construct minimax optimal private estimators. Minimax optimality however fails to shed light on an algorithm's performance on individual (non-worst-case) instances p and simple minimax-optimal DP estimators can have poor empirical performance on real distributions. We then study this problem from an instance-optimality viewpoint, where the algorithm's error on p is compared to the minimum achievable estimation error over a small local neighborhood of p. Under natural notions of local neighborhood, we propose algorithms that achieve instance-optimality up to constant factors, with and without a differential privacy constraint. Our upper bounds rely on (private) variants of the Good-Turing estimator. Our lower bounds use additive local neighborhoods that more precisely captures the hardness of distribution estimation in KL divergence, compared to ones considered in prior works.

1 Introduction

Accurately estimating a discrete distribution (over d symbols) from the empirical samples is a fundamental task in statistical machine learning. An especially important distribution estimation objective is the Kullback-Leibler (KL) divergence error, as it is crucial in promoting diversity and smoothness by penalizing zero-mass assignment to unseen symbols. Noticeably in speech-recognition and language modeling communities, negative log likelihood on a test set (which up to translation is the KL divergence) is the measure that has been found to best correlate with the performance of the model [13, 14], and is therefore the standard loss being optimized for. Additionally, KL error is well-established in the coding community [18, 42] in the context of data compression (i.e., identifying encoding that represents the information with fewer bits). Due to this reason, KL distribution estimation has been extensively studied [62, 11, 10, 51, 52, 39, 49].

However, overly fine-grained release of statistics make it possible to infer the memberships [37, 55, 28, 56] or even reconstruct the values [20, 12] of input data records. To this end, differential privacy [26] offers a powerful mathematical definition to provably control the associated privacy risk.

Definition 1.1 (Differential Privacy (DP) [26])**.** A (randomized) algorithm $\mathcal{A}$ is (ε, δ)-differentially private ((ε, δ)-DP) if for all neighboring datasets x, x' that differ in at most one sample, and any measurable output set T, $\Pr[\mathcal{A}(x) \in T] \leq e^{\varepsilon} \cdot \Pr[\mathcal{A}(x') \in T] + \delta$.

Our work aims to provide an understanding of the KL error of private distribution estimation. Below we present the problem setting and summarize known results and our contributions.

[*]Research done while at Apple

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

Problem Setting Let $p \in \Delta(d) \coloneqq \{p \in \mathbb{R}^d : p_1, \cdots, p_d \geq 0 \text{ and } \sum_{i=1}^d p_i = 1\}$ be an unknown distribution over d symbols. Let $x \in \mathbb{N}^d$ be the histogram representation of a dataset consisting of empirical samples drawn from distribution p, where x_i denotes the count of symbol i in the dataset. Following prior works [2, 49], we assume that the count of each symbol i is independently drawn from a Poisson distribution with mean np_i, i.e., $(x_1, \cdots, x_d) \sim \text{Poi}(np) \coloneqq \text{Poi}(np_1) \times \cdots \times \text{Poi}(np_d)$. This assumption is a convenient choice for analysis as it ensures independent sampling of each symbol, while enjoying the benefit of being equivalent to sampling from multinomial distribution when conditioned on a fixed dataset size n. Our goal is to design an algorithm that accurately estimates the unknown distribution p in KL divergence given a sampled dataset $x \sim \text{Poi}(np)$.

Minimax Optimality Results (and Their Limitations) We start by looking at the (private) KL distribution estimation problem from the standard minimax optimality objective, where the goal is to design estimators $\mathcal{A}$ that minimizes the KL error on the *worst-case* distribution instance.

$$\min_{\mathcal{A}} \max_{p \in \Delta(d)} \mathbb{E}\left[\text{KL}(p, \mathcal{A}(x))\right], \text{KL}(p \| \mathcal{A}(x)) = \sum_i p_i \log \frac{p_i}{\mathcal{A}(x)_i} \tag{1}$$

where $\text{KL}(p\|\mathcal{A}(x)) = \sum_i p_i \log \frac{p_i}{\mathcal{A}(x)_i}$, and the expectation is over the randomness of estimator $\mathcal{A}$ and over the sampling of $x \sim \text{Poi}(np)$. This minimax objective (1) is well-studied for non-DP algorithms, where simple add-constant estimators are proved to be minimax optimal with $O\left(\ln\left(1 + \frac{d}{n}\right)\right)$ rates [10, 52] (see Appendix C.1 for a complete discussion of the existing results). For the DP setting, to the best of our knowledge, there are no known results for the KL minimax rates. Nevertheless, in Appendix C.2, we show that a similarly simple algorithm (that truncates the Laplace perturbations of empirical counts) achieves the minimax optimal $O\left(\ln\left(1 + \frac{d}{\varepsilon n}\right)\right)$ rates.

However, such simple estimators achieve poor performance in experiments (Section 4) on commonly occurring distributions such as power-law distributions. Similar observations, i.e., the poor performance of the minimax-optimal add-constant estimator (compared to the practical Good-Turing [33] estimator), have long existed in the non-DP setting. This is intuitively because minimax optimality only captures the *worst-case* error over all possible distributions, thus failing to indicate whether an algorithm performs well for each *non-worst-case* distribution p.

Instance-optimality Instance-optimality is a promising framework to address the above limitation of minimax optimality and shed light on the per-instance provable performance of an estimator. In the non-DP setting, many works have studied instance-optimality in different contexts, such as local minimax estimation [21, 47, 23], competitive distribution estimation [2, 49] or instance-by-instance analysis [59]. Remarkably, the seminal work of [49] proved that a simple variant of Good-Turing estimators is nearly instance-optimal, in that it estimates every distribution nearly as well as the best estimator designed with prior knowledge of the distribution up to a permutation. The recent work by Feldman, McMillan, Sivakumar and Talwar [30], further generalize this instance-optimality definition to other natural definitions of prior knowledge in the language of per-instance neighborhood. Formally, we follow Feldman, McMillan, Sivakumar and Talwar [30] and define instance-optimality as follows.

Definition 1.2 (Instance-Optimality to Neighborhood Map N [30])**.** We say an estimator $\mathcal{A}$ is instance-optimal with respect to a neighborhood map N if:

$$\begin{aligned} &\forall\, p \in \Delta(d) : \mathbb{E}[\text{KL}(p, \mathcal{A}(x))] \leq O\left(\text{lower}(p, n, N)\right) \\ &\quad \text{where } \text{lower}(p, n, N) \overset{def}{=} \min_{\mathcal{A}'} \max_{q \in N(p)} \mathbb{E}\left[\text{KL}(q, \mathcal{A}'(x))\right] \text{ for a neighborhood } N(p) \text{ of } p. \end{aligned} \tag{2}$$

That is, instance-optimality says that the algorithm $\mathcal{A}$ is competitive with any hypothetical algorithm $\mathcal{A}'$ that has auxiliary knowledge about the neighborhood $N(p)$ of the input distribution p. We can further constrain the algorithm to be (ε, δ)-DP in establishing the per-instance lower bound as follows.

$$\text{lower}_{\varepsilon,\delta}(p, n, N) = \min_{\mathcal{A} \text{ is } (\varepsilon,\delta)\text{-DP}} \max_{q \in N(p)} \mathbb{E}\left[\text{KL}(q, \mathcal{A}(x))\right] \tag{3}$$

This is the lower bound that we will use for establishing instance-optimality of private algorithms.

Limitations of Prior Instance-Optimality Objectives for KL Error In instance-optimality results, the smaller the neighborhood, the stronger the result. This is because we are proving that our algorithm is competitive with a hypothetical algorithm that already knows that the input distribution is in the neighborhood. However, if the neighborhood is too small, then no algorithm can be instance-optimal, as the baseline estimator's knowledge of the instance is overly precise. (As an extreme example, when $N(p) = \{p\}$ contains only the target distribution p, the per-instance lower bounds (2) and (3) trivially become zero (as achieved by an estimator that always outputs $\mathcal{A}(x) = p$ regardless of the input dataset).) Below we review existing choices of neighborhood $N(p)$ in the literature, and show they are either too large or too small to accurately capture the per-instance estimation hardness in KL.

- Permutation neighborhood [2, 49, 59]: $N_\pi(p) = \{q : \exists \text{ permutation } \pi \text{ s.t. } q_i = p_{\pi \circ i}, \forall i \in [d]\}$, i.e., all distributions obtained by permuting the probability values. On the one hand, prior work [30] has argued that this neighborhood is too large, as practical estimators often have stronger knowledge about which symbols are more frequent than others (e.g., the word "and" is often more frequent than "differential", thus assuming the two words are permutable is not reasonable). On the other hand, permutation neighborhood can be too small, in that it provably does not allow for multiplicative instance optimality guarantees, as discussed in Section 2.2. Indeed, permutation neighborhood assumes precise knowledge of the set of true probability values, which can be overly strong for any realistic estimator to satisfy.
- Two-point neighborhood [23, 47, 8]: $N_2(p) = \{p, q\}$ contains the target instance p and another alternative distribution q that is "close" to p. This neighborhood reduces the per-instance hardness to binary hypothesis testing of whether the observed samples are from p or q. Under appropriate notions of "close", this neighborhood allows matching per-instance upper and lower bounds for estimating one-parameter (exponential) families distribution under central DP [47] and local DP [23], as well as for general one-dimensional statistical estimation problems [8]. However, for high-dimensional problems, this neighborhood is provably too small, as the per-instance lower bound remain constant under growing data dimension, despite the growing hardness for distribution estimation under higher dimension. For example, there exist instances (such as long-tailed instances in our particular distribution estimation problem) whose estimation error inevitably grows with data dimension (e.g., of scale $\ln d/16$ in our Theorem G.9 construction), indicating that the two-point neighborhood and multiplicative neighborhood are too small. Indeed testing can be provably easier than learning in simple settings.
- Multiplicative neighborhood [30]: $N_\times(p) = \{q : \frac{1}{2}p_i \le q_i \le 2p_i, \forall i \in [d]\}$. It allows matching per-instance upper and lower bounds for distribution estimation in Wasserstein distance. However, this neighborhood is too small to capture the KL error: the per-instance lower bound is at most a constant (since an algorithm that is tailored to $N(p)$ can achieve $O(1)$ error), whereas such an error bound is provably unachievable for some distributions (e.g., our Theorem G.9 construction).

Besides the above three representative types of neighborhoods, earlier works [7, 6] on instance-optimal DP estimation also considered other variants of neighborhoods that have similar limitations. [7] only deals with one-dimensional quantities, and [6] discusses high-dimensional problems defined on the dataset (rather than on the distribution as in our work). The notion of local minimax optimality there includes all datasets at a certain distance, whereas their second notion only competes with unbiased algorithms. These limitations call for new definitions of local neighborhood to precisely capture the per-instance hardness of distribution estimation in KL divergence, which we study in this paper.

Our Main Contributions: Instance-Optimal Results

- We define instance-optimality objectives that more precisely capture the hardness of private KL distribution estimation. This is by constructing new neighborhoods (4) and (6) that additively perturb the probability values p_i of individual symbols. We use small perturbation scales – $1/n$, $1/n\epsilon$, and $\sqrt{p_i/n}$ – to ensure the dataset (sampled by p) could plausibly have come from other distribution in the neighborhood. (See Section 2 and 3 for more rationales on the neighborhood sizes.) This is stronger than permutation neighborhood [2, 49, 59] in allowing auxiliary knowledge of frequency order among symbols, and is thus a very strong notion of instance-optimality. Furthermore, we show in Section 3.1 that our

additive neighborhoods are (up to constants) the smallest that still allow for instance-optimal algorithms.

- Under such additive neighborhoods, we propose a new non-DP algorithm and prove it to be instance-optimal (Section 2.1). Our algorithm resembles the idea of the Good-Turing estimator (which is known to be instance-optimal in prior works), while ensuring significantly smaller sensitivity to adjacent datasets (in the DP sense). This reduced sensitivity is the key that makes our algorithm easier to privatize. We then propose a DP version of this algorithm and show that it is instance-optimal amongst DP algorithms (Section 3.1).
- We validate the performance of our instance-optimal estimators via experiments (Section 4), and show the reward from studying instance-optimality: while the Add-constant (DP) algorithm is already minimax-optimal, our instance-optimal algorithms achieve significantly better performance on many instances of practical interest, such as power-law distributions and real-word token distributions.

1.1 Technical Contributions

Generalized Assouad's method for decomposable statistical distance Standard tools for proving lower bounds, such as (DP) Assouad's method [65, 3, 30], only apply to symmetric statistical distance, which is not satisfied by KL divergence. To prove KL per-instance lower bound, we propose generalized (DP) Assouad's method (Theorem A.3 and A.4) that applies to general *decomposable statistical distance* (Definition A.1). This allows us to prove strong per-instance lower bounds.

Reducing the Sensitivity of Good-Turing Estimator via "Sampling Twice" The challenge of privatizing the prior (near) instance-optimal Good-Turing estimators lie in its excessively high sensitivity to neighboring datasets. At the core, Good-Turing estimator is motivated by observing that estimating "unseen" symbols as zero-probability is biased, as the symbols that appeared exactly once would intuitively have similar probability values (which are non-zero). To correct this bias, it recursively uses the counts of symbols with frequency $t+1$ to estimate the probability of symbols that appeared t times, for all low-frequency symbols (e.g., $t = 1, \cdots, n^{1/3}$ in Orlitsky and Suresh [49]). Such computations suffer from high sensitivity, as one record could change the combined counts of symbols with frequency t by $t = n^{1/3}$ in the worst-case. To address this limitation, we perform bias-correction via an alternative "sampling twice" approach: partitioning the dataset into two halves, using one for identifying "unseen" symbols and the other for estimating their combined mass. The resulting Algorithm 1 is conceptually simpler than prior Good-Turing estimators, while achieving tight instance-optimality guarantees (up to constants) and being empirically competitive in experiments. Crucially, this "Sampling Twice" design reduces sensitivity to just one (making the estimator easy to privatize), and is the key to achieving instance-optimality in the DP setting.

Effectively Privatizing Good-Turing Estimator via Calibrated Thresholding The Good-Turing estimator is based on the intuition that the probability values of symbols that appeared zero times (in the dataset) are similar to those of symbols that appeared exactly once. However, this simple zero-or-one thresholding does not remain effective for private algorithms, because DP estimates cannot differ significantly on neighboring input datasets. This forces us to use a larger threshold. To identify the best threshold for DP distribution estimation, given every possible threshold, we separately analyze (Theorem E.2 and G.6) the DP estimation error due to False Negatives (FN) (i.e., high-probability symbols being below-threshold) versus False Positives (FP) (i.e., low-probability symbols being above-threshold). This analysis allows us to choose a threshold that balances the FN and FP errors, and achieves DP instance-optimality up to a constant factor (Section 3.1).

1.2 Other Related Works

The Good-Turing estimator, originally developed by [33] and simplified by [31, 49], has been widely observed to yield empirically accurate distribution estimates (especially for language modeling tasks [40, 13] under large vocabulary size). Several later works prove it to be minimax optimal [46, 22, 50] as well as instance-optimal [2, 49] for discrete distribution estimation in various metrics. However, to our knowledge, no prior works have designed or analyzed DP variants of Good-Turing estimator, which is the main technical innovation of this paper.

For DP discrete distribution estimation, minimax optimality is well-studied across a variety of models for DP, including central DP [19, 3], local DP [24, 38, 64, 25, 1] and user-level DP [44]. Existing works also focus on different error metrics, such as total variation distance (i.e., ℓ_1 error) [19] and ℓ_2 error [3] (see a summary of results in [3, Table 1-2]). Our work falls into the central DP setting, and is the first to study the KL divergence error. Additionally, we establish the stronger instance-optimality (rather than minimax optimality). The closest work to our paper in the literature is [30], where the authors study instance-optimal density estimation in Wasserstein distance (that covers discrete distribution estimation in ℓ_1 error as a special case). The results in [30] are largely incomparable to our analysis, due to the lack of (tight) conversions between total variation distance error and KL divergence error. (See Appendix C.2 for a nuanced comparison between prior TV distance lower bounds and our KL lower bound for DP minimax distribution estimation.) Indeed, achieving optimality under KL often requires non-trivial change to the algorithm and analysis compared to ℓ_1 error, as evidenced by the abundant literature on distribution estimation in KL [10, 51, 52, 2, 49].

A closely related problem is frequency estimation (a.k.a. histogram estimation), where the goal is to privately and accurately estimate the empirical distribution. DP histogram estimation is extensively studied in a variety of models for DP, including central DP [34, 63], local DP [9, 61], and shuffle DP [29, 15, 32]. Due to sampling error, algorithms for empirical frequency estimation typically do not directly yield good utility for distribution estimation.

2 Tighter Instance-Optimality for Non-DP Estimation

We first define instance-optimality under additive neighborhoods, that get around the limitations of previously studied neighborhood notions as discussed in Section 1.

Neighborhood Choices As discussed in Section 1, a good neighborhood should be as small as possible and contain distributions that are similarly hard to estimate compared to the target distribution instance. Additive neighborhood is thus a natural choice as intuitively, the hardness of distribution estimation does not change a lot when perturbing each symbol's probability by a small amount. The key question is how small should the scale be. To this end, our main design choices are

1. For every symbol, we allow up to t/n perturbation, for a small $t \geq 1$. This only changes its expected count up to t, and thus from the the empirically sampled count, it is hard to distinguish the perturbed distribution from the target distribution.
2. For symbol with large $p_i > t/n$, we allow a larger $\sqrt{p_i/n}$ perturbation, for a small $t \geq 1$. This captures the fact that with constant probability, counts sampled from $\mathrm{Poi}(np_i)$ and $\mathrm{Poi}(np_i + \sqrt{np_i})$ are "indistinguishable" due to the statistical variance in sampling. Indeed these are related to the standard confidence intervals for binomial estimation [16, 4].

On top of these design choices, we try to reduce the size of the neighborhood as much as possible. Thus we add a constraint that the *combined mass* of small symbols (with $p_i \leq \frac{t}{n}$ given a small $t \geq 1$) should not change by more than $\frac{t}{n}$ – this added condition only makes the results stronger (as the resulting neighborhood is smaller). As a result, we obtain the following additive neighborhood.

$$N_+(p) = \left\{ q : \forall i \in [d],\, |q_i - p_i| \leq \min\left\{ \frac{t}{n}, \sqrt{\frac{p_i}{n}} \right\} \text{ and } \sum_{i:p_i \leq \frac{t}{n}} q_i \leq \max\left\{ \frac{t}{n}, \sum_{i:p_i \leq \frac{t}{n}} p_i \right\} \right\} \quad (4)$$

Per-Instance Lower Bound We next prove per-instance lower bound under this neighborhood $N_+(p)$. Our analysis requires generalized variants of Assouad's method (discussed in Section 1.1). We defer all proofs to Appendix D, and only present the per-instance lower bound below.

Theorem 2.1. *Let $p \in \Delta(d)$ be an arbitrarily fixed distribution instance. Let $N_+(p)$ be the additive neighborhood defined in* (4)*. If $d \geq 2$ and $n \geq 4$, then we have*

$$\textit{lower}(p, n, N_+) \geq \Omega\left(\frac{\ln(1 + d_{small}(L'))}{n} + p_{small}(L') \ln\left(1 + \frac{d_{small}(L')}{np_{small}(L')}\right) + \sum_i \min\left\{ p_i, \frac{1}{n} \right\} \right) \quad (5)$$

for any $t \geq 1$ and any set $L' \subseteq [d]$, where $p_{small}(L') = \sum_{i \in L', p_i \leq \frac{t}{n}} p_i$ and $d_{small}(L') = \sum_{i \in L', p_i \leq \frac{t}{n}} 1$.

	KL Error Bound	Reference
lower(p, n, N_π)	$\Omega\left(\sum_{t=0}^{\infty} \mathop{\mathbb{E}}_{x\sim\text{Poi}(np)}\left[\sum_{i:x_i=t} p_i \ln\left(\frac{p_i \sum_{j:x_j=t} 1}{\sum_{j:x_j=t} p_j}\right)\right]\right)$	[49, Lemma 4 & 5] N_π: permutation neighborhood
Add-constant	lower$(p, n, N_\pi) + \Omega\left(\min\left\{n^{-1/3}, \frac{d}{n}\right\}\right)$	[2, Lemma 1]
Good-Turing	lower$(p, n, N_\pi) + O\left(\min\left\{n^{-1/3}, \frac{d}{n}\right\}\right)$	[49, Theorem 1]
Smoothed Good-Turing	lower$(p, n, N_\pi) + O\left(\min\left\{n^{-1/2}, \frac{d}{n}\right\}\right)$	[49, Theorem 2] [2, Theorem 2]
Any estimator	lower$(p, n, N_\pi) + \Omega\left(\min\left\{n^{-2/3}, \frac{d}{n}\right\}\right)$	[49, Theorem 3]

Table 1: Prior instance-optimality results of Non-DP estimators

Algorithm 1 Non-DP "Sampling Twice"

Input: Data partition ratio $\alpha = 0.5$. Independently sampled datasets $x \sim \text{Poi}(\alpha \cdot np)$ and $x' \sim \text{Poi}((1-\alpha)\cdot np)$ s.t. $x + x' \sim \text{Poi}(np)$. Threshold $\tau = 0$.
Thresholding: $L = \{i \in [d] : x_i \le \tau\}$
Estimate combined mass for symbols in L: $\tilde{c} = \max\left\{\sum_{i\in L} x'_i, 1\right\}$
Truncate individual estimates: let $\tilde{x}_i = \max\{x'_i, 1\}$ for $i = 1, \cdots, d$
Return $\mathcal{A}(x)$ with $\mathcal{A}(x)_i = \begin{cases} \frac{1}{N}\cdot\tilde{c}\cdot\frac{\tilde{x}_i}{\sum_{i\in L}\tilde{x}_i} & i \in L \\ \frac{1}{N}\cdot\tilde{x}_i & i \notin L \end{cases}$ where $N = \tilde{c} + \sum_{i\notin L}\tilde{x}_i$

To interpret this lower bound, first observe that (5) is always smaller than the $\Omega\left(1+\frac{d}{n}\right)$ minimax lower bound, especially when there are many small symbols (with $p_i < \frac{1}{n}$) that jointly takes a small combined mass $p_{small}(L')$. As an extreme example, imagine a highly concentrated distribution $p = (1/3, 2/3, 0, \cdots, 0)$, then our per-instance lower bound (5) is as small as $\frac{\ln(1+d)}{n}$. This is significantly smaller than the $\ln(1+\frac{d}{n})$ minimax lower bound for high-dimensional setting ($d \to \infty$), thus correctly indicating that p is an extremely easy-to-estimate distribution. Achieving error upper bound that matches this small per-instance lower bound (5), then serves as a strong requirement for instance-optimal algorithm to fulfill (which we prove in the next section).

2.1 An Instance-Optimal "Sampling Twice" Estimator

We now present a simple "sampling twice" Algorithm 1, and prove that it is instance-optimal up to constants. Compared to the minimax optimal add-constant estimator, this algorithm observes that simply estimating "unseen" symbols as zero-probability (or constant-probability) is heavily biased, and use a conceptual "sampling-twice" procedure to correct the bias of low-frequency symbols.

As discussed in Section 1.1, this bias-correction idea originates from prior (near) instance-optimal Good-Turing algorithm [33, 31, 49] and maintains their utility benefit. The key benefit of our "sampling-twice" design lies in significantly reducing the estimator's sensitivity to neighboring dataset (making it easier to privatize), enabling instance-optimality under DP (as we will show in Section 3).

Instance-optimality Guarantee In Theorem E.2, we prove a per-instance KL error upper bounds for the "sampling twice" estimator (Algorithm 1). In Corollary E.3, we further prove that this upper bound can be rewritten as $\mathop{\mathbb{E}}_{x\sim\text{Poi}(np)}[\text{KL}(p, \mathcal{A}(x))] \le O\left(\text{lower}(p, n, N_+)\right)$ for any choice of neighborhood size $t \ge 1$ s.t. $t \cdot e^{-t} \le 1/\ln d$. Specifically, we can choose $t = \min\{1, 2\ln\ln d\}$.

2.2 Comparison to Prior Instance-Optimality Results

In this section, we discuss and compare with prior instance-optimality results, which mainly cover two representative non-DP estimators – the add-constant estimator and the Good-Turing (GT) estimator.

Add-constant Estimator Given by $\mathcal{A}(x) = x_i + c, \forall i \in [d]$ for a constant $c > 0$, the add-constant estimator is one of the oldest and simplest distribution estimator. Its variants cover several known estimators, such as Laplace smoothing ($c = 1$) and Krichevsky-Trofimov estimator [42] ($c = 1/2$).

Good-Turing Estimator [33] The Good-Turing estimator [33], when combined with add-constant estimator, has long been observed to yield strong empirical performance. Many variants of GT estimator exist in the literature [33, 31, 49]. In the comparison, we use the simplified variant of Good-Turing estimators proposed in Orlitsky and Suresh [49], that is also provably (near) instance-optimal under the "permutation neighborhood", i.e. all distributions obtained by permuting the distribution.

We summarize these prior instance-optimality results under permutation neighborhood in Table 1. Apart from the fact that this neighborhood may be too large to be appropriate for some applications (as discussed in Section 1), the last row of Table 1 shows a lower bound for the suboptimality – it shows that no algorithm can do better than an additive $\min\left\{n^{-2/3}, \frac{d}{n}\right\}$ compared to a hypothetical algorithm that already knows the neighborhood. Observe that this additive gap is a lot larger than the per-instance lower bounds for certain distributions, e.g., highly concentrated distributions. Thus, multiplicative instance optimality is not achievable with permutation neighborhood. By contrast, tight instance-optimality (up to constant) is achievable under our additive neighborhood, as proved in Section 2.1. We will also discuss in Section 3.1 that our additive neighborhoods are the smallest (up to constants) that can still allow for the possibility of any DP instance-optimal algorithms.

Additionally, the existing analysis for Good-Turing would result in a $\ln n/n$ additive gap compared to the lower bound we obtain under additive neighborhood. The main gap is that the Good-Turing's error for combined probability estimate is $\ln n/n$ (as in Lemma 19 of [49]) due to delicate correlations between partitioned buckets (of small symbols) and their combined mass estimates, rather than $1/n$ in our sampling-twice estimator (as in (191) that applies Lemma B.10) facilitated by using fresh counts for small symbols to estimate combined mass. We will add this discussion in the revised version to clarify the comparison.

3 DP Instance-Optimality

In this section, we present the first results for DP instance-optimality of KL distribution estimation.

Neighborhood Choices To define appropriate DP instance-optimality objective, we modify the N_+ neighborhood (4) with perturbations calibrated to the privacy parameters. For a small $t \geq 1$, let

$$N_{\leq \frac{t}{n\varepsilon}}(p, n) = \left\{ q : |q_i - p_i| \leq \frac{t}{n\varepsilon} \text{ for any } i \in [d] \text{ and } \sum_{i:p_i \leq \frac{t}{n\varepsilon}} q_i \leq \max\left\{ \frac{t}{n\varepsilon}, \sum_{i:p_i \leq \frac{t}{n\varepsilon}} p_i \right\} \right\} \quad (6)$$

Compared to (4), the main change is that we allow the perturbation scale for individual symbols to be proportional to $1/\varepsilon$. This corresponds to the obliviousness of (ε, δ)-DP algorithm to dataset changes up to $O\left(\frac{1}{\varepsilon}\right)$ hamming distance (which allows an additional $\frac{1}{n\varepsilon}$ change in probability).

DP Per-Instance Lower Bound Under the above neighborhood $N_{\leq \frac{t}{n\varepsilon}}(p, n)$ that is calibrated to (ε, δ)-DP guarantee, we prove the below per-instance lower bound.

Theorem 3.1. *Let $p \in \Delta(d)$ be an arbitrary distribution instance, and let $N_{\leq \frac{t}{n\varepsilon}}(p)$ be the additive neighborhood defined in* (6)*. If $\delta \leq \varepsilon$, $d \geq 2$, and $n\varepsilon \geq 1$, then we have*

$$\begin{aligned} \textit{lower}_{\varepsilon,\delta}\left(p, n, N_{\leq \frac{t}{n\varepsilon}}\right) \geq \Omega\Bigg(& \sum_i \min\left\{ p_i, \frac{1}{p_i} \cdot \frac{1}{n^2\varepsilon^2} \right\} + \frac{\ln\left(1 + d_{small}(L')\right)}{n\varepsilon} \\ & + p_{small}(L') \cdot \ln\left(1 + \frac{d_{small}(L')}{n\varepsilon \cdot p_{small}(L')}\right) \Bigg) \end{aligned} \quad (7)$$

for any $t \geq 1$ and any set $L' \subseteq [d]$, where $p_{small}(L') = \sum_{i \in L': p_i \leq \frac{t}{n\varepsilon}} p_i$ and $d_{small}(L') = \sum_{i \in L': p_i \leq \frac{t}{n\varepsilon}} 1$.

Algorithm 2 ε-DP "Sampling Twice" (Instance-Optimal)

Inputs: Data partition ratio $\alpha = 0.5$. Independently sampled datasets $x \sim \text{Poi}(\alpha \cdot np)$ and $x' \sim \text{Poi}((1-\alpha) \cdot np)$ s.t. $x + x' \sim \text{Poi}(np)$. Threshold $\tau = 4 \ln d$.
$L = \emptyset$
for symbol $i = 1, \cdots, d$ **do**
  Private Thresholding: If $\tilde{x}_i := x_i + z_i \leq \frac{\tau}{\min\{\varepsilon, 1\}}$ for $z_i \sim \text{Lap}\left(0, \frac{1}{\varepsilon}\right)$, add i to L
end for
Estimate small symbols' combined mass: $\tilde{c} = \max\left\{\sum_{i \in L} x'_i + Lap\left(0, \frac{1}{\varepsilon}\right), \frac{1}{\min\{\varepsilon, 1\}}\right\}$
Estimate individual large symbols: for $i \in [d] \setminus L$, $\tilde{x}'_i = x'_i + z'_i$ for $z'_i \sim \text{Lap}\left(0, \frac{1}{\varepsilon}\right)$
Truncation: $\bar{x}_i = \max\left\{\tilde{x}_i, \frac{1}{\min\{\varepsilon, 1\}}\right\}$ for $i \in L$; $\bar{x}_i = (1-\alpha)\left(\max\left\{\tilde{x}_i, \frac{1}{\min\{\varepsilon, 1\}}\right\} + \max\left\{\tilde{x}'_i, \frac{1}{\min\{\varepsilon, 1\}}\right\}\right)$ for $i \in [d] \setminus L$
Return $\mathcal{A}(x)$ with $\mathcal{A}(x)_i = \begin{cases} \frac{1}{N} \cdot \tilde{c} \cdot \frac{\bar{x}'_i}{\sum_{i \in L} \bar{x}'_i} & i \in L \\ \frac{1}{N} \cdot \bar{x}'_i & i \notin L \end{cases}$ where $N = \tilde{c} + \sum_{i \notin L} \bar{x}_i$

The proofs are deferred to Appendix F. The lower bound (7) consists of three terms:

1. Cost of privacy for **large symbols** $p_i \geq \frac{1}{n\varepsilon}$ is $\frac{1}{n^2\varepsilon^2 p_i}$, which is significantly smaller than the corresponding non-DP lower bound $\Omega(\frac{1}{n})$ in (5) when p_i is large enough, i.e., privacy is free for sufficiently large p_i.
2. Cost of privacy for **combined mass estimate of small symbols** $p_i \leq \frac{t}{n\varepsilon}$ is $\frac{\ln\left(1+d_{small}(L')\right)}{n\varepsilon}$, which is non-zero whenever there exist small symbols in L'.
3. Cost of privacy for **small symbols** $p_i < \frac{t}{n\varepsilon}$ is $p_{small}(L') \cdot \ln\left(1 + \frac{d_{small}(L')}{n\varepsilon \cdot p_{small}(L')}\right) + \sum_{i < \frac{1}{n\varepsilon}} p_i$, which is larger than the corresponding non-DP lower bound in (5) for small symbols whenever $\varepsilon \ll 1$, i.e., in the high privacy regime.

3.1 Privatized Instance-Optimal "Sampling Twice" Estimator

We then propose a privatized variant of "sampling-twice" Algorithm 2 by applying the Laplace Mechanism [26]. The ε-DP guarantee follows from standard results in DP by observing that the ℓ_1-sensitivity of $(x_1, \cdots, x_d, x'_1, \cdots, x'_d)$ is one. The main novel ingredient (as discussed in Section 1.1) is to choose a calibrated threshold for privately selecting a set of small symbols. This selected set is then used for estimating the combined mass of small symbols on fresh samples. An accurate combined mass estimate is the key to reducing KL error.

Instance-optimality Guarantee In Theorem G.6, we prove a per-instance upper bound for Algorithm 2. In Corollary G.7, we further prove that this upper bound can be controlled by the per-instance lower bound under the combination of our Non-DP addditive neighborhood $N_+(p)$ in (4) and our (ε, δ)-DP calibrated neighborhood in (6). Specifically, under neighborhood size $t = 24 \ln d$, we prove that $\mathop{\mathbb{E}}_{x \sim \text{Poi}(np)} [\text{KL}(p, \mathcal{A}(x))] \leq O\left(\text{lower}\left(p, n, N_+ \cup N_{\leq \frac{t}{n\varepsilon}}\right)\right)$.

Discussions on the Neighborhood Size We have established instance-optimality under local neighborhood $N_{\frac{24 \ln d}{n\varepsilon}}$. It is a natural question to ask whether the size of such neighborhoods could be further reduced to enable stronger instance-optimality notions. In Theorem G.9, we prove that this neighborhood size is necessary up to constants—no DP estimator can be instance-optimal for a neighborhood $N_{\leq \frac{\gamma \cdot \ln d}{n\varepsilon}}$ with $\gamma \leq o(1)$.

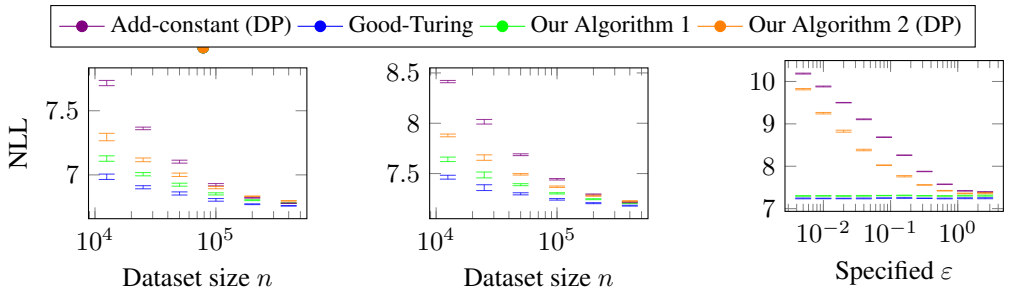
Discussions on Approximate DP variants Our per-instance lower bounds hold for (ε, δ)-approx-DP with $\delta < \varepsilon$, and match the upper bound achieved by pure-DP algorithms. This indicates that our pure-DP algorithm, while enjoying better privacy guarantees, is also instance-optimal under approximate DP. Similar phenomena have previously been observed for the histogram estimation

problem under approx-DP, where in the asymptotic sense, approx DP (e.g., Gaussian mechanism) offers little advantage over pure DP (e.g., Laplace).

In the practical sense, however, the Gaussian mechanism may be practically better in some regime of parameters (due to the lighter distribution tails compared to Laplace mechanism). And our algorithm (after substituting Laplace mechanism with Gaussian mechanism and changing threshold to calibrate to delta) achieves instance optimality up to a $\log(1/\delta)$ factor (see the Gaussian variants of our algorithm and its instance-optimality proof in Appendix G.5).

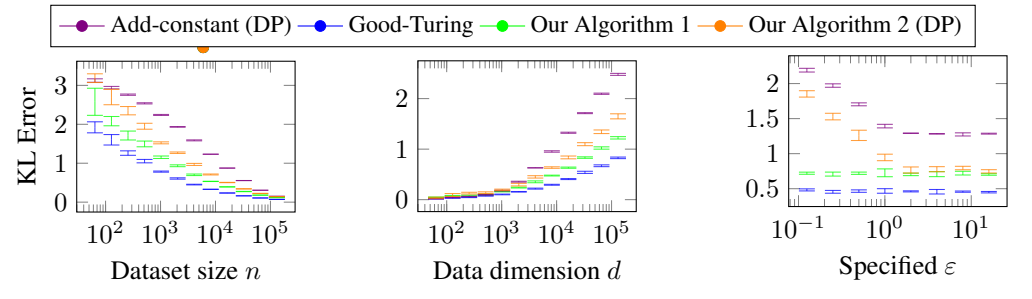
4 Experiments

We now evaluate the performance of our algorithms and compare them with baselines. All experiments are performed on a MacOS intergrated CPU ($\leq$ 30 minutes) with 18GB RAM. All reported performance numbers are averaged across five random trials of data sampling and estimators.



(a) $d = 30522$, $\varepsilon = 1$ (Bert) (b) $d = 50257$, $\varepsilon = 1$ (GPT2) (c) $n = 10^5$, $d = 50257$ (GPT2)

Figure 1: (Reddit Token Distribution Estimation) KL error versus dataset size n, distribution dimension d, and DP guarantee ε for our methods compared with the simple minimax optimal Add-constant (DP) baseline, and the strongest non-DP baseline of prior (near) instance-optimal Good-Turing estimator.



(a) $d = 50000$, $\varepsilon = 1$ (b) $n = 2000$, $\varepsilon = 1$ (c) $n = 1000$, $d = 10000$

Figure 2: (Power law distribution $p_i \propto \frac{1}{i}$) KL error versus dataset size n, distribution dimension d, and DP guarantee ε for our methods compared with the simple minimax optimal Add-constant (DP) baseline, and the strongest non-DP baseline of prior (near) instance-optimal Good-Turing estimator.

Datasets We evaluate on both synthetic and real-world data distributions. For synthetic data, we evaluate power law distributions $p_i \propto \frac{1}{i^\beta}$ over a range of parameters $\beta > 0$. We choose power law because of its practical relevance: word frequencies from a text corpus have long been observed to roughly follow power law distributions [54, 53, 48, 17]. For real-word data, we experiment on randomly drawn tokens from Reddit [60], Enron-email [41] and MMLU [36, 35], where each token is one (sensitive) record. We chose Reddit and Enron-email datasets because they are user-specific and thus bear a natural notion of privacy risk (compared to e.g., wikipedia), and are standard and widely used text datasets in the private learning literature [58, 43, 45]. We additionally evaluate on MMLU to simulate diverse text domains. To vary the distribution dimension (specified by the number of all possible tokens), we use two types of tokenizers (GPT2 and Bert).

One challenge for evaluating real-word dataset is the unknown ground-truth distribution p (as only empirical samples are given). To address this, we independently sample two equal-size datasets x

and x', thus ensuring $\mathbb{E}\left[x'_i/\|x'\|_1\right] = p_i$ for any $i \in [d]$. We then compute that

$$\mathbb{E}\left[KL\left(p, \mathcal{A}(x)\right)\right] = \mathbb{E}\left[\sum_{i=1}^{d} p_i \ln\left(\frac{p_i}{\mathcal{A}(x)_i}\right)\right] = \underbrace{\sum_{i=1}^{d} p_i \ln\left(p_i\right)}_{\text{Negative Entropy of } p} \underbrace{-\mathbb{E}\left[\sum_{i=1}^{d} \frac{x'_i}{\|x'\|_1} \cdot \ln\left(\mathcal{A}(x)_i\right)\right]}_{\text{Negative Log Likelihood (NLL)}}$$

Observe that the entropy term is independent of the estimator. Thus in experiments we only report the negative log likelihood ratio term to compare different estimators $\mathcal{A}$.

Hyperparameters Although our proposed Algorithm 1 and 2 are provably instance-optimal, they separately use two disjoint fractions of the dataset. Thus in the extreme scenarios when all symbols are above or below threshold (e.g., when n, d are exceedingly large or small), Algorithm 1 and 2 incur twice as much noise compared to the add-one estimator, incurring a suboptimal multiplicative constant of two. To address this limitation, we perform grid search for the optimal hyperparameters over $\alpha \in \{0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.99\}$ and $\tau \in \{0, 0.0625, 0.125, 0.25, 0.5, 1, 2, 4\} \times \ln d$, and use the tuned hyperparameters $\tau = 0, \alpha = 0.5$ for Algorithm 1, and $\tau = \min\left\{\frac{1}{\varepsilon}, 1.0\right\} \times \ln d, \alpha = 0.9$ for Algorithm 2 in all experiments.

Observations We show results on power law distributions $p_i \propto \frac{1}{i^\beta}$ for $\beta = 1, 1.5, 2$ (Figure 2, 3, and 4), Reddit corpus (Figure 1), Enron-email corpus(Figure 5); and MMLU corpus (Figure 6). In all experiments, our DP instance-optimal Algorithm 2 consistently outperforms the simple minimax optimal Add-constant (DP) baseline, and our non-DP instance-optimal Algorithm 1 is consistently competitive (within constants) to the strongest prior non-DP baseline of (near) instance-optimal Good-Turing [50]. The gain of our algorithms are especially significant for real-world datasets, validating the effectiveness of our algorithms. We also remark that instance optimality means that our algorithm will provably adapt to any input distribution, and no other algorithm can be significantly better (across a neighborhood).

5 Conclusion

We provide tight instance-optimality analysis for private KL distribution estimation, in terms of achieving provably competitive error to the best possible estimator in a small additive local neighborhood of each instance. Furthermore, our constructed neighborhood's size is necessary up to constants for instance-optimality on the *worst-case* instances. Additionally, we proved instance-optimality up to constants, and leave open the question of whether exact instance-optimality is achievable. Such results, if possible, would require improving the constants in per-instance privacy lower bounds and/or designing better estimators.

Acknowledgements

The authors thank Audra McMillan and Satchit Sivakumar for insightful discussions on the literature, and Daogao Liu for valuable feedback on earlier drafts. Jiayuan Ye is supported by the Apple Scholars in AI/ML PhD fellowship.

References

[1] Jayadev Acharya and Ziteng Sun. Communication complexity in locally private distribution estimation and heavy hitters. In *International Conference on Machine Learning*, pages 51–60. PMLR, 2019.

[2] Jayadev Acharya, Ashkan Jafarpour, Alon Orlitsky, and Ananda Theertha Suresh. Optimal probability estimation with applications to prediction and classification. In *Conference on Learning Theory*, pages 764–796. PMLR, 2013.

[3] Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *Algorithmic Learning Theory*, pages 48–78. PMLR, 2021.

[4] Alan Agresti and Brent A Coull. Approximate is better than "exact" for interval estimation of binomial proportions. *The American Statistician*, 52(2):119–126, 1998.

[5] David Aldous. Random walks on finite groups and rapidly mixing markov chains. In *Séminaire de Probabilités XVII 1981/82: Proceedings*, pages 243–297. Springer, 1983.

[6] Hilal Asi and John C Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. *Advances in neural information processing systems*, 33:14106–14117, 2020.

[7] Hilal Asi and John C Duchi. Near instance-optimality in differential privacy. *arXiv preprint arXiv:2005.10630*, 2020.

[8] Hilal Asi, John C Duchi, Saminul Haque, Zewei Li, and Feng Ruan. Universally instance-optimal mechanisms for private statistical estimation. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 221–259. PMLR, 2024.

[9] Raef Bassily and Adam Smith. Local, private, efficient protocols for succinct histograms. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pages 127–135, 2015.

[10] Dietrich Braess and Thomas Sauer. Bernstein polynomials and learning theory. *Journal of Approximation Theory*, 128(2):187–206, 2004.

[11] Dietrich Braess, Jürgen Forster, Tomas Sauer, and Hans U Simon. How to achieve minimax expected kullback-leibler distance from an unknown finite distribution. In *International Conference on Algorithmic Learning Theory*, pages 380–394. Springer, 2002.

[12] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[13] Stanley F Chen and Joshua Goodman. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language*, 13(4):359–394, 1999.

[14] Stanley F Chen, Douglas Beeferman, and Roni Rosenfeld. Evaluation metrics for language models. 1998.

[15] Albert Cheu, Adam Smith, Jonathan Ullman, David Zeber, and Maxim Zhilyaev. Distributed differential privacy via shuffling. In *Advances in Cryptology–EUROCRYPT 2019: 38th Annual International Conference on the Theory and Applications of Cryptographic Techniques, Darmstadt, Germany, May 19–23, 2019, Proceedings, Part I 38*, pages 375–403. Springer, 2019.

[16] Charles J Clopper and Egon S Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

[17] Brian Conrad and Michael Mitzenmacher. Power laws for monkeys typing randomly: the case of unequal probabilities. *IEEE Transactions on information theory*, 50(7):1403–1414, 2004.

[18] Thomas Cover. Broadcast channels. *IEEE Transactions on Information Theory*, 18(1):2–14, 1972.

[19] Ilias Diakonikolas, Moritz Hardt, and Ludwig Schmidt. Differentially private learning of structured discrete distributions. *Advances in Neural Information Processing Systems*, 28, 2015.

[20] Irit Dinur and Kobbi Nissim. Revealing information while preserving privacy. In *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 202–210, 2003.

[21] David L Donoho and Iain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455, 1994.

[22] Evgeny Drukh and Yishay Mansour. Concentration bounds for unigram language models. *Journal of Machine Learning Research*, 6(8), 2005.

[23] John C Duchi and Feng Ruan. The right complexity measure in locally private estimation: It is not the fisher information. *The Annals of Statistics*, 52(1):1–51, 2024.

[24] John C Duchi, Michael I Jordan, and Martin J Wainwright. Local privacy, data processing inequalities, and minimax rates. *arXiv preprint arXiv:1302.3203*, 2013.

[25] John C Duchi, Michael I Jordan, and Martin J Wainwright. Minimax optimal procedures for locally private estimation. *Journal of the American Statistical Association*, 113(521):182–201, 2018.

[26] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006.

[27] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

[28] Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *2015 IEEE 56th Annual Symposium on Foundations of Computer Science*, pages 650–669. IEEE, 2015.

[29] Úlfar Erlingsson, Vitaly Feldman, Ilya Mironov, Ananth Raghunathan, Kunal Talwar, and Abhradeep Thakurta. Amplification by shuffling: From local to central differential privacy via anonymity. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2468–2479. SIAM, 2019.

[30] Vitaly Feldman, Audra McMillan, Satchit Sivakumar, and Kunal Talwar. Instance-optimal private density estimation in the wasserstein distance. *arXiv preprint arXiv:2406.19566*, 2024.

[31] William A Gale and Geoffrey Sampson. Good-turing frequency estimation without tears. *Journal of quantitative linguistics*, 2(3):217–237, 1995.

[32] Badih Ghazi, Noah Golowich, Ravi Kumar, Rasmus Pagh, and Ameya Velingker. On the power of multiple anonymous messages: Frequency estimation and selection in the shuffle model of differential privacy. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pages 463–488. Springer, 2021.

[33] Irving J Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(3-4):237–264, 1953.

[34] Michael Hay, Vibhor Rastogi, Gerome Miklau, and Dan Suciu. Boosting the accuracy of differentially-private histograms through consistency. *arXiv preprint arXiv:0904.0942*, 2009.

[35] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[36] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[37] Nils Homer, Szabolcs Szelinger, Margot Redman, David Duggan, Waibhav Tembe, Jill Muehling, John V Pearson, Dietrich A Stephan, Stanley F Nelson, and David W Craig. Resolving individuals contributing trace amounts of dna to highly complex mixtures using high-density snp genotyping microarrays. *PLoS genetics*, 4(8):e1000167, 2008.

[38] Peter Kairouz, Keith Bonawitz, and Daniel Ramage. Discrete distribution estimation under local privacy. In *International Conference on Machine Learning*, pages 2436–2444. PMLR, 2016.

[39] Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In *Conference on Learning Theory*, pages 1066–1100. PMLR, 2015.

[40] Slava Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE transactions on acoustics, speech, and signal processing*, 35(3): 400–401, 1987.

[41] Bryan Klimt and Yiming Yang. The enron corpus: A new dataset for email classification research. In Jean-François Boulicaut, Floriana Esposito, Fosca Giannotti, and Dino Pedreschi, editors, *Machine Learning: ECML 2004*, pages 217–226, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-30115-8.

[42] Raphail Krichevsky and Victor Trofimov. The performance of universal encoding. *IEEE Transactions on Information Theory*, 27(2):199–207, 1981.

[43] Xuechen Li, Florian Tramer, Percy Liang, and Tatsunori Hashimoto. Large language models can be strong differentially private learners. *arXiv preprint arXiv:2110.05679*, 2021.

[44] Yuhan Liu, Ananda Theertha Suresh, Felix Xinnan X Yu, Sanjiv Kumar, and Michael Riley. Learning discrete distributions: user vs item-level privacy. *Advances in Neural Information Processing Systems*, 33:20965–20976, 2020.

[45] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE, 2023.

[46] David A McAllester and Robert E Schapire. On the convergence rate of good-turing estimators. In *COLT*, pages 1–6, 2000.

[47] Audra McMillan, Adam Smith, and Jon Ullman. Instance-optimal differentially private estimation. *arXiv preprint arXiv:2210.15819*, 2022.

[48] Michael Mitzenmacher. A brief history of generative models for power law and lognormal distributions. *Internet mathematics*, 1(2):226–251, 2004.

[49] Alon Orlitsky and Ananda Theertha Suresh. Competitive distribution estimation: Why is good-turing good. *Advances in Neural Information Processing Systems*, 28, 2015.

[50] Alon Orlitsky, Narayana P Santhanam, and Junan Zhang. Always good turing: Asymptotically optimal probability estimation. *Science*, 302(5644):427–431, 2003.

[51] Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6): 1191–1253, 2003.

[52] Liam Paninski. Variational minimax estimation of discrete distributions under kl loss. *Advances in Neural Information Processing Systems*, 17, 2004.

[53] Steven T Piantadosi. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21:1112–1130, 2014.

[54] David MW Powers. Applications and explanations of zipf's law. In *New methods in language processing and computational natural language learning*, 1998.

[55] Sriram Sankararaman, Guillaume Obozinski, Michael I Jordan, and Eran Halperin. Genomic privacy and limits of individual detection in a pool. *Nature genetics*, 41(9):965–967, 2009.

[56] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[57] Michael Short. Improved inequalities for the poisson and binomial distribution and upper tail quantile functions. *International Scholarly Research Notices*, 2013, 2013.

[58] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, pages 587–601, 2017.

[59] Gregory Valiant and Paul Valiant. An automatic inequality prover and instance optimal identity testing. *SIAM Journal on Computing*, 46(1):429–455, 2017.

[60] Michael V"olske, Martin Potthast, Shahbaz Syed, and Benno Stein. TL;DR: Mining Reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pages 59–63, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4508. URL `https://www.aclweb.org/anthology/W17-4508`.

[61] Tianhao Wang, Jeremiah Blocki, Ninghui Li, and Somesh Jha. Locally differentially private protocols for frequency estimation. In *26th USENIX Security Symposium (USENIX Security 17)*, pages 729–745, 2017.

[62] Qun Xie and Andrew R Barron. Minimax redundancy for the class of memoryless sources. *IEEE Transactions on Information Theory*, 43(2):646–657, 1997.

[63] Jia Xu, Zhenjie Zhang, Xiaokui Xiao, Yin Yang, Ge Yu, and Marianne Winslett. Differentially private histogram publication. *The VLDB journal*, 22:797–822, 2013.

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

[65] Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pages 423–435. Springer, 1997.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: All claims in the abstract are elaborated in the introduction, and all claims in the introduction are followed by references to their discussed sections.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the conclusion Section 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.

- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: See the statements in Section 2 and 3.1 and proofs in the appendices.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Algorithm pseudocodes are given in Algorithm 1 and 2, and hyperparameters are given in Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often

one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not provide access to the code, but only uses opensource datasets in Section 4, and provides algorithm pseudocode in Algorithm 1 and 2.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Hyperparameters are given in Section 4

   Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: All our experiments (Figure 2, 1 and Appendix G.4) report 1-sigma error bars across five runs.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

Justification: We reviewed the NeurIPS Code of Ethics and ensured it is respected.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: We do not see potential such risks of this paper, as we only used synthetic and open-source datasets for validation.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: We do not foresee any no potential ethical harms, as the paper only uses synthetic and open-source datasets for validation.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All used datasets are referenced in Section 4, and have licenses that allow research use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: LLMs are only used for checking grammar and formatting.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Contents

A Tools for Lower Bounds

Below we first define decomposable statistical distance.

Definition A.1 (Decomposable Statistical Distance)**.** Let $d \in \mathbb{N}$ and let $dist : \Delta(d) \times \Delta(d) \to \mathbb{R}$ be a non-negative function such that $dist(p, q) = 0$ if and only if $p = q$. We say $dist$ is decomposable, if for any disjoint sets of symbols $\mathcal{B}_1, \cdots, \mathcal{B}_k \subseteq [d]$, and for any distributions $p, q \in \Delta(d)$, it is the case that

$$dist(p, q) \geq \sum_{i=1}^{k} \left(\sum_{j \in \mathcal{B}_i} p_j \right) \cdot dist(p|_i, q|_i) \tag{8}$$

where for any distribution $p \in \Delta(d)$, and any $i = 1, \cdots, k$, we have denoted

$$p|_i = \begin{cases} \frac{p_j}{\sum_{j \in \mathcal{B}_i} p_j} & \text{if } j \in \mathcal{B}_i \\ 0 & \text{otherwise} \end{cases} \quad \text{for } i = 1, \cdots, k$$

Lemma A.2 (KL divergence is decomposable)**.** *KL divergence is decomposable, that is, for any disjoint sets of symbols $\mathcal{B}_1, \cdots, \mathcal{B}_k$ and any $p, q \in \Delta(d)$, it is the case that*

$$KL(p, q) \geq \sum_{i=1}^{k} \left(\sum_{j \in \mathcal{B}_i} p_j \right) \cdot KL\left(p|_i, q|_i\right) \tag{9}$$

Proof. Let $p, q \in \Delta(d)$. Denote $\mathcal{B}^c = [d] \setminus (\cup_{i=1}^{k} \mathcal{B}_i)$, then by definition we compute

$$\begin{aligned} KL(p, q) =& \sum_{i=1}^{k} \sum_{j \in \mathcal{B}_i} p_j \ln\left(\frac{p_j}{q_j}\right) + \sum_{j \in \mathcal{B}^c} p_j \ln\left(\frac{p_j}{q_j}\right) && (10) \\ =& \sum_{i=1}^{k} \left(\sum_{j \in \mathcal{B}_i} p_j \right) \cdot KL\left(p|_i, q|_i\right) + \sum_{i=1}^{k} \left(\sum_{j \in \mathcal{B}_i} p_j \right) \cdot \ln\left(\frac{\sum_{j \in \mathcal{B}_i} p_j}{\sum_{j \in \mathcal{B}_i} q_j}\right) && (11) \\ &+ \left(\sum_{j \in \mathcal{B}^c} p_j \right) \cdot KL\left(p|_{\mathcal{B}^c}, q|_{\mathcal{B}^c}\right) + \left(\sum_{j \in \mathcal{B}^c} p_j \right) \cdot \ln\left(\frac{\sum_{j \in \mathcal{B}^c} p_j}{\sum_{j \in \mathcal{B}^c} q_j}\right) && (12) \\ \geq& \sum_{i=1}^{k} \left(\sum_{j \in \mathcal{B}_i} p_j \right) \cdot KL\left(p|_i, q|_i\right) && (13) \end{aligned}$$

where the last inequality is by non-negativity of KL divergence $KL\left(p|_{\mathcal{B}^c}, q|_{\mathcal{B}^c}\right)$ and $\sum_{i=1}^{k} \left(\sum_{j \in \mathcal{B}_i} p_j \right) \cdot \ln\left(\frac{\sum_{j \in \mathcal{B}_i} p_j}{\sum_{j \in \mathcal{B}_i} q_j}\right) + \left(\sum_{j \in \mathcal{B}^c} p_j \right) \cdot \ln\left(\frac{\sum_{j \in \mathcal{B}^c} p_j}{\sum_{j \in \mathcal{B}^c} q_j}\right)$. □

Theorem A.3 (Generalized Assouad's method for decomposable statistical distance)**.** *Let $dist$ be a decomposable statistic distance as per Definition A.1. Let $\mathcal{B}_1, \cdots, \mathcal{B}_k \subseteq [d]$ be k disjoint sets of symbols. For each $i = 1, \cdots, k$, let $\mathcal{P}_i$ be a set of distributions supported on $\mathcal{B}_i$. For fixed $w_1, \cdots, w_k \geq 0$ such that $\sum_{i=1}^{k} w_i = 1$, let $\mathcal{P}$ be the following composed packing set:*

$$\mathcal{P} = \left\{ q \coloneqq \sum_{i=1}^{k} w_i p^i \middle| p^i \in \mathcal{P}_i \right\}. \tag{14}$$

If the following two conditions hold:

1. *There exists a non-negative function f such that*

$$\text{For any } q \in \Delta(d), \quad \frac{1}{|\mathcal{P}_i|}\sum_{p\in\mathcal{P}_i} \mathbf{1}_{dist(p,q)\geq f(\mathcal{P}_i)} \geq \frac{1}{2}, \tag{15}$$

2. *For each $i \in [k]$, there exists $\tau_i \geq 0$ and $\bar{p}^i \in \mathcal{P}_i$, such that for any fixed $p^j \in \mathcal{P}_j, j \neq i$, it holds that*

$$\int \left(\min_{p^i\in\mathcal{P}_i} \frac{d\mathcal{S}(n, \sum_{j=1}^k w_j p^j)}{d\mathcal{S}(n, w_i\bar{p}^i + \sum_{j\neq i} w_j p^j)} \right) d\mathcal{S}(n, w_i\bar{p}^i + \sum_{j\neq i} w_j p^j) \geq \tau_i \tag{16}$$

where we have denoted $\mathcal{S}(n, p)$ as the distribution of histogram representation of dataset sampled from distribution $p \in \Delta(d)$ with target sample size n.

Then for any fixed $\varepsilon > 0$, $\delta \leq \varepsilon$,

$$\min_{\mathcal{A} \text{ is } (\varepsilon,\delta)\text{-DP}} \max_{p\in\mathcal{P}} \mathop{\mathbb{E}}_{x\sim\mathcal{S}(n,p)} [dist(p, \mathcal{A}(x))] \geq \frac{1}{2}\sum_{i=1}^k w_i \cdot \tau_i \cdot f(\mathcal{P}_i) \tag{17}$$

Proof. We will reduce the estimation problem over all d symbols to the estimation problem over each bucket $\mathcal{B}_i$. For any distribution $p \in \Delta(d)$, denote its conditional distribution on $\mathcal{B}_i$ as

$$p|_i = \begin{cases} \frac{p_j}{\sum_{j\in\mathcal{B}_i} p_j} & \text{if } j \in \mathcal{B}_i \\ 0 & \text{otherwise} \end{cases} \text{ for } i = 1, \cdots, k. \tag{18}$$

And for any (ε, δ)-DP estimator $\mathcal{A}$ given dataset x supported on $\mathcal{B}$, similarly denote

$$\mathcal{A}(x)|_i = \begin{cases} \frac{\mathcal{A}(x)_j}{\sum_{j\in\mathcal{B}_i} \mathcal{A}(x)_j} & \text{if } j \in \mathcal{B}_i \\ 0 & \text{otherwise} \end{cases} \text{ for } i = 1, \cdots, k. \tag{19}$$

Then by definition, for any fixed $i = 1, \cdots, k$, we have

$$\frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \mathop{\mathbb{E}}_{x\sim\mathcal{S}(n,p),\mathcal{A}} [dist(p|_i, \mathcal{A}(x)|_i)] \geq \frac{f(\mathcal{P}_i)}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \mathop{\mathbb{E}}_{x\sim\mathcal{S}(n,p),\mathcal{A}} \left[\mathbf{1}_{dist(p|_i,\mathcal{A}(x)|_i)\geq f(\mathcal{P}_i)}\right] \tag{20}$$

$$= \frac{f(\mathcal{P}_i)}{\prod_{j=1}^k |\mathcal{P}_j|} \sum_{p^j\in\mathcal{P}_j, j\neq i} \sum_{p^i\in\mathcal{P}_i} \mathop{\mathbb{E}}_{x\sim\mathcal{S}\left(n, \sum_{l=1}^d w_l p^l\right)} \left[\mathbb{E}_{\mathcal{A}}\left[\mathbf{1}_{dist(p^i,\mathcal{A}(x)|_i)\geq f(\mathcal{P}_i)}\right]\right] \tag{21}$$

$$\geq \frac{f(\mathcal{P}_i)}{\prod_{j=1}^k |\mathcal{P}_j|} \sum_{p^j\in\mathcal{P}_j, j\neq i} \mathop{\mathbb{E}}_{\bar{x}\sim\mathcal{S}\left(n, w_i\bar{p}^i + \sum_{l\neq i} w_l p^l\right)} \left[\min_{p^i\in\mathcal{P}_i} \frac{d\mathcal{S}(n, \sum_{j=1}^k w_j p^j)}{d\mathcal{S}(n, w_i\bar{p}^i + \sum_{j\neq i} w_j p^j)}(\bar{x}) \cdot \sum_{p^i\in\mathcal{P}_i} \mathbb{E}_{\mathcal{A}}\left[\mathbf{1}_{dist(p^i,\mathcal{A}(x)|_i)\geq f(\mathcal{P}_i)}\right] \right] \tag{22}$$

$$\geq \frac{f(\mathcal{P}_i)}{2} \cdot \mathop{\mathbb{E}}_{\bar{x}\sim\mathcal{S}\left(n, w_i\bar{p}^i + \sum_{l\neq i} w_l p^l\right)} \left[\min_{p^i\in\mathcal{P}_i} \frac{d\mathcal{S}(n, \sum_{j=1}^k w_j p^j)}{d\mathcal{S}(n, w_i\bar{p}^i + \sum_{j\neq i} w_j p^j)}(\bar{x}) \right] \tag{23}$$

$$\geq \frac{\tau_i}{2} \cdot f(\mathcal{P}_i) \tag{24}$$

where (20) is by Markov inequality, (22) is by definition and by moving the sum of $p^i \in \mathcal{P}_i$ inside expectation (as $\bar{x}$ and $\min_{p^i\in\mathcal{P}_i} \frac{d\mathcal{S}(n, \sum_{j=1}^k w_j p^j)}{d\mathcal{S}(n, w_i\bar{p}^i + \sum_{j\neq i} w_j p^j)}(\bar{x})$ are independent of the choice of p^i), (23) is by the packing assumption (23), and (24) is by the dataset distance assumption (16).

We now use (23) to prove lower bound on the expected KL error for estimating the whole distribution over all possible distributions in the packing set $\mathcal{P}$. By Lemma A.2, it follows that

$$\frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}} \mathop{\mathbb{E}}_{x\sim\mathcal{S}(n,p),\mathcal{A}} [dist(p, \mathcal{A}(x))] \geq \frac{1}{|\mathcal{P}|}\sum_{p\in\mathcal{P}}\sum_{i=1}^k w_i \cdot \mathop{\mathbb{E}}_{x\sim\mathcal{S}(n,p),\mathcal{A}} [dist(p|_i, \mathcal{A}(x)|_i)] \tag{25}$$

$$\geq \frac{1}{2}\sum_{i=1}^k w_i \cdot \tau_i \cdot f(\mathcal{P}_i) \tag{26}$$

where the last inequality (26) is by (23). This suffice to prove (17) by observing that average is smaller than maximum. □

Theorem A.4 (Generalized DP Assouad's method for decomposable statistical distance)**.** *Let $dist$ be a decomposable statistic distance as per Definition A.1. Let $\mathcal{B}_1, \cdots, \mathcal{B}_k \subseteq [d]$ be k disjoint sets of symbols. For each $i = 1, \cdots, k$, let $\mathcal{P}_i$ be a set of distributions supported on $\mathcal{B}_i$. For fixed $w_1, \cdots, w_k \geq 0$ such that $\sum_{i=1}^k w_i = 1$, let $\mathcal{P}$ be the following composed packing set:*

$$\mathcal{P} = \left\{ q \coloneqq \sum_{i=1}^{k} w_i p^i \middle| p^i \in \mathcal{P}_i \right\}. \tag{27}$$

For any $p \in \Delta(d)$, denote $\mathcal{S}(n, p)$ as the distribution of histogram representation of dataset sampled from p with target sample size n. If the following two conditions hold:

1. *There exists a non-negative function f such that*

$$\text{For any } q \in \Delta(d), \quad \frac{1}{|\mathcal{P}_i|} \sum_{p \in \mathcal{P}_i} \mathbf{1}_{dist(p,q) \geq f(\mathcal{P}_i)} \geq \frac{1}{2}, \tag{28}$$

2. *For each $i \in [k]$, there exists $\tau_i \geq 0$ and $\bar{p}^i \in \mathcal{P}_i$, such that for any fixed $p^j \in \mathcal{P}_j, j = 1, \cdots, k$, it holds that*

$$\mathbb{E}_{(x,\bar{x})} \left[\|x - \bar{x}\|_1 \right] \leq \tau_i \tag{29}$$

for a coupling $(x, \bar{x})$ between distributions $\mathcal{S}(n, \sum_{j=1}^k w_j p^j)$ and $\mathcal{S}(n, w_i \bar{p}^i + \sum_{j \neq i} w_j p^j)$.

Then for any fixed $\varepsilon > 0$, $\delta \leq \varepsilon$,

$$\min_{\mathcal{A} \text{ is } (\varepsilon,\delta)\text{-DP}} \max_{p \in \mathcal{P}} \mathop{\mathbb{E}}_{x \sim \mathcal{S}(n,p)} \left[dist(p, \mathcal{A}(x)) \right] \geq \sum_{i=1}^{k} w_i \cdot \left(\frac{1}{10} - 4\varepsilon \cdot \tau_i \right) \cdot f(\mathcal{P}_i) \tag{30}$$

Proof. We will reduce the estimation problem over all d symbols to the estimation problem over each bucket $\mathcal{B}_i$. For any distribution $p \in \Delta(d)$, we denote $p|_i$ as its conditional distribution on $\mathcal{B}_i$ as defined in (18). And for any (ε, δ)-DP estimator $\mathcal{A}$ given dataset x supported on $\mathcal{B}$, we denote $\mathcal{A}(x)|_i$ as the conditional estimate on $\mathcal{B}_i$ as defined in (19). Then by definition, for any fixed $i = 1, \cdots, k$, we have

$$\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathop{\mathbb{E}}_{x \sim \mathcal{S}(n,p), \mathcal{A}} \left[dist(p|_i, \mathcal{A}(x)|_i) \right] \geq \frac{f(\mathcal{P}_i)}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathop{\mathbb{E}}_{x \sim \mathcal{S}(n,p), \mathcal{A}} \left[\mathbf{1}_{dist(p|_i, \mathcal{A}(x)|_i) \geq f(\mathcal{P}_i)} \right] \tag{31}$$

$$= \frac{f(\mathcal{P}_i)}{\prod_{j=1}^k |\mathcal{P}_j|} \sum_{p^j \in \mathcal{P}_j, j \neq i} \sum_{p^i \in \mathcal{P}_i} \mathop{\mathbb{E}}_{x \sim \mathcal{S}\left(n, \sum_{l=1}^d w_l p^l\right)} \left[\Pr_{\mathcal{A}} \left[dist(p^i, \mathcal{A}(x)|_i) \geq f(\mathcal{P}_i) \right] \right] \tag{32}$$

$$\geq \frac{f(\mathcal{P}_i)}{\prod_{j=1}^k |\mathcal{P}_j|} \sum_{p^j \in \mathcal{P}_j, j \neq i} \sum_{p^i \in \mathcal{P}_i} \mathop{\mathbb{E}}_{(x,\bar{x})} \left[e^{-\varepsilon \cdot \|x - \bar{x}\|_1} \cdot \Pr_{\mathcal{A}} \left[dist(p^i, \mathcal{A}(\bar{x})|_i) \geq f(\mathcal{P}_i) \right] - \delta \cdot \|x - \bar{x}\|_1 \right] \tag{33}$$

$$\geq \frac{f(\mathcal{P}_i)}{\prod_{j=1}^k |\mathcal{P}_j|} \sum_{p^j \in \mathcal{P}_j, j \neq i} \sum_{p^i \in \mathcal{P}_i} \mathop{\mathbb{E}}_{(x,\bar{x})} \left[e^{-\varepsilon \cdot \tau_i \cdot C} \cdot \Pr_{\mathcal{A}} \left[dist(p^i, \mathcal{A}(\bar{x})|_i) \geq f(\mathcal{P}_i) \right] - \delta \cdot \tau_i \cdot C - \mathbf{1}_{\|x - \bar{x}\|_1 > \tau_i \cdot C} \right] \tag{34}$$

$$= \frac{f(\mathcal{P}_i)}{\prod_{j=1}^k |\mathcal{P}_j|} \sum_{p^j \in \mathcal{P}_j, j \neq i} \left(e^{-\varepsilon \cdot \tau_i \cdot C} \cdot \mathop{\mathbb{E}}_{\bar{x}} \left[\sum_{p^i \in \mathcal{P}_i} \Pr_{\mathcal{A}} \left[dist(p^i, \mathcal{A}(\bar{x})|_i) \geq f(\mathcal{P}_i) \right] \right] - \delta \cdot \tau_i \cdot C \cdot |\mathcal{P}_i| \right.$$
$$\left. - \sum_{p^i} \Pr_{(x,\bar{x})} \left[\|x - \bar{x}\|_1 > \tau_i \cdot C \right] \right) \tag{35}$$

$$\geq \left(\frac{e^{-\varepsilon \cdot \tau_i \cdot C}}{2} - \varepsilon \cdot \tau_i \cdot C - \frac{1}{C} \right) \cdot f(\mathcal{P}_i) \tag{36}$$

$$> \left(\frac{1}{10} - 4\varepsilon \cdot \tau_i \right) \cdot f(\mathcal{P}_i) \tag{37}$$

where (31) is by Markov inequality, (33) is by group privacy and holds for arbitrary constant $C > 0$, (35) is by moving the sum of $p^i \in \mathcal{P}_i$ inside expectation (as $\bar{x}$ is independent of the choice of p^i given coupling $(x, \bar{x})$ between distributions $\mathcal{S}(n, \sum_{j=1}^d w_j p^j)$ and $\mathcal{S}(n, w_i \bar{p}^i + \sum_{j \neq i} w_j p^j)$), (36) is by the packing assumption (28) and $\delta \leq \varepsilon$ and the dataset distance assumption (29), and (37) is by choosing $C = \frac{1}{4\varepsilon \cdot \tau_i}$.

We now use (37) to prove a lower bound on the expected error for estimating the whole distribution over all possible distributions in the packing set $\mathcal{P}$. By Lemma A.2, it follows that

$$\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathop{\mathbb{E}}_{x \sim \mathcal{S}(n,p), \mathcal{A}} [dist(p, \mathcal{A}(x))] \geq \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \sum_{i=1}^{k} w_i \cdot \mathop{\mathbb{E}}_{x \sim \mathcal{S}(n,p), \mathcal{A}} [dist(p|_i, \mathcal{A}(x)|_i)] \tag{38}$$

$$\geq \sum_{i=1}^{k} w_i \cdot \left(\frac{1}{10} - 4\varepsilon \cdot \tau_i \right) \cdot f(\mathcal{P}_i) \tag{39}$$

where the last inequality (39) is by (37). This suffice to prove (30) by observing that average is smaller than maximum. □

Finally, we provide two useful constructions of packing set that satisfy (15) and (28).

Lemma A.5 (Packing over Two Symbols)**.** *Let $\mathcal{B} = \{j_1, j_2\}$ be a set of two symbols. Given $0 \leq \Delta \leq a \leq 1$, let $\mathcal{P} \coloneqq \{p, p^-\}$ be a packing set containing the following two distributions on $\mathcal{B}$.*

$$p(j) = \begin{cases} a & j = j_1 \\ 1 - a & j = j_2 \\ 0 & j \in [d] \setminus \{j_1, j_2\} \end{cases} \quad \textit{and } p^-(j) = \begin{cases} a - \Delta & j = j_1 \\ 1 - a + \Delta & j = j_2 \\ 0 & j \in [d] \setminus \{j_1, j_2\} \end{cases}$$

If $\Delta < \frac{a}{2}$, then

$$\textit{For any } q \in \Delta(d) \quad \frac{1}{2} \left(\mathbf{1}_{KL(p,q) \geq \frac{\Delta^2}{8a}} + \mathbf{1}_{KL(p^-,q) \geq \frac{\Delta^2}{8a}} \right) \geq \frac{1}{2} \tag{40}$$

Proof. We will prove this claim by separating the analysis for different q.

1. If $q_{j_1} \leq a - \frac{\Delta}{2}$: by definition, we compute that

$$KL(p, q) = a \ln \left(\frac{a}{q_{j_1}} \right) + (1 - a) \ln \left(\frac{1 - a}{q_{j_2}} \right)$$

$$\geq a - q_{j_1} + \frac{1}{2} \cdot \frac{(a - q_{j_1})^2}{a} + (1 - a) - q_{j_2} \tag{41}$$

$$\geq \frac{1}{8} \cdot \frac{\Delta^2}{a} \tag{42}$$

   where (41) is by $\ln(1 + x) \leq x - \frac{x^2}{2}$ for $x \leq 0$, and by $\ln(1 + x) \leq x$ for $x \geq 0$, and the last inequality is by $q_{j_1} + q_{j_2} \leq 1$ and by using the condition that $q_{j_1} \leq a - \frac{\Delta}{2}$.

2. If $q_{j_1} > a - \frac{\Delta}{2}$: by definition, we compute that

$$KL(p^-, q) = (a - \Delta) \ln \left(\frac{a - \Delta}{q_{j_1}} \right) + (1 - a + \Delta) \ln \left(\frac{1 - a + \Delta}{q_{j_2}} \right)$$

$$\geq a - \Delta - q_{j_1} + \frac{1}{4} \cdot \frac{(a - \Delta - q_{j_1})^2}{a - \Delta} + (1 - a + \Delta) - q_{j_2} \tag{43}$$

$$\geq \frac{1}{8} \cdot \frac{\Delta^2}{a} \tag{44}$$

   where (43) is by $\ln(1 + x) \leq x - \frac{x^2}{4}$ for $0 \leq x < \frac{1}{2}$, and by $\ln(1 + x) \leq x$ for $x \geq 0$, and the last inequality is by $q_{j_1} + q_{j_2} \leq 1$ and by using the condition that $q_{j_1} > a - \frac{\Delta}{2}$.

□

Lemma A.6 (Dirac Distribution Packing over κ Symbols)**.** *Let $\mathcal{B} = \{j_1, \cdots, j_\kappa\}$ be a set of $\kappa \geq 2$ symbols. Let $\mathcal{P} \coloneqq \{p, p^-\}$ be a packing set containing the following dirac distributions on $\mathcal{B}$.*

$$p^l(j) = \begin{cases} 1 & j = j_l \\ 0 & j \in [d] \setminus \{j_l\} \end{cases} \text{ for } l = 1, \cdots, \kappa$$

Then it holds that

$$\text{For any } q \in \Delta(d) \quad \frac{1}{\kappa} \sum_{l=1}^{\kappa} \mathbf{1}_{KL(p^l, q) \geq \ln\left(1+\frac{\kappa}{4}\right)} \geq \frac{1}{2} \tag{45}$$

Proof. We prove by contradiction, suppose that there exists $q \in \Delta(d)$, such that

$$\sum_{l=1}^{\kappa} \mathbf{1}_{KL(p^l, q) \geq \ln\left(1+\frac{\kappa}{4}\right)} < \frac{\kappa}{2} \tag{46}$$

Then by definition of KL divergence, we have

$$\sum_{l=1}^{\kappa} \mathbf{1}_{q_{j_l} \leq \frac{1}{1+\frac{\kappa}{4}}} < \frac{\kappa}{2} \tag{47}$$

Observe that $\sum_{l=1}^{\kappa} \mathbf{1}_{q_{j_l} \leq \frac{1}{1+\frac{\kappa}{4}}}$ is integer, (47) implies that

$$\sum_{l=1}^{\kappa} \mathbf{1}_{q_{j_l} \leq \frac{1}{1+\frac{\kappa}{4}}} \leq \begin{cases} 0 & \kappa = 2 \\ \frac{2\kappa}{3} - 1 & \kappa \geq 3 \end{cases} \tag{48}$$

Thus

$$\sum_{l=1}^{\kappa} q_{j_l} \geq \frac{1}{1+\frac{\kappa}{4}} \cdot \sum_{l=1}^{\kappa} \mathbf{1}_{q_{j_l} > \frac{1}{1+\frac{\kappa}{4}}} = \frac{1}{1+\frac{\kappa}{4}} \cdot \left(\kappa - \sum_{l=1}^{\kappa} \mathbf{1}_{q_{j_l} \leq \frac{1}{1+\frac{\kappa}{4}}}\right) \tag{49}$$

$$\geq \begin{cases} \frac{4}{3} & \kappa = 2 \\ \frac{1+\frac{\kappa}{3}}{1+\frac{\kappa}{4}} & \kappa \geq 3 \end{cases} > 1 \tag{50}$$

where (50) is by (47). This contradicts $q \in \Delta(d)$. □

B Useful Lemmas for Poisson and Laplace Random Variables

Lemma B.1 (Expectation of inverse Poisson random variable)**.** *Let $p \in [0, 1], m \in \mathbb{N}$, and let $x \sim Poi(mp)$. Then*

$$\mathbb{E}\left[\frac{1}{x+1}\right] \leq \frac{1}{mp} \tag{51}$$

Proof. By definition, we compute that

$$\mathbb{E}\left[\frac{1}{x+1}\right] = \sum_{t=0}^{\infty} \frac{1}{t+1} \cdot e^{-mp} \cdot \frac{(mp)^t}{t!} = \frac{1}{mp} \sum_{t=1}^{\infty} e^{-mp} \cdot \frac{(mp)^t}{t!} = \frac{1}{mp} \cdot \left(1 - e^{-mp}\right) \leq \frac{1}{mp} \tag{52}$$

□

Lemma B.2 (Tail bound for Sum of Poisson and Laplace Random Variables)**.** *For any $a, b, c > 0$. Suppose that $x \sim Poi(a)$, $z \sim Lap(0, b)$, then we have that*

$$\Pr[x + z \leq c] \leq \frac{4}{3} e^{\left(-\frac{a}{3} + \frac{c}{2}\right) \cdot \frac{1}{\max\{b,1\}}} \quad \text{and} \quad \Pr[x + z \geq c] \leq \frac{4}{3} e^{\frac{a-c}{2} \cdot \frac{1}{\max\{b,1\}}}. \tag{53}$$

Proof. We first compute the moment generating function of $x+z$, i.e., the convolution of Poisson random variable and Laplace noise. For any $\theta \in [-\frac{1}{b}, \frac{1}{b}]$, by definition, we have that

$$M_{x+z}(\theta) = \mathbb{E}\left[e^{\theta\cdot(x+z)}\right] = \mathbb{E}\left[e^{\theta\cdot x}\right]\cdot\mathbb{E}\left[e^{\theta\cdot Lap(0,b)}\right] = e^{a(e^{\theta}-1)}\cdot\frac{1}{1-b^2\theta^2} \tag{54}$$

Then, by using Markov inequality, for $\theta \in [0, \frac{1}{b})$, we have

$$\Pr[\tilde{x}\le c] = \Pr\left[e^{-\theta(x+z)} \ge e^{-\theta c}\right] \le \frac{M_{x+z}(-\theta)}{e^{-\theta c}} = \frac{e^{a(e^{-\theta}-1)+\theta c}}{1-b^2\theta^2} \tag{55}$$

By choosing $\theta = \frac{1}{2\max\{b,1\}}$, then by observing $e^{-\theta}-1 \le -\frac{2}{3}\theta$ for $\theta = \frac{1}{2\max\{b,1\}} < \frac{1}{2}$, we prove that

$$\Pr[\tilde{x}\le c] \le \frac{4}{3}\cdot e^{\left(-\frac{a}{3}+\frac{c}{2}\right)\cdot\frac{1}{\max\{b,1\}}} \tag{56}$$

Similarly, we prove a bound for the upper tail by Markov inequality.

$$\Pr[\tilde{x}\ge c] = \Pr\left[e^{\theta(x+z)} \ge e^{\theta c}\right] \le \frac{M_{x+z}(\theta)}{e^{\theta c}} = \frac{e^{a(e^{\theta}-1)-\theta c}}{1-\theta^2 b^2} \le \frac{4}{3}e^{\frac{a-c}{2}\cdot\frac{1}{\max\{b,1\}}} \tag{57}$$

where the last inequality is by choosing $\theta = \frac{1}{2\max\{b,1\}}$. □

Lemma B.3 (Bias of Truncated Laplace Random Variable)**.** *Let $\lambda \ge 0$, $n \in \mathbb{N}$. Let $x \sim$ Poi(λ) and let $Z \sim$ Lap$(0,b)$. Then the noisy estimator given by $\tilde{x} = \max\{x+Z, c\}$ satisfies*

$$0 \le \mathbb{E}[\tilde{x}-\lambda] \le b + c \tag{58}$$

Proof. We first prove the left inequality in Lemma B.3. By definition,

$$\mathbb{E}[\tilde{x}-\lambda] = \mathbb{E}[\tilde{x}-x] = \mathbb{E}[Z\cdot\mathbf{1}_{x+Z\ge c} + (c-x)\cdot\mathbf{1}_{x+Z<c}] = \mathbb{E}[Z] + \mathbb{E}[(c-x-Z)\cdot\mathbf{1}_{x+Z<c}] \ge 0$$

We then prove the right inequality in Lemma B.3. By $\tilde{x} = \max\{x+Z, c\}$, we compute that

$$\mathbb{E}[\tilde{x}-\lambda] = \mathbb{E}[\tilde{x}-x] \le \mathbb{E}[\max\{c, |Z|\}] \le \mathbb{E}[|Z|] + c = b + c \tag{59}$$

□

Lemma B.4 (Conditional Bias under Thresholding)**.** *Let X be a random variable over $\mathbb{R}$. Then for any $c \in \mathbb{R}$,*

$$\mathbb{E}[X|X\ge c] \ge \mathbb{E}[X] \quad \textit{and}\ \mathbb{E}[X|X\le c] \le \mathbb{E}[X] \tag{60}$$

Proof. We first prove the first inequality in Lemma B.4. If $c \ge \mathbb{E}[X]$, then $\mathbb{E}[X|x\ge c] \ge c \ge \mathbb{E}[X]$. If $c < \mathbb{E}[X]$, then by definition,

$$\mathbb{E}[(X-\mathbb{E}[X])\cdot\mathbf{1}_{X>c}] \ge \mathbb{E}[(X-\mathbb{E}[X])] = 0 \tag{61}$$

where the inequality is by $x - \mathbb{E}[X] \le c - \mathbb{E}[X] \le 0$ for $x \le c$. The proof for the second inequality in Lemma B.4 is similar. □

Corollary B.5 (Conditional Bias of Truncated Sum of Poisson and Laplace Random Variable)**.** *Let $b > 0$, $\lambda > 0$, $c, d \in \mathbb{R}\cup\{-\infty, +\infty\}$. Let $x \sim$ Poi(λ) and let $Z \sim$ Lap$(0,b)$. Then the noisy estimator given by $\tilde{x} = \max\{x+Z, c\}$ satisfies*

$$\mathbb{E}[\tilde{x}-\lambda | x+Z\le d] \le b + c \tag{62}$$

and

$$\mathbb{E}[\tilde{x}-\lambda | x+Z\ge d] \ge 0 \tag{63}$$

Proof. We first prove (62). By Lemma B.4, we have that

$$\mathbb{E}\left[\tilde{x}-\lambda | x+Z \leq d\right] \leq \mathbb{E}\left[\tilde{x}-\lambda\right] \leq b+c$$

where the last inequality is by Lemma B.3. We then prove (63). By Lemma B.4, we have that

$$\mathbb{E}\left[\tilde{x}-\lambda | x+Z \geq d\right] \geq \mathbb{E}\left[\tilde{x}-\lambda\right] \geq 0$$

where the last inequality is by Lemma B.3. □

Lemma B.6. *Let $b > 0$ and $c \in \mathbb{R}$, and let $Z \sim$ Lap$(0, b)$. Then*

$$\mathbb{E}\left[Z | Z \geq c\right] \leq \max\{c, 0\} + b \tag{64}$$

Proof. We separate our discussion for $c \geq 0$ and $c < 0$.

1. If $c \geq 0$, we have that

$$\Pr[Z = z | Z \geq c] = \begin{cases} \frac{\frac{1}{2b}e^{-\frac{z}{b}}}{\frac{1}{2}e^{-\frac{c}{b}}} = \frac{1}{b}e^{-\frac{z-c}{b}} & z \geq c \\ 0 & z < c \end{cases} \tag{65}$$

   Thus

$$\mathbb{E}\left[Z | Z \geq c\right] = \int_{c}^{+\infty} z \cdot \frac{1}{b}e^{-\frac{z-c}{b}} dz \leq O\left(c+b\right) \tag{66}$$

2. If $c < 0$, we have that

$$\Pr[Z = z | Z \geq c] \leq \frac{\frac{1}{2b}e^{-\frac{|z|}{b}}}{\frac{1}{2}} = \frac{1}{b}e^{-\frac{|z|}{b}} \tag{67}$$

   Thus

$$\mathbb{E}\left[Z | Z \geq c\right] \leq \int_{c}^{+\infty} \frac{1}{b}e^{-\frac{|z|}{b}} dz \leq \int_{-\infty}^{+\infty} \frac{1}{b}e^{-\frac{|z|}{b}} dz \leq O\left(b\right) \tag{68}$$

□

Lemma B.7. *Let $\lambda > 0$, $b > 0$ and $c \in \mathbb{R}$. Let $X \sim$ Poi(λ) and $Z \sim$ Lap$(0, b)$ be independent random variables. Then*

$$\mathbb{E}\left[X+Z | X+Z \geq c\right] \leq \lambda + O\left(b + \max\{c, 0\}\right) \tag{69}$$

Proof. By definition,

$$\begin{aligned} \mathbb{E}\left[X+Z | X+Z \geq c\right] =& \sum_{t=0}^{\infty} \Pr[X=t] \cdot \left(t + \mathbb{E}\left[Z | Z \geq c-t\right]\right) && (70) \\ \leq& \mathbb{E}[X] + \sum_{t=0}^{\infty} \Pr[X=t] \cdot O\left(\max\{c-t, 0\} + b\right) && (71) \\ \leq& \lambda + O\left(b + \max\{c, 0\}\right) && (72) \end{aligned}$$

where (71) is by applying Lemma B.6

□

Lemma B.8 (KL Divergence between Poisson Distributions [57, Theorem 2])**.** *Let $m, k > 0$ be fixed. Then the KL divergence between two Poisson distributions Poi(m) and Poi(k) satisfies*

$$KL\left(Poi\left(m\right), Poi\left(k\right)\right) = m - k + k \cdot \ln\left(\frac{k}{m}\right) \tag{73}$$

Lemma B.9 (Total Variation Distance between Poisson Distributions)**.** *Let $\lambda_1, \lambda_2 > 0$ be fixed. If $\lambda_1 \le \lambda_2$, Then*

$$TV\left(Poi\left(\lambda_1\right), Poi\left(\lambda_2\right)\right) \le \frac{1}{2}\sqrt{\frac{(\lambda_1-\lambda_2)^2}{\lambda_1}} \tag{74}$$

Proof. By Pinsker's inequality, we have

$$\begin{aligned}
\mathrm{TV}\left(\mathrm{Poi}\left(\lambda_1\right), \mathrm{Poi}\left(\lambda_2\right)\right) \le & \sqrt{\frac{\mathrm{KL}\left(\mathrm{Poi}\left(\lambda_2\right), \mathrm{Poi}\left(\lambda_1\right)\right)}{2}} && (75)\\
\le & \sqrt{\frac{\lambda_2-\lambda_1+\lambda_1\cdot\ln\left(\frac{\lambda_1}{\lambda_2}\right)}{2}} && (76)\\
\le & \frac{1}{2}\sqrt{\frac{(\lambda_1-\lambda_2)^2}{\lambda_1}} && (77)
\end{aligned}$$

where (76) is by Lemma B.8, and the last inequality is by $\ln(1+x) \ge x - \frac{x^2}{2}$ for $x \ge 0$. □

Lemma B.10. *Let $p > 0$, $m \in \mathbb{N}$, $c_1, c_2 \in \mathbb{R} \cup \{+\infty, -\infty\}$, $b \ge 0$, and $c \ge \max\{b, 1\}$. Let $x \sim$ Poi(mp) and let $Z \sim$ Lap$(0, b)$ be independent of x. Then the noisy estimator given by $\tilde{x} = \max\{x + Z, c\}$ satisfies*

$$\mathbb{E}\left[\mathbf{1}_{c_1 \le x+Z \le c_2} \cdot \left(p\ln\left(\frac{mp}{\tilde{x}}\right) + \frac{\tilde{x}-mp}{m}\right)\right] \le O\left(\frac{1}{m} + \frac{b^2+c^2}{m\cdot\max\{c, mp\}}\right) \tag{78}$$

Proof. We separate the discussions for $p \le \frac{c}{m}$ and $p > \frac{c}{m}$.

- If $p \le \frac{c}{m}$, by $\ln(1+t) \le t$ for any $t > -1$, we have that

$$\begin{aligned}
& \mathbb{E}\left[\mathbf{1}_{c_1 \le x+Z \le c_2} \cdot \left(p\ln\left(\frac{mp}{\tilde{x}}\right) + \frac{\tilde{x}-mp}{m}\right)\right] && (79)\\
\le & \mathbb{E}\left[\mathbf{1}_{c_1 \le x+Z \le c_2} \cdot \left(\frac{(mp-\tilde{x})^2}{m\tilde{x}} + \frac{mp-\tilde{x}}{m} + \frac{\tilde{x}-mp}{m}\right)\right] && (80)\\
\le & \mathbb{E}\left[\frac{(mp-\tilde{x})^2}{mc}\right] && (81)\\
\le & \mathbb{E}\left[\frac{2(mp-x)^2 + 2(x-\tilde{x})^2}{mc}\right] && (82)\\
\le & \mathbb{E}\left[\frac{2mp}{mc} + \frac{4b^2+2c^2}{mc}\right] \le O\left(\frac{1}{m} + \frac{b^2+c^2}{mc}\right) && (83)
\end{aligned}$$

  where the (81) is by $\tilde{x} \ge c$ and by $(mp-\tilde{x})^2 \ge 0$, (82) is by $(a+b)^2 \le 2a^2 + 2b^2$, and the last inequality is by $\mathbb{E}[(mp-x)^2] = mp$, $\mathbb{E}\left[(x-\tilde{x})^2\right] \le \mathbb{E}[Z^2 + c^2] = 2b^2 + c^2$, and $p \le \frac{c}{m}$.

- If $p > \frac{c}{m}$: Observe that by $\ln(y) \ge y - 1 - (y-1)^2$ for any $y \ge \frac{1}{3}$, we have that

$$\begin{aligned}
& \mathbb{E}\left[\mathbf{1}_{c_1 \le x+Z \le c_2} \cdot \left(p\ln\left(\frac{mp}{\tilde{x}}\right) + \frac{\tilde{x}-mp}{m}\right) \cdot \mathbf{1}_{x+Z \ge \frac{1}{3}mp}\right] && (84)\\
\le & \mathbb{E}\left[\mathbf{1}_{c_1 \le x+Z \le c_2} \cdot \left(-p\cdot\frac{\tilde{x}-mp}{mp} + p\cdot\frac{(\tilde{x}-mp)^2}{(mp)^2} + \frac{\tilde{x}-mp}{m}\right) \cdot \mathbf{1}_{x+Z \ge \frac{1}{3}mp}\right] && (85)\\
= & \mathbb{E}\left[\mathbf{1}_{c_1 \le x+Z \le c_2} \cdot \frac{(\tilde{x}-mp)^2}{m^2p} \cdot \mathbf{1}_{x+Z \ge \frac{1}{3}mp}\right] && (86)\\
\le & \frac{2\mathbb{E}\left[Z^2 + c^2 + (x-mp)^2\right]}{m^2p} = O\left(\frac{b^2+c^2}{m^2p} + \frac{1}{m}\right) && (87)
\end{aligned}$$

where the first inequality in (87) is by $(a+b)^2 \le 2(a^2+b^2)$ and $(\tilde{x}-x)^2 \le Z^2+c^2$, and by $Z^2+c^2+(x-mp)^2 \ge 0$; the last equality in (87) is by $\mathbb{E}\left[Z^2\right] = \frac{1}{\varepsilon^2}$ and $\mathbb{E}\left[(x-mp)^2\right] = mp$.

On the other hand, by Lemma B.2, we have that $\Pr[x+Z < \frac{1}{3}mp] \le O\left(e^{-\frac{mp}{3}\cdot\frac{1}{\max\{b,1\}}}\right)$. Thus we have that

$$\mathbb{E}\left[\mathbf{1}_{c_1\le x+Z\le c_2}\cdot\left(p\ln\left(\frac{mp}{\tilde{x}}\right)+\frac{\tilde{x}-x}{m}\right)\cdot\mathbf{1}_{x+Z<\frac{1}{3}mp}\right] \le \frac{p\ln\left(\frac{mp}{c}\right)+p}{e^{\frac{mp}{3}\cdot\frac{1}{\max\{b,1\}}}} \tag{88}$$

$$=\frac{\max\{b,1\}}{m}\cdot\frac{x\ln(x)+x}{e^{x/3}}\Big|_{x=\frac{mp}{\max\{b,1\}}}+\ln\left(\frac{\max\{b,1\}}{c}\right)\cdot\frac{\max\{b,1\}}{n}\cdot\frac{x}{e^{x/3}}\Big|_{x=\frac{mp}{\max\{b,1\}}} \tag{89}$$

$$\le O\left(\frac{\max\{b,1\}^2}{m^2\cdot p}\right)\le O\left(\frac{c^2}{m^2\cdot p}\right) \tag{90}$$

where the inequality in (90) is by $\frac{x\ln(x)}{e^{x/6}} \le O\left(\frac{1}{x}\right)$ and $\frac{x}{e^{x/6}} \le O\left(\frac{1}{x}\right)$ for $x \ge 0$, and by $c \ge \max\{b,1\}$ This suffice to prove the bound in the statement.

□

C Deferred Minimax Optimality Results

C.1 Recap: Non-DP Minimax Results

The minimax rate for non-DP distribution estimation in KL divergence error is well-studied [10, 51, 52], as summarized below in Table 2.

Estimator/Bound	Expected KL Error	Reference
Upper Bound (Add-Constant Estimator)	$\ln\left(1+\frac{d}{n}\right)$	[10, Theorem 8] Recap: Theorem C.2
Lower Bound	$\Omega\left(\ln\left(1+\frac{d}{n}\right)\right)$	Recap: Theorem C.1

Table 2: Non-DP Minimax Rates for Distribution Estimation in KL Divergences

For completeness, below we offer simple proofs for the minimax upper and lower bounds.

C.1.1 Recap: Non-DP Minimax KL Error Lower Bound

The minimax lower bound for distribution estimation in KL divergence error is well-understood to be $\Omega(1+\frac{d}{n})$, e.g., see the variational lower bounds in [52]. For completeness, below we provide an alternative proof via the tools in this paper, i.e., the generalized DP Assouad's method for decomposable statistical distance (in our case KL divergence) in Lemma A.2.

Theorem C.1 (Lower Bound for Non-DP Estimation)**.** *Let d, n be fixed. Then for any estimator $\mathcal{A}$, we have that*

$$\max_{p}\mathop{\mathbb{E}}_{x\sim \textit{Mult}(n,p)}\Big[KL(p\|\mathcal{A}(x))\Big] \ge \begin{cases}\Omega\left(\frac{d-1}{n}\right) & d-1\le 2n\\ \Omega\left(\ln\left(\frac{d-1}{n}\right)\right) & d-1>2n\end{cases} = \Omega\left(\ln\left(1+\frac{d}{n}\right)\right) \tag{91}$$

Proof sketch The idea is to design p to be uniform over a random support of symbols with small probability mass, and then reduce the problem to the difficulty of inferring the support given limited samples.

Proof. • **Dense Case** $d-1\le 2n$**:** Assume $d \mod 2 = 1$ for convenience. We decompose d symbols into $\frac{d-1}{2}$ buckets $\mathcal{B}_i = \{2i-1, 2i\}$ for $i = 1, \cdots, \frac{d-1}{2}$. We then construct the

packing as follows. For $i = 1, \cdots, \frac{d-1}{2}$.

$$\text{For } i = 1, \cdots, \frac{d-1}{2} \quad \mathcal{P}_i = \{\delta_{2i-1}, \delta_{2i}\} \tag{92}$$

where δ_i means point distribution on symbol i. We then construct a set of distributions.

$$\mathcal{P} = \left\{ p = \sum_{i=1}^{\frac{d-1}{2}} \frac{1}{n} \cdot p^i + \left(1 - \frac{d-1}{2n}\right) \cdot \delta_d : p^i \in \mathcal{P}_i \text{ for any } i = 1, \cdots, \frac{d-1}{2} \right\} \tag{93}$$

Then by Lemma A.2, for any algorithm $\mathcal{A}$ we have that

$$\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathop{\mathbb{E}}_{x \sim \text{Mult}(n,p)} [KL(p, \mathcal{A}(x))] \geq \frac{1}{n|\mathcal{P}|} \sum_{i=1}^{\frac{d-1}{2}} \sum_{p \in \mathcal{P}} \mathop{\mathbb{E}}_{x \sim \text{Mult}(n,p)} [KL(p|_i, \mathcal{A}(x)|_i)] \tag{94}$$

$$\geq \frac{1}{ne|\mathcal{P}|} \sum_{i=1}^{\frac{d-1}{2}} \sum_{p \in \mathcal{P}} \mathop{\mathbb{E}}_{x \sim \text{Mult}(n,p)} [KL(p|_i, \mathcal{A}(x)|_i) | x_{2i-1} = x_{2i} = 0] \tag{95}$$

$$\geq \frac{1}{ne} \sum_{i=1}^{\frac{d-1}{2}} \ln(2) = \frac{(d-1)\ln(2)}{2ne} \tag{96}$$

$$\tag{97}$$

Thus for any $\mathcal{A}$, there must exist one $p \in \mathcal{P}$ such that $\mathop{\mathbb{E}}_{x \sim \text{Mult}(n,p)} [KL(p, \mathcal{A}(x))] \geq \Omega(\frac{d-1}{n})$.

- **Sparse Case:** $d - 1 > 2n$**:** Denote $\kappa = \lfloor \frac{d-1}{n} \rfloor \geq 2$ and $k = n$ for convenience. We decompose d symbols into k buckets $\mathcal{B}_i = \{\kappa \cdot i - \kappa + 1, \kappa \cdot i\}$ for $i = 1, \cdots, k$. We then construct the packing as follows.

$$\text{For } i = 1, \cdots, k \quad \mathcal{P}_i = \{\delta_{\kappa \cdot i - \kappa + 1}, \cdots, \delta_{\kappa \cdot i}\} \tag{98}$$

where δ_i means point distribution on symbol i. We then construct a set of distributions.

$$\mathcal{P} = \left\{ p = \sum_{i=1}^{k} \frac{1}{n} \cdot p^i : p^i \in \mathcal{P}_i \text{ for any } i = 1, \cdots, k \right\} \tag{99}$$

Then by Lemma A.2, for any algorithm $\mathcal{A}$ we have that

$$\frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} \mathop{\mathbb{E}}_{x \sim \text{Mult}(n,p)} [KL(p, \mathcal{A}(x))] \geq \frac{1}{n|\mathcal{P}|} \sum_{i=1}^{k} \sum_{p \in \mathcal{P}} \mathop{\mathbb{E}}_{x \sim \text{Mult}(n,p)} [KL(p|_i, \mathcal{A}(x)|_i)] \tag{100}$$

$$\geq \frac{1}{ne|\mathcal{P}|} \sum_{i=1}^{k} \sum_{p \in \mathcal{P}} \mathop{\mathbb{E}}_{x \sim \text{Mult}(n,p)} [KL(p|_i, \mathcal{A}(x)|_i) | x_{\kappa i - \kappa + 1} = \cdots = x_{\kappa i} = 0] \tag{101}$$

$$\geq \frac{1}{ne} \sum_{i=1}^{k} \ln(\kappa) = \frac{k \ln \kappa}{ne} = \frac{1}{e} \ln\left(\frac{d-1}{n}\right) \tag{102}$$

$$\tag{103}$$

Thus for any $\mathcal{A}$, there must exist one $p \in \mathcal{P}$ such that $\mathop{\mathbb{E}}_{x \sim \text{Mult}(n,p)} [KL(p, \mathcal{A}(x))] \geq \Omega(\ln\left(\frac{d-1}{n}\right))$.

□

C.1.2 Recap: Non-DP Minimax KL Error Upper Bound

The minimax upper bound for distribution estimation is well-studied, and various works [10, 51, 52] have shown that simple add-constant estimators could achieve the optimal $O\left(\ln\left(1 + \frac{d}{n}\right)\right)$ minimax KL error. Below we offer a simple proof for completeness.

Theorem C.2 (Non-DP Distribution Estimation KL Upper Bound)**.** *There exists estimator $\mathcal{A}$ such that for any $p \in \Delta(d)$, given n empirical data samples $x \sim$ Mult(n, p), it satisfies that*

$$\mathop{\mathbb{E}}_{x\sim \textit{Mult}(n,p)}\Big[KL(p\|\mathcal{A}(x))\Big] \le \ln\left(1+\frac{d}{n}\right) \tag{104}$$

Proof. We use that the following simple add-constant estimator.

$$\mathcal{A}(x)_i = \frac{1}{1+\frac{d}{n}} \cdot \frac{x_i+1}{n} \tag{105}$$

Then by definition

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim \text{Mult}(n,p)}\Big[KL(p\|\mathcal{A}(x))\Big] =& \sum_{i=1}^{d} p_i \mathop{\mathbb{E}}_{x\sim \text{Mult}(n,p)}\Big[\ln\left(\frac{np_i}{x_i+1}\right)\Big] + \ln\left(1+\frac{d}{n}\right) && (106)\\
\le& \sum_{i=1}^{d} p_i \ln\left(np_i \cdot \mathop{\mathbb{E}}_{x\sim \text{Mult}(n,p)}\Big[\frac{1}{x_i+1}\Big]\right) + \ln\left(1+\frac{d}{n}\right) && (107)\\
\le& \sum_{i=1}^{d} \ln\left(\frac{np_i}{\mathop{\mathbb{E}}_{x\sim \text{Mult}(n,p)}[x_i]}\right) + \ln\left(1+\frac{d}{n}\right) && (108)\\
=& \ln\left(1+\frac{d}{n}\right) && (109)
\end{aligned}$$

where (107) is by concavity of the function $\ln(x)$, and (108) is by Lemma C.3. □

Lemma C.3. *Let $s_1, \cdots, s_K$ be K independent Bernoulli random variables with parameter $p_1, \cdots, p_K$. Let $c \ge 1$ be a positive constant. Then we have that*

$$\mathbb{E}\left[\frac{1}{c+\sum_{k=1}^{K} s_k}\right] \le \frac{1}{\sum_{k=1}^{K} p_k} \tag{110}$$

Proof. Conditioned on any fixed value for $s_3, \cdots, s_K$, denote $c' = c + \sum_{k=3}^{K} s_k$ we have that

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{c+\sum_{k=1}^{K} s_k}\Big| s_3, \cdots, s_K\right] = \frac{p_1p_2}{c'+2} + \frac{p_1(1-p_2)+(1-p_1)p_2}{c'+1} + \frac{(1-p_1)(1-p_2)}{c'} && (111)\\
=&p_1p_2\left(\frac{1}{c'+2} - \frac{2}{c'+1} + \frac{1}{c'}\right) + (p_1+p_2)\cdot\left(\frac{1}{c'+1} - \frac{1}{c'}\right) + \frac{1}{c'} && (112)\\
\le&\left(\frac{p_1+p_2}{2}\right)^2\left(\frac{1}{c'+2} - \frac{2}{c'+1} + \frac{1}{c'}\right) + (p_1+p_2)\cdot\left(\frac{1}{c'+1} - \frac{1}{c'}\right) + \frac{1}{c'} && (113)
\end{aligned}$$

where the last inequality is by $\frac{1}{c'+2} - \frac{2}{c'+1} + \frac{1}{c'} > 0$ for $c' > 0$. By similar argument, we have that

$$\mathbb{E}\left[\frac{1}{c+\sum_{k=1}^{K} s_k}\right] \le \mathbb{E}\left[\frac{1}{c+s}\right] \le \mathbb{E}\left[\frac{1}{1+s}\right] \tag{114}$$

where $s \sim \text{Bin}(K, \bar{p})$ where $\bar{p} = \frac{1}{K}\sum_{k=1}^{K} p_k$. Then by simple algebraic tricks, we have that

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{c+\sum_{k=1}^{K} s_k}\right] \le& \sum_{j=0}^{K}\binom{K}{j}\bar{p}^j(1-\bar{p})^{K-j}\cdot\frac{1}{j+1} && (115)\\
=&\frac{1}{(K+1)\bar{p}}\sum_{j=0}^{K}\binom{K+1}{j+1}\bar{p}^{j+1}(1-\bar{p})^{K-j} && (116)\\
<&\frac{1}{K\bar{p}} = \frac{1}{\sum_{k=1}^{K} p_k} && (117)
\end{aligned}$$

□

Method	KL Error Bound	Conditions	Reference
Any (ε, δ)-DP	$\Omega\left(\ln(1+\frac{d}{n\min\{\varepsilon,1\}})\right)$	$d \geq 4, \delta \leq \varepsilon$	Theorem C.4
Add-constant ((ε, δ)-DP) (Laplace Mechanism)	$O\left(\ln\left(1+\frac{d}{n\min\{\varepsilon,1\}}\right)\right)$	$\varepsilon > 0, \delta \geq 0$	Theorem C.5

Table 3: DP Minimax KL error bounds

C.2 Our DP Minimax Results

We first summarize our derived DP minimax rates in Table 3 – observe that simple variants of Laplace mechanism is minimax optimal.

We comment that our KL minimax lower bound is stronger than naive conversions of prior TV distance lower bounds to KL lower bounds – applying Pinsker inequality to the prior $\Omega\left(\frac{d}{\varepsilon n}\right)$ TV lower bounds [19, 3] only gives a KL lower bound of $\Omega\left(\min\left\{\frac{d^2}{\varepsilon^2 n^2}, 1\right\}\right)$, which is significantly weaker than our bound $\Omega\left(\ln\left(1+\frac{d}{\varepsilon n}\right)\right)$ in Table 3, especially for large d.

Below we present the proofs for the minimax lower bounds and upper bounds.

C.2.1 DP Minimax Lower Bound

Theorem C.4 (Minimax Lower Bound for DP estimation)**.** *Let $d, n \in \mathbb{N}$, $\varepsilon \geq 0$, and $\delta \in [0,1]$ be fixed. If $d \geq 4$, $\delta \leq \varepsilon$, then*

$$\max_{p\in\Delta(d)} \mathop{\mathbb{E}}_{x\sim Poi(n,p)}\Big[KL(p, \mathcal{A}(x))\Big] \geq \Omega\left(\ln\left(1+\frac{d}{\varepsilon n}\right)\right) \tag{118}$$

Proof. Let $\kappa, k \in \mathbb{N}$ be defined as follows.

$$\kappa = \begin{cases} 2 & \frac{d}{160n\varepsilon} \leq 2 \\ \lfloor \frac{d}{160n\varepsilon} \rfloor & \frac{d}{160n\varepsilon} > 2 \text{ and } 80n\varepsilon \geq 1 \\ \lfloor \frac{d}{2} \rfloor & \frac{d}{160n\varepsilon} > 2 \text{ and } 80n\varepsilon < 1 \end{cases} \text{ and } k = \begin{cases} \lfloor \frac{d}{4} \rfloor & \frac{d}{160n\varepsilon} \leq 2 \\ \lfloor 80n\varepsilon \rfloor & \frac{d}{160n\varepsilon} > 2 \text{ and } 80n\varepsilon \geq 1 \\ 1 & \frac{d}{160n\varepsilon} > 2 \text{ and } 80n\varepsilon < 1 \end{cases}. \tag{119}$$

Thus by $d \geq 2$, we have

$$\kappa \geq 2 \quad \text{and} \quad k \geq 1 \quad \text{and} \quad \kappa \cdot k \leq \frac{d}{2} \quad \text{and} \quad \frac{k}{160n\varepsilon} \leq \frac{1}{2} \tag{120}$$

For $i = 1, \cdots, k$, we construct a packing set of distributions supported on symbols $\mathcal{B}_i = \{k \cdot i - \kappa + 1, \cdots, \delta_{\kappa \cdot i}\}$ as follows.

$$\text{For } i = 1, \cdots, k \quad \mathcal{P}_i = \{\delta_{\kappa \cdot i - \kappa + 1}, \cdots, \delta_{\kappa \cdot i}\} \tag{121}$$

We construct the following set of distributions that lie in the additive neighborhood of p.

$$\mathcal{P} = \left\{ q = \left(1 - \frac{k}{160n\varepsilon}\right) \cdot q^c + \sum_{i=1}^{k} w_i \cdot q^i : w_i = \frac{1}{160n\varepsilon} \text{ and } q^i \in \mathcal{P}_i \text{ for any } i = 1, \cdots, k \right\} \tag{122}$$

where

$$q_c(j) = \begin{cases} 0 & j = 1, \cdots, \kappa \cdot k \\ \frac{1}{d - \kappa \cdot k} & j = \kappa \cdot k + 1, \cdots, d \end{cases} \tag{123}$$

One can verify that the distributions in $\mathcal{P}$ are well-defined (i.e., normalized). Now by applying Lemma A.6 to $\mathcal{P}_i$ for each $i = 1, \cdots, k$, we prove that for any $q' \in \Delta(d)$, it holds that

$$\frac{1}{|\mathcal{P}_i|} \sum_{q \in \mathcal{P}_i} \mathbf{1}_{KL(q,q') \geq f(\mathcal{P}_i)} \geq \frac{1}{2} \text{ where } f(\mathcal{P}_i) = \ln\left(1 + \frac{\kappa}{4}\right) \tag{124}$$

Thus the first condition of Theorem A.4 holds. Below we analyze the second condition of Theorem A.4. For each $i \in [k]$ and fixed $q^j \in \mathcal{P}_j, j = 1, \cdots, k$ and fixed $\bar{q}^i \in \mathcal{P}_i$. Let $x \sim \text{Poi}\left(n, \left(1 - \frac{k}{160n\varepsilon}\right) \cdot q^c + \sum_{j=1}^{k} w_j \cdot q^j\right)$ and $y \sim \text{Poi}\left(n, \left(1 - \frac{k}{160n\varepsilon}\right) \cdot q^c + w_i \cdot \bar{q}^i + \sum_{j \neq i} w_j \cdot q^j\right)$ be independent Poisson random variables. Then we could construct the following coupling $(x, \bar{x})$ between distributions $\text{Poi}\left(n, \left(1 - \frac{k}{160n\varepsilon}\right) \cdot q^c + \sum_{j=1}^{k} w_j \cdot q^j\right)$ and $\text{Poi}\left(n, \left(1 - \frac{k}{160n\varepsilon}\right) \cdot q^c + w_i \cdot \bar{q}^i + \sum_{j \neq i} w_j \cdot q^j\right)$:

$$\bar{x}_l = \begin{cases} y_l & l = \kappa \cdot i - \kappa + 1, \cdots, \kappa \cdot i \\ x_l & l \in [d] \setminus \{\kappa \cdot i - \kappa + 1, \cdots, \kappa \cdot i\} \end{cases} \tag{125}$$

By definition, we compute that

$$\mathbb{E}\left[\|x - \bar{x}\|_1\right] = \sum_{l=\kappa \cdot i - \kappa + 1}^{\kappa \cdot i} \mathbb{E}\left[|x_l - y_l|\right] \leq \sum_{l=\kappa \cdot i - \kappa + 1}^{\kappa \cdot i} \mathbb{E}\left[x_l + y_l\right] = \frac{1}{80\varepsilon} \coloneqq \tau \tag{126}$$

where the inequality is by triangle inequality for ℓ_1 distance. By applying Theorem A.4 under our proved conditions (124) and (126), we finally prove that

$$\max_{q \in \mathcal{P}} \mathbb{E}\left[KL(q, \mathcal{A}(x))\right] \geq \sum_{i=1}^{k} w_i \cdot \left(\frac{1}{10} - 4\varepsilon \cdot \tau\right) \cdot f(\mathcal{P}_i) = \frac{k}{160n\varepsilon} \cdot \frac{1}{20} \cdot \ln\left(1 + \frac{\kappa}{4}\right) \tag{127}$$

$$= \begin{cases} \frac{1}{160 \cdot 20n\varepsilon} \lfloor \frac{d}{4} \rfloor \cdot \ln\left(1 + \frac{1}{2}\right) & \frac{d}{160n\varepsilon} \leq 2 \\ \frac{1}{160 \cdot 20n\varepsilon} \cdot \lfloor 80n\varepsilon \rfloor \cdot \ln\left(1 + \frac{1}{4} \lfloor \frac{d}{160 \cdot n\varepsilon} \rfloor\right) & \frac{d}{160n\varepsilon} > 2 \text{ and } 80n\varepsilon \geq 1 \\ \frac{1}{160 \cdot 20n\varepsilon} \cdot \ln\left(1 + \frac{1}{4} \lfloor \frac{d}{2} \rfloor\right) & \frac{d}{160n\varepsilon} > 2 \text{ and } 80n\varepsilon < 1 \end{cases} \tag{128}$$

$$\geq \begin{cases} \frac{d}{160 \cdot 20n\varepsilon \cdot 8} \ln\left(1 + \frac{1}{2}\right) & \frac{d}{160n\varepsilon} \leq 2 \\ \frac{1}{80} \cdot \lfloor 80n\varepsilon \rfloor \cdot \ln\left(1 + \frac{d}{8 \cdot 160 \cdot n\varepsilon}\right) & \frac{d}{160n\varepsilon} > 2 \text{ and } 80n\varepsilon \geq 1 \\ \frac{1}{160 \cdot 20n\varepsilon} \cdot \ln\left(1 + \frac{d}{16}\right) & \frac{d}{160n\varepsilon} > 2 \text{ and } 80n\varepsilon < 1 \end{cases} \tag{129}$$

$$\geq \Omega\left(\ln\left(1 + \frac{d}{n\varepsilon}\right)\right) \tag{130}$$

where (128) is by using the definitions of κ and k in (119), (129) is by $\lfloor x \rfloor \geq \frac{x}{2}$ for any $x \geq 1$, (130) is by $\ln(1 + \lambda x) \leq \lambda \cdot \ln(1 + x) \leq x$ for $0 \leq \lambda \leq 1$ and $x > 0$, and by $\lambda \ln(1 + x) \geq \ln(1 + \lambda x)$ for $\lambda > 1$ and $x > 0$. □

C.2.2 DP Minimax Upper Bound

Theorem C.5 (Upper Bound - Pure DP)**.** *There exists an (ε, δ)-DP estimator $\mathcal{A}$ such that for any fixed $d, n \in \mathbb{N}$,*

$$\mathop{\mathbb{E}}_{x \sim Poi(n,p)}\left[KL(p \| \mathcal{A}(x))\right] \leq O\left(\ln\left(1 + \frac{d}{n \min\{1, \varepsilon\}}\right)\right) \tag{131}$$

Proof. This can be achieved by a simple estimator that combines an add-constant estimator and Laplace mechanism as follows.

$$\mathcal{A}(x)_i = \frac{1}{N} \cdot \frac{\tilde{x}_i}{n} \text{ where } \tilde{x}_i = \max\left\{x_i + z_i, \frac{1}{\min\{1, \varepsilon\}}\right\} \tag{132}$$

where $z \sim Lap\left(0, \frac{1}{\varepsilon}\right)^d$, and $N = \sum_i \tilde{x}_i$ is the normalization constant. The $(\varepsilon, 0)$-DP guarantee follows by observing that the ℓ_1-sensitivity of vector release $(x_1, \cdots, x_d)$ is one, and by applying the DP guarantee for Laplace Mechanism in [26, Theorem 1]. Below we focus on bounding the KL

Neighborhood N	lower(p, n, N) (2)	Conditions	Reference
$N_{\le t/n}(p)$ (136)	$\Omega\Big(\frac{\ln(1+d_{small}(L'))}{n} + p_{small}(L')\ln\Big(1+\frac{d_{small}(L')}{np_{small}(L')}\Big)\Big)$	$L' \subseteq [d], t \ge 1$ $d \ge 2, n \ge 4$	Theorem D.3
$N_{stat}(p)$ (137)	$\Omega\Big(\sum_i \min\big\{p_i, \frac{1}{n}\big\}\Big)$	$d \ge 2$	Theorem D.4

Table 4: Non-DP per-instance lower bounds, where $p_{small}(L') = \sum_{i\in L', p_i \le \frac{t}{n}} p_i$ and $d_{small}(L') = \sum_{i\in L', p_i\le \frac{t}{n}} 1$.

error. By concavity of $\ln(t)$ on $t > 0$, we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{x\sim\text{Poi}(n,p)}[KL(p\|\mathcal{A}(x))] \le& \mathop{\mathbb{E}}_{x\sim\text{Poi}(n,p)}\left[\sum_i p_i \ln\left(\frac{p_i\cdot n}{\tilde{x}_i}\right)\right] + \ln\left(1+\sum_i \frac{\mathbb{E}\left[\tilde{x}_i - p_i\right]}{n}\right) \\
\le& \mathop{\mathbb{E}}_{x\sim\text{Poi}(n,p)}\left[\sum_{i:p_i > \frac{1}{n\min\{1,\varepsilon\}}} p_i\ln\left(\frac{p_i\cdot n}{\tilde{x}_i}\right) + \frac{\tilde{x}_i - p_i}{n}\right] + \ln\left(1+\sum_{i:p_i<\frac{1}{n\min\{1,\varepsilon\}}}\frac{\mathbb{E}\left[\tilde{x}_i-p_i\right]}{n}\right) \quad (133)\\
\le& O\left(\frac{\sum_{i:p_i>\frac{1}{n\min\{1,\varepsilon\}}} 1}{n\min\{1,\varepsilon\}}\right) + \ln\left(1+O\left(\frac{\sum_{i:p_i<\frac{1}{n\min\{1,\varepsilon\}}}1}{n\min\{1,\varepsilon\}}\right)\right) \quad (134)\\
\le& O\left(1+\frac{d}{n\min\{1,\varepsilon\}}\right) \quad (135)
\end{aligned}$$

where (133) is by $p_i\ln\left(\frac{p_i\cdot n}{\tilde{x}_i}\right) \le 0$ for $p_i < \frac{1}{n\min\{1,\varepsilon\}}$, by $\mathbb{E}[\tilde{x}_i - x_i] \ge 0$ for any i (by Lemma B.3), and by $\ln(1+x+y) \le x + \ln(1+y)$ for any $x, y > 0$; (134) is by applying Lemma B.10 under setting $m = n$, $c_1 = -\infty$, $c_2 = +\infty$, $b = \frac{1}{\varepsilon}$ and $c = \frac{1}{\min\{1,\varepsilon\}}$ for $p_i > \frac{1}{n\min\{1,\varepsilon\}}$ and by applying Lemma B.3 for $p_i < \frac{1}{n\min\{1,\varepsilon\}}$; and (135) is by $a + \ln(1+b) \le \ln(1+2a+2b)$ for $0 < a < 1$ and $b > 0$. □

D Deferred Proofs for Non-DP Per-Instance Lower Bound Theorem 2.1

For ease of presentation and understanding, we break up the neighborhood $N_+(p)$ into two sub-neighborhoods: one $N_{\le\frac{t}{n}} \subseteq N_+(p)$ with the same small perturbation for every symbol, and the other $N_{stat} \subseteq N_+(p)$ (based on statistical variance) with larger perturbations for large symbols, defined as follows.

$$N_{\le\frac{t}{n}}(p) = \left\{q : |q_i - p_i| \le \frac{t}{n} \text{ for any } i\in[d] \text{ and } \sum_{i:p_i\le\frac{t}{n}} q_i \le \max\left\{\frac{t}{n}, \sum_{i:p_i\le\frac{t}{n}} p_i\right\}\right\} \quad (136)$$

$$N_{stat}(p) = \left\{q : |q_i - p_i| \le \min\left\{p_i, \sqrt{\frac{p_i}{n}}\right\} \text{ for any } i\in[d]\right\} \cup \left\{q : \sum_{i=1}^d |q_i - p_i| \le \frac{1}{n}\right\} \quad (137)$$

We then prove per-instance lower bounds under the sub-neighborhoods $N_{\le\frac{t}{n}}(p)$ and $N_{stat}(p)$ respectively. By definition (2), their average is a lower bound for the per-instance lower bound under $N_+(p)$, i.e., lower$(p, n, N_+) \ge \frac{1}{2}\cdot$lower$(p, n, N_{\le\frac{t}{n}}) + \frac{1}{2}\cdot$lower$(p, n, N_{stat})$. (This is because one can construct a distribution over hard instances, choosing the hard instance(s) in $N_{\le\frac{t}{n}}(p)$ and $N_{stat}(p)$ with $1/2$ probability respectively.) Our per-instance lower bounds under the two sub-neighborhood are summarized in Table 4.

D.1 Useful Lemmas

Lemma D.1 (Coupling Lemma [5, Lemma 3.6])**.** *Let random variables Z_1, Z_2 have distributions ν_1, ν_2. Then*

$$TV(\nu_1, \nu_2) \le \Pr[Z_1 \neq Z_2]. \tag{138}$$

Conversely, given probability distributions ν_1, ν_2, there exists (Z_1, Z_2) such that

$$TV(\nu_1, \nu_2) = \Pr[Z_1 \neq Z_2] \tag{139}$$

where Z_1, Z_2 have distributions ν_1, ν_2 respectively.

Lemma D.2 (Total Variation Inequality between Product Measures)**.** *Let μ_1, μ_2 be probability distributions over Ω and let μ_1', μ_2' be probability distributions over Ω'. Denote $\mu_1 \times \mu_1'$ as the product measure of μ_1 and μ_1' over $\Omega \times \Omega'$, and similarly denote $\mu_2 \times \mu_2'$ as the product measure of μ_2 and μ_2' over $\Omega \times \Omega'$. Then*

$$TV(\mu_1 \times \mu_1', \mu_2 \times \mu_2') \le TV(\mu_1, \mu_2) + TV(\mu_1', \mu_2') \tag{140}$$

Proof. By the coupling lemma Lemma D.1, there exists (Z_1, Z_1') such that $Z_1 \sim \mu_1$, $Z_1' \sim \mu_1'$ and $\mathrm{TV} = \Pr[Z_1 \neq Z_1']$. Similarly, there exists (Z_2, Z_2') such that $Z_2 \sim \mu_2$, $Z_2' \sim \mu_2'$ and $\mathrm{TV} = \Pr[Z_2 \neq Z_2']$. Let $Z = (Z_1, Z_1')$ and $Z' = (Z_2, Z_2')$. Then $Z \sim \mu_1 \times \mu_1'$ and $Z' \sim \mu_2 \times \mu_2'$. By again using the coupling lemma Lemma D.1, we prove that

$$\mathrm{TV}(\mu_1 \times \mu_1', \mu_2 \times \mu_2') \le \Pr[Z \neq Z'] \le \Pr[Z_1 \neq Z_1'] + \Pr[Z_2 \neq Z_2'] = \mathrm{TV}(\mu_1, \mu_2) + \mathrm{TV}(\mu_1', \mu_2') \tag{141}$$

where the second inequality is by union bound. □

D.2 Non-DP Per-Instance Lower Bound under $N_{\le \frac{t}{n}}(p)$

We now prove the per-instance lower bound in Table 4 under additive neighborhood $N_{\le \frac{t}{n}}(p)$ for low-probability symbols.

Theorem D.3. *Let $p \in \Delta(d)$ be fixed. For $t > 0$, let $N_{\le \frac{t}{n}}(p)$ be the below additive local neighborhood of p as defined in* (136)*.*

$$N_{\le \frac{t}{n}}(p) = \left\{ q : |q_i - p_i| \le \frac{t}{n} \text{ and } \sum_{i: p_i \le \frac{t}{n}} q_i \le \max\left\{ \frac{t}{n}, \sum_{i: p_i \le \frac{t}{n}} p_i \right\} \right\} \tag{142}$$

If $t \ge 1$, $d \ge 2$ and $n \ge 4$, then for any $L' \subseteq [d]$,

$$\max_{q \in N_{\le \frac{t}{n}}(p)} \mathop{\mathbb{E}}_{x \sim Poi(n,q)} KL(q, \mathcal{A}(x)) \ge \Omega\left(\frac{\ln(1 + d_{small}(L'))}{n} + p_{small}(L') \ln\left(1 + \frac{d_{small}(L')}{n p_{small}(L')}\right)\right) \tag{143}$$

where $p_{small}(L') = \sum_{i \in L', p_i \le \frac{t}{n}} p_i$ and $d_{small}(L') = \sum_{i \in L', p_i \le \frac{t}{n}} 1$.

Proof. We separate the discussions for different $d_{small}(L')$.

1. If $d_{small}(L') = 0$, then $p_{small}(L') = 0$ and thus (143) trivially holds.
2. If $1 \le d_{small}(L') \le 3$: Without loss of generality, assume that $\{i \in L' : p_i \le \frac{t}{n}\} = \{1, \cdots, d_{small}(L')\}$.
   (a) If $\max_{i \in [d] \setminus [d_{small}(L')]} p_i \le \frac{1}{n}$: Then the neighborhood N_{stat} defined in (137) is a subset of $N_{\le \frac{t}{n}}$ defined in (142), i.e., $N_{stat} \subseteq N_{\le \frac{t}{n}}$. Thus (143) holds by observing that (143) is dominated by the lower bound in Theorem D.4. Specifically, by $\ln(x) \le x - 1$ for any $x > 0$ and by $d_{small}(L') \le 3$, we prove that $\frac{\ln(1+d_{small}(L'))}{n} + p_{small}(L') \ln\left(1 + \frac{d_{small}(L')}{n p_{small}(L')}\right) \le \frac{2 d_{small}(L')}{n} \le \frac{6}{n} \le O\left(\sum_{i=1}^{d} \min\left\{p_i, \frac{1}{n}\right\}\right)$ where the last inequality is by $d \ge 2$.

(b) If $\max_{i\in[d]\setminus[d_{small}(L')]} p_i > \frac{1}{n}$, without loss of generality, assume that $p_d > \frac{1}{n}$: Then we construct a packing set of distributions $\mathcal{P} = \{p^+, p^-\}$ that contains the following two distributions.

$$p^+(j) = \begin{cases} \frac{1}{2n} & j = 1 \\ p_j + \frac{p_1}{d-1} & j = 2, \cdots, d-1 \\ p_d - \frac{1}{2n} + \frac{p_1}{d-1} & j = d \end{cases} \tag{144}$$

and

$$p^-(j) = \begin{cases} \frac{1}{4n} & j = 1 \\ p_j + \frac{p_1}{d-1} & j = 2, \cdots, d-1 \\ p_d - \frac{1}{4n} + \frac{p_1}{d-1} & j = d \end{cases} \tag{145}$$

One can validate that the distributions in $\mathcal{P}$ are well-defined and are in the additive neighborhood $N_{\leq \frac{t}{n}}(p)$ (by observing that $p_1 \leq \frac{t}{n}$ and $d \geq 2$, and thus $\frac{p_1}{d-1} \leq \frac{t}{n}$). By applying Lemma A.5 (under setting $a = \frac{1}{2n}$ and $\Delta = \frac{1}{4n}$), we further prove that for any $q' \in \Delta(d)$,

$$\frac{1}{|\mathcal{P}|}\sum_{q\in\mathcal{P}} \mathbf{1}_{KL(q,q')\geq f(\mathcal{P})} \geq \frac{1}{2} \text{ where } f(\mathcal{P}) = \frac{\Delta^2}{a} = \frac{1}{8n} \tag{146}$$

Thus the first condition of Theorem A.3 holds. Below we analyze the second condition of Theorem A.3. By definition of $\mathcal{P}$ we compute that

$$\int \left(\min_{q\in\mathcal{P}} \frac{d\text{Poi}(n,q)}{d\text{Poi}(n,p^+)}\right) d\text{Poi}\left(n,p^+\right) = \int \min\left\{d\text{Poi}\left(n,p^+\right), d\text{Poi}\left(n,p^-\right)\right\} \tag{147}$$

$$= 1 - \text{TV}\left(\text{Poi}\left(n,p^+\right), \text{Poi}\left(n,p^-\right)\right) \tag{148}$$

$$\geq 1 - \text{TV}\left(\text{Poi}\left(n\cdot\frac{1}{2n}\right), \text{Poi}\left(n\cdot\frac{1}{4n}\right)\right) - \text{TV}\left(\text{Poi}\left(np_d - \frac{1}{2} + \frac{np_1}{d-1}\right), \text{Poi}\left(np_d - \frac{1}{4} + \frac{np_1}{d-1}\right)\right) \tag{149}$$

$$\geq 1 - \frac{1}{2}\sqrt{\frac{n^2\left(\frac{1}{4n}\right)^2}{n\cdot\frac{1}{4n}}} - \frac{1}{2}\sqrt{\frac{\left(\frac{1}{4}\right)^2}{np_d - \frac{1}{2} + \frac{p_1}{d-1}}} \tag{150}$$

$$\geq 1 - \frac{1}{4} - \frac{1}{4\sqrt{2}} > \frac{1}{2} \coloneqq \tau \tag{151}$$

where (147) is by definition for $\mathcal{P}$ that only contains two distributions, (148) is by using the definition of total variation distance, (149) is by the total variation inequality for product measures (Lemma D.2), (150) is by Lemma B.9. Thus the second condition of Theorem A.3 holds. By applying Theorem A.3 under our proved conditions (146) and (151), we finally prove that

$$\max_{q\in\mathcal{P}} \mathbb{E}\left[KL(q, \mathcal{A}(x))\right] \geq \frac{1}{2}\cdot\tau\cdot f(\mathcal{P}) = \frac{1}{32n} \tag{152}$$

$$\geq \Omega\left(\frac{\ln(1+d_{small}(L'))}{n} + p_{small}(L')\cdot\ln\left(1 + \frac{d_{small}(L')}{np_{small}(L')}\right)\right) \tag{153}$$

where (153) is by observing that $\frac{\ln(1+d_{small}(L'))}{n} + p_{small}(L')\ln\left(1+\frac{d_{small}(L')}{np_{small}(L')}\right) \leq \frac{2d_{small}(L')}{n} \leq \frac{6}{n}$ (due to $\ln(x) \leq x-1$ for any $x > 0$ and the condition that $d_{small}(L') \leq 3$).

3. If $d_{small}(L') \geq 4$: Without loss of generality, assume that $\{i \in L' : p_i \leq \frac{t}{n}\} = \{1, \cdots, d_{small}(L')\}$. For brevity, denote $\hat{d} = \lfloor \frac{d_{small}(L')}{2} \rfloor \geq 2$ and $\hat{p} = \sum_{i=1}^{\hat{d}} p_i$. Without

loss of generality, also assume that $p_1 \geq \cdots \geq p_{d_{small}(L')}$, then

$$\hat{d} \geq \frac{d_{small}(L')}{3} \text{ and } \hat{p} \geq \frac{p_{small}(L')}{3} \tag{154}$$

Let $\kappa, k \in \mathbb{N}$ be defined as follows.

$$\kappa = \begin{cases} 2 & \frac{\hat{d}}{n\hat{p}} \leq 2 \\ \hat{d} & \frac{\hat{d}}{n\hat{p}} > 2 \text{ and } n\hat{p} < 1 \\ \lfloor \frac{\hat{d}}{n\hat{p}} \rfloor & \frac{\hat{d}}{n\hat{p}} > 2 \text{ and } n\hat{p} \geq 1 \end{cases} \quad k = \begin{cases} \lfloor \frac{\hat{d}}{2} \rfloor & \frac{\hat{d}}{n\hat{p}} \leq 2 \\ 1 & \frac{\hat{d}}{n\hat{p}} > 2 \text{ and } n\hat{p} < 1 \\ \lfloor n\hat{p} \rfloor & \frac{\hat{d}}{n\hat{p}} > 2 \text{ and } n\hat{p} \geq 1 \end{cases}. \tag{155}$$

Thus by definition, we have

$$\kappa \geq 2 \quad \text{and} \quad k \geq 1 \quad \text{and} \quad \kappa \cdot k \leq \hat{d} \leq \frac{d_{small}(L')}{2} \quad \text{and} \quad \frac{k}{n} \leq \max\left\{\hat{p}, \frac{1}{n}\right\} \tag{156}$$

For $i = 1, \cdots, k$, we construct a packing set of distributions supported on symbols $\mathcal{B}_i = \{k \cdot i - \kappa + 1, \cdots, \delta_{\kappa \cdot i}\}$ as follows.

$$\text{For } i = 1, \cdots, k \quad \mathcal{P}_i = \{\delta_{\kappa \cdot i - \kappa + 1}, \cdots, \delta_{\kappa \cdot i}\} \tag{157}$$

We construct the following set of distributions that lie in the additive neighborhood of p.

$$\mathcal{P} = \left\{ q = \left(1 - \frac{k}{n}\right) \cdot q^c + \sum_{i=1}^{k} w_i \cdot q^i : w_i = \frac{1}{n} \text{ and } q^i \in \mathcal{P}_i \text{ for any } i = 1, \cdots, k \right\} \tag{158}$$

where

$$q_c(j) = \begin{cases} 0 & j = 1, \cdots, \hat{d} \\ \frac{1}{1-\frac{k}{n}} \cdot \left(p_j + \frac{\max\left\{\hat{p}, \frac{1}{n}\right\} - \frac{k}{n}}{d_{small}(L') - \hat{d}} \right) & j = \hat{d}+1, \cdots, d_{small}(L') \\ \frac{p_j}{1-\frac{k}{n}} \cdot \left(1 + \frac{\hat{p} - \max\left\{\hat{p}, \frac{1}{n}\right\}}{1 - p_{small}(L')} \right) & j = d_{small}(L')+1, \cdots, d \end{cases} \tag{159}$$

One can verify that the distributions in $\mathcal{P}$ are well-defined (i.e., normalized) and lie in the neighborhood $N_{\leq \frac{t}{n}}(p)$ defined in (136). This is by $p_j \leq \frac{t}{n}$ for $j = 1, \cdots, \hat{d}$, and by $0 \leq \frac{\max\left\{\hat{p}, \frac{1}{n}\right\} - \frac{k}{n}}{d_{small}(L') - \hat{d}} \leq \max\left\{\frac{\hat{p}}{\hat{d}}, \frac{1}{n}\right\} \leq \frac{t}{n}$ under (156) and $\hat{d} < d_{small}(L')$, and by $0 \geq \frac{\hat{p} - \max\left\{\hat{p}, \frac{1}{n}\right\}}{1 - p_{small}(L')} \geq -\frac{1/n}{1-3/n} \geq -1$ under (154) and $n \geq 4$, and by observing that $0 \geq p_j \cdot \frac{\hat{p} - \max\left\{\hat{p}, \frac{1}{n}\right\}}{1 - p_{small}(L')} \geq \hat{p} - \max\left\{\hat{p}, \frac{1}{n}\right\} \geq -\frac{1}{n}$ for any $j = d_{small}(L') + 1, \cdots, d$.

Now by applying Lemma A.6 to $\mathcal{P}_i$ for each $i = 1, \cdots, k$, we prove that for any $q' \in \Delta(d)$, it holds that

$$\frac{1}{|\mathcal{P}_i|} \sum_{q \in \mathcal{P}_i} \mathbf{1}_{KL(q,q') \geq f(\mathcal{P}_i)} \geq \frac{1}{2} \text{ where } f(\mathcal{P}_i) = \ln\left(1 + \frac{\kappa}{4}\right) \tag{160}$$

Thus the first condition of Theorem A.3 holds. Below we analyze the second condition of Theorem A.3. By definition of $\mathcal{P}$, for any fixed $i \in [k]$ and any fixed $\bar{p} \coloneqq \left(1 - \frac{k}{n}\right) \cdot q^c + \sum_{i=1}^{k} \frac{1}{n} \cdot \bar{q}^i \in \mathcal{P}$, we compute that

$$\int \left(\min_{q^i \in \mathcal{P}_i} \frac{d\text{Poi}\left(n, \bar{p} - \frac{1}{n}\bar{q}^i + \frac{1}{n}q^i\right)}{d\text{Poi}\left(n, \bar{p}\right)} \right) d\text{Poi}\left(n, \bar{p}\right) = \int \min_{q^i \in \mathcal{P}_i} d\text{Poi}\left(n, \bar{p} - \frac{1}{n}\bar{q}^i + \frac{1}{n}q^i\right) \tag{161}$$

$$\geq \min_{q^i \in \mathcal{P}_i} \Pr_{x \sim \text{Poi}(n, \bar{p} - \frac{1}{n}\bar{q}^i + \frac{1}{n}q^i)} \left[x_{\kappa \cdot i - \kappa + 1} = \cdots = x_{\kappa \cdot i}\right] = \frac{1}{e} \coloneqq \tau \tag{162}$$

where (162) is by probability mass function of Poisson distributions $\mathrm{Poi}(n \cdot 0)$ and $\mathrm{Poi}\left(n \cdot \frac{1}{e}\right)$. By applying Theorem A.3 under our proved conditions (160) and (162), we finally prove that

$$\max_{q \in \mathcal{P}} \mathbb{E}\left[KL(q, \mathcal{A}(x))\right] \geq \frac{1}{2} \cdot \tau \cdot \sum_{i=1}^{k} w_i \cdot f(\mathcal{P}_i) = \frac{k}{2en} \ln\left(1 + \frac{\kappa}{4}\right) \tag{163}$$

$$= \begin{cases} \frac{1}{2en} \lfloor \frac{\hat{d}}{2} \rfloor \cdot \ln\left(1 + \frac{1}{2}\right) & \frac{\hat{d}}{n\hat{p}} \leq 2 \\ \frac{1}{2en} \cdot \ln\left(1 + \frac{\hat{d}}{4}\right) & \frac{\hat{d}}{n\hat{p}} > 2 \text{ and } n\hat{p} < 1 \\ \frac{1}{2en} \cdot \lfloor n\hat{p} \rfloor \cdot \ln\left(1 + \frac{1}{4} \lfloor \frac{\hat{d}}{n\hat{p}} \rfloor\right) & \frac{\hat{d}}{n\hat{p}} > 4 \text{ and } n\hat{p} \geq 1 \end{cases} \tag{164}$$

$$\geq \begin{cases} \Omega\left(\frac{\hat{d}}{n}\right) & \frac{\hat{d}}{n\hat{p}} \leq 4 \\ \Omega\left(\frac{1}{n} \cdot \ln\left(1 + \hat{d}\right)\right) & \frac{\hat{d}}{n\hat{p}} > 4 \text{ and } n\hat{p} < 1 \\ \Omega\left(\hat{p} \cdot \ln\left(1 + \frac{\hat{d}}{n\hat{p}}\right)\right) & \frac{\hat{d}}{n\hat{p}} > 4 \text{ and } n\hat{p} \geq 1 \end{cases} \tag{165}$$

$$\geq \Omega\left(\frac{\ln(1 + \hat{d}}{n} + \hat{p} \cdot \ln\left(1 + \frac{\hat{d}}{n\hat{p}}\right)\right) \tag{166}$$

$$\geq \Omega\left(\frac{\ln(1 + d_{small}(L'))}{n} + p_{small}(L') \cdot \ln\left(1 + \frac{d_{small}(L')}{n p_{small}(L')}\right)\right) \tag{167}$$

where (164) is by using the definitions of κ and k in (155), (165) is by $\lfloor x \rfloor \geq \frac{x}{2}$ for any $x \geq 1$, (166) is by $\lambda \ln(1 + x) \geq \ln(1 + \lambda x)$ for $\lambda \geq 1$ and $x > 0$, and by $\ln(1 + \lambda x) \leq \lambda \cdot \ln(1 + x) \leq x$ for $0 \leq \lambda \leq 1$ and $x > 0$, and the last inequality is by applying (154).

□

D.3 Non-DP Per-Instance Lower Bound under $N_{stat}(p)$

We now prove the per-instance lower bound in Table 4 under neighborhood $N_{stat}(p)$ (137).

Theorem D.4. *Let $n, d \in \mathbb{N}$ and $p \in \Delta(d)$ be fixed. Let $N_{stat}(p)$ be the following additive local neighborhood of p as defined in* (137)*.*

$$N_{stat}(p) = \left\{q : |q_i - p_i| \leq \min\left\{p_i, \sqrt{\frac{p_i}{n}}\right\} \text{ for any } i \in [d]\right\} \cup \left\{q : \sum_{i=1}^{d} |q_i - p_i| \leq \frac{1}{n}\right\} \tag{168}$$

Then if $d \geq 2$, it holds that

$$\max_{q \in N_{stat}(p)} \mathop{\mathbb{E}}_{x \sim Poi(n,q)} KL(q, \mathcal{A}(x)) \geq \Omega\left(\sum_i \min\left\{p_i, \frac{1}{n}\right\}\right) \tag{169}$$

Proof. We will apply Theorem A.3 to prove the lower bound. Below we first construct the packing set. Without loss of generality, assume that $p_1 \geq \cdots \geq p_d$. For brevity, denote

$$w_k = p_{2k-1} + p_{2k} \quad \text{and} \quad \Delta_k = \frac{1}{2} \min\left\{p_{2k}, \sqrt{\frac{p_{2k}}{n}}\right\} \text{ for } k = 1, \cdots, \lfloor \frac{d}{2} \rfloor \tag{170}$$

Let $\mathcal{P}$ be the following packing set of distributions.

$$\mathcal{P} = \left\{q \coloneqq \sum_{k=1}^{\lfloor \frac{d}{2} \rfloor} w_k \cdot q^k : q_k \in \mathcal{P}_k \coloneqq \left\{q_+^k, q_-^k\right\}\right\} \tag{171}$$

where

$$q_-^k(j) = \begin{cases} \frac{1}{w_k} \cdot p_{2k-1} & j = 2k-1 \\ \frac{1}{w_k} \cdot p_{2k} & j = 2k \\ 0 & j \in [d] \setminus \{2k-1, 2k\} \end{cases} \quad \text{and} \quad q_+^k(j) = \begin{cases} \frac{1}{w_k} \cdot (p_{2k-1} + \Delta_k) & j = 2k-1 \\ \frac{1}{w_k} \cdot (p_{2k} - \Delta_k) & j = 2k \\ 0 & j \in [d] \setminus \{2k-1, 2k\} \end{cases}$$

One can verify that distributions in the packing set $\mathcal{P}$ are normalized and lie in the statistical neighborhood $N_{stat}(p)$ by observing that $\min\left\{p_{i-1}, \sqrt{\frac{p_{i-1}}{n}}\right\} \geq \min\left\{p_i, \sqrt{\frac{p_i}{n}}\right\}$ for any i. By applying Lemma A.5 (under setting $a = \frac{p_{2k-1}}{w_k}$ and $\Delta = \frac{\Delta_k}{w_k}$), we further prove that for any $q \in \Delta(d)$,

$$\frac{1}{|\mathcal{P}_k|}\sum_{q^k \in \mathcal{P}_k} \mathbf{1}_{KL(q^k,q)\geq f(\mathcal{P}_k)} \geq \frac{1}{2} \text{ where } f(\mathcal{P}_k) = \frac{\left(\frac{\Delta_k}{w_k}\right)^2}{8\cdot\frac{p_{2k}}{w_k}} = \frac{1}{32 w_k}\cdot\min\left\{p_{2k}, \frac{1}{n}\right\} \tag{172}$$

Thus the first condition of Theorem A.3 holds. We now analyze the second condition of Theorem A.3. For any k and any fixed $q^l \in \mathcal{P}_l, l \neq k$, by definition of $\mathcal{P}_k$ we compute that

$$\int \left(\min_{q^k\in\mathcal{P}_k} \frac{d\text{Poi}\left(n, \sum_{l=1}^{\lfloor \frac{d}{2}\rfloor} w_l\cdot q^l\right)}{d\text{Poi}\left(n, w_k\cdot q_+^k + \sum_{l\neq k} w_l\cdot q^l\right)}\right) d\text{Poi}\left(n, w_k\cdot q_+^k + \sum_{l\neq k} w_l\cdot q^l\right) \tag{173}$$

$$= \int \min\left\{d\text{Poi}\left(n, w_k\cdot q_+^k + \sum_{l\neq k} w_l\cdot q^l\right), d\text{Poi}\left(n, w_k\cdot q_-^k + \sum_{l\neq k} w_l\cdot q^l\right)\right\} \tag{174}$$

$$= 1 - \text{TV}\left(\text{Poi}\left(n, w_k\cdot q_+^k + \sum_{l\neq k} w_l\cdot q^l\right), \text{Poi}\left(n, w_k\cdot q_-^k + \sum_{l\neq k} w_l\cdot q^l\right)\right) \tag{175}$$

$$\geq 1 - \text{TV}\left(\text{Poi}\left(np_{2k-1}\right), \text{Poi}\left(np_{2k-1} + n\Delta_k\right)\right) - \text{TV}\left(\text{Poi}\left(np_{2k}\right), \text{Poi}\left(np_{2k} - n\Delta_k\right)\right) \tag{176}$$

$$\geq 1 - \frac{1}{2}\sqrt{\frac{n^2\Delta_k^2}{np_{2k-1}}} - \frac{1}{2}\sqrt{\frac{n^2\Delta_k^2}{np_{2k} - n\Delta_k}} \tag{177}$$

$$\geq 1 - \sqrt{\frac{\min\{np_{2k}, 1\}}{2}} \geq \frac{1}{4} \coloneqq \tau_k \tag{178}$$

where (174) is because $\mathcal{P}_k$ only contains two distributions, (175) is by using the definition of total variation distance, (176) is by the total variation inequality for product measures (Lemma D.2), (177) is by Lemma B.9, (178) is by $p_{2k-1} \geq p_{2k}$ and by definition (170) of $\Delta_k = \frac{1}{2}\min\left\{p_{2k}, \sqrt{\frac{p_{2k}}{n}}\right\} \leq \frac{p_{2k}}{2}$. Thus the second condition of Theorem A.3 holds. By applying Theorem A.3 under our proved conditions (172) and (178), we finally prove that

$$\max_{q\in\mathcal{P}} \mathbb{E}\left[KL(q, \mathcal{A}(x))\right] \geq \frac{1}{2}\sum_{k=1}^{\lfloor\frac{d}{2}\rfloor} w_k\cdot\tau_k\cdot f(\mathcal{P}_k) = \frac{1}{128}\sum_{k=1}^{\lfloor\frac{d}{2}\rfloor}\min\left\{p_{2k}, \frac{1}{n}\right\} \tag{179}$$

$$\geq \frac{1}{256}\sum_{i=2}^{d}\min\left\{p_i, \frac{1}{n}\right\} \tag{180}$$

where (180) is by observing that for $p_1 \geq \cdots \geq p_d$, it is the case that $\sum_{k=1}^{\lfloor\frac{d}{2}\rfloor}\min\left\{p_{2k}, \frac{1}{n}\right\} \geq \sum_{k=1}^{\lfloor\frac{d}{2}\rfloor}\min\left\{p_{\min\{2k+1,d\}}, \frac{1}{n}\right\}$. We now separate the discussions for the remaining (at most two) symbols.

1. If $\sum_{i=2}^{d}\min\left\{p_i, \frac{1}{n}\right\} \geq \frac{1}{2n}$, then (180) suffice to prove the bound (169) in the statement (by observing that there is only one remaining symbol).

2. If $\sum_{i=2}^{d}\min\left\{p_i, \frac{1}{n}\right\} < \frac{1}{2n}$, then it must be the case that $p_1 > 1 - \frac{1}{2n}$, $p_2 < \frac{1}{2n}$, and $\sum_{i=1}^{d}\min\left\{p_i, \frac{1}{n}\right\} < \frac{3}{2n}$. By repeating the proof for a new packing $\mathcal{P}' = \{\hat{p}^-, p\}$ where

$$\hat{p}(j) = \begin{cases} p_1 - \frac{1}{2n} & j = 1 \\ p_2 + \frac{1}{2n} & j = 2 \\ p_j & j = 3, \cdots, d \end{cases} \tag{181}$$

we similarly prove a new lower bound of $\frac{1}{256}\cdot\frac{1}{n}$. One can also validate that the new packing $\mathcal{P}'$ is also in the neighborhood N_{stat} (137) This suffice to prove (169) in the statement.

□

E Deferred Proofs for Non-DP Per-Instance Upper Bounds

E.1 Non-DP Per-Instance Upper Bound (Sampling Twice Algorithm)

We will use the following lemma for bounding the estimation error on zero-count symbols.

Lemma E.1 (Error of Algorithm 1 on zero-count Symbols)**.** *Let $p \in \Delta(d)$ be a discrete distribution over d symbols. Let $n \in \mathbb{N}$. Then Algorithm 1 satisfies*

$$\mathbb{E}\left[\sum_{i\in L} p_i \cdot \ln\left(\frac{\frac{p_i}{\sum_{i\in L} p_i}}{\frac{\tilde{x}_i}{\sum_{i\in L}\tilde{x}_i}}\right)\right] \le \mathbb{E}\left[\left(\sum_{i\in L} p_i\right)\ln\left(1+\frac{2\sum_{i\in L}1}{\sum_{i\in L} np_i}\right)\right] \tag{182}$$

Proof. By definition, we have that

$$\mathbb{E}\left[\sum_{i\in L} p_i \cdot \ln\left(\frac{\frac{p_i}{\sum_{i\in L} p_i}}{\frac{\tilde{x}_i}{\sum_{i\in L}\tilde{x}_i}}\right)\right] = \underbrace{\mathbb{E}\left[\sum_{i\in L} p_i \cdot \ln\left(\frac{np_i/2}{\tilde{x}_i}\right)\right]}_{①} + \underbrace{\mathbb{E}\left[\left(\sum_{i\in L} p_i\right)\ln\left(\frac{\sum_{i\in L}\tilde{x}_i}{\sum_{i\in L} np_i/2}\right)\right]}_{②} \tag{183}$$

We first analyze ①. By concavity of $\ln(t)$ over $t>0$ and by $\tilde{x}_i \le x'_i + 1$, we have that

$$① \le \mathbb{E}_L\left[\sum_{i\in L} p_i \cdot \ln\left(\frac{np_i}{2}\cdot\mathbb{E}\left[\frac{1}{x'_i+1}\right]\right)\right] \le 0 \tag{184}$$

where the last inequality is by Lemma B.1 under $m=\frac{n}{2}$. We then analyze ②. By concavity of $\ln(t)$ over $t>0$, we have that

$$② \le \mathbb{E}_L\left[\sum_{i\in L} p_i \ln\left(1+\frac{\sum_{i\in L}\mathbb{E}[\tilde{x}_i - np_i/2]}{\sum_{i\in L} np_i/2}\right)\right] \tag{185}$$

$$\le \mathbb{E}_L\left[\left(\sum_{i\in L} p_i\right)\ln\left(1+\frac{2\sum_{i\in L}1}{\sum_{i\in L} np_i}\right)\right] \tag{186}$$

where the last inequality is by applying Lemma B.3 with $b=0$ and $c=1$. □

We are now ready to prove the Non-DP per-instance upper bound for the non-dp "sampling twice" algorithm Algorithm 1.

Theorem E.2 (Per-Instance Upper Bound - Sampling Twice Algorithm)**.** *Let $\mathcal{A}$ be the estimator given by Algorithm 1. If $d \ge 2$, then for any n and any $p \in \Delta(d)$,*

$$\mathop{\mathbb{E}}_{x\sim Poi(n,p)}\Big[KL(p\|\mathcal{A}(x))\Big] \le O\left(\mathbb{E}\left[\sum_{i\in L} p_i\cdot\ln\left(1+\frac{\sum_{j\in L}1}{n\sum_{j\in L}p_j}\right)\right] + \sum_i \min\left\{p_i, \frac{1}{n}\right\}\right) \tag{187}$$

Proof. By definition,

$$\mathop{\mathbb{E}}_{x,x'\sim\text{Poi}(np)}[KL(p\|\mathcal{A}(x))] \tag{188}$$

$$=\mathbb{E}\left[\sum_{i\in L} p_i \ln\left(\frac{p_i/\sum_{j\in L} p_j}{\tilde{x}_i/\sum_{j\in L}\tilde{x}_j}\right)\right]+\mathbb{E}\left[\left(\sum_{i\in L} p_i\right)\ln\left(\frac{\sum_{i\in L} p_i}{\frac{\tilde{c}}{n/2}}\right)+\sum_{i\notin L} p_i\ln\left(\frac{p_i}{\frac{\tilde{x}_i}{n/2}}\right)+\ln\left(\frac{\tilde{c}+\sum_{i\notin L}\tilde{x}_i}{n/2}\right)\right] \tag{189}$$

$$\leq\mathbb{E}\left[\sum_{i\in L} p_i\cdot\ln\left(1+\frac{2\sum_{j\in L}1}{\sum_{j\in L} np_j}\right)\right]+\mathbb{E}\left[\left(\sum_{i\in L} p_i\right)\ln\left(\frac{\sum_{i\in L} p_i}{\frac{\tilde{c}}{n/2}}\right)+\frac{\tilde{c}-\sum_{i\in L} np_i/2}{n/2}\right]$$

$$+\mathbb{E}\left[\sum_{i\notin L} p_i\ln\left(\frac{p_i}{\frac{\tilde{x}_i}{n/2}}\right)+\sum_{i\notin L}\frac{\tilde{x}_i-np_i/2}{n/2}\right] \tag{190}$$

$$\leq O\left(\mathbb{E}\left[\sum_{i\in L} p_i\cdot\ln\left(1+\frac{\sum_{j\in L}1}{n\sum_{j\in L} p_j}\right)\right]\right)+\frac{1}{n/2}+\mathbb{E}\left[\sum_{i\notin L}\frac{1}{n/2}\right] \tag{191}$$

$$\leq O\left(\mathbb{E}\left[\sum_{i\in L} p_i\cdot\ln\left(1+\frac{\sum_{j\in L}1}{n\sum_{j\in L} p_j}\right)\right]+\sum_i\min\left\{p_i,\frac{1}{n}\right\}\right) \tag{192}$$

where (190) is by $\ln(t)\leq t-1$ for any $t>0$ and by applying Lemma E.1, (191) is by applying Lemma B.10 under $m=n/2$, $c_1=-\infty$, $c_2=+\infty$, $b=0$ and $c=1$, and the last inequality is by $\Pr[i\notin L]=1-e^{-np_i}\leq\min\{np_i,1\}$, and by $\sum_i\min\{p_i,\frac{1}{n}\}\geq\frac{1}{n}$ for $d\geq 2$. □

E.2 Proof for Matching Lower and Upper Bound

Corollary E.3. *Let $\mathcal{A}$ be the estimator given by Algorithm 1. Let $N_{stat}(p)$ and $N_{\leq\frac{t}{n}}(p)$ be the additive neighborhoods defined in* (137) *and* (136) *respectively. Then for any n and any $p\in\Delta(d)$,*

$$\mathop{\mathbb{E}}_{x\sim Poi(n,p)}\Big[KL(p\|\mathcal{A}(x))\Big]\leq O\left(\textit{lower}(p,n,N_{stat})+\textit{lower}(p,n,N_{\leq\frac{t}{n}})\right) \tag{193}$$

for any choice of neighborhood size $t\geq 1$ such that $t\cdot e^{-t}\leq\frac{1}{\ln d}$. Specifically, we can always choose $t=\min\{1,2\ln\ln d\}$.

Proof. We will use the upper bound given by Theorem E.2. Observe that by $\frac{\sum_{j:x_j=0}1}{\sum_{j:x_j=0}p_j}\leq\frac{\sum_{j:p_j\leq\frac{t}{n},x_j=0}1}{\sum_{j:p_j\leq\frac{t}{n},x_j=0}p_j}$, we have

$$\mathbb{E}\left[\sum_{i\in L:p_i\leq\frac{t}{n}} p_i\ln\left(1+\frac{\sum_{j\in L}1}{n\sum_{j\in L}p_j}\right)\right]\leq\mathbb{E}\left[\sum_{i\in L:p_i\leq\frac{t}{n}} p_i\ln\left(1+\frac{\sum_{j\in L:p_j\leq\frac{t}{n}}1}{n\sum_{j\in L:p_j\leq\frac{t}{n}}p_j}\right)\right] \tag{194}$$

$$\leq\text{lower}(p,n,N_{\leq\frac{t}{n}})\text{ in Theorem D.3} \tag{195}$$

Additionally, by definition, we compute that

$$\mathbb{E}\left[\sum_{i\in L:p_i>\frac{t}{n}} p_i\ln\left(1+\frac{\sum_{j\in L}1}{n\sum_{j\in L}p_j}\right)\right]\leq\mathbb{E}\left[\sum_{i\in L:p_i>\frac{t}{n},x_i=0} p_i\ln\left(1+\frac{d}{t}\right)\right] \tag{196}$$

$$=\sum_{i:p_i>\frac{t}{n}} p_i e^{-np_i}\ln\left(1+\frac{d}{t}\right) \tag{197}$$

$$\leq O\left(\sum_{i:p_i>\frac{t}{n}}\frac{1}{n}\right)=\text{lower}(p,n,N_{stat})\text{ in Theorem F.2} \tag{198}$$

Neighborhood N	$\text{lower}_{\varepsilon,\delta}(p,n,N)$ (3)	Conditions	Reference
$N_{\le \frac{1}{n\varepsilon}}$ (6)	$\Omega\left(\sum_{i:p_i<\frac{1}{n\varepsilon}} p_i + \sum_{i:p_i\ge\frac{1}{n\varepsilon}} \frac{1}{p_i}\cdot\frac{1}{n^2\varepsilon^2}\right)$	$\delta\le\varepsilon, d\ge 2$	Theorem F.1
$N_{\le \frac{t}{n\varepsilon}}$ (6)	$\Omega\Bigg(\frac{\ln\left(1+d_{small}(L')\right)}{n\varepsilon}+ p_{small}(L')\cdot\ln\left(1+\frac{d_{small}(L')}{n\varepsilon\cdot p_{small}(L')}\right)\Bigg)$	$L'\subseteq[d], t\ge 1, d\ge 2$; $\delta\le\varepsilon, n\varepsilon\ge 1$	Theorem F.2

Table 5: DP per-instance lower bounds, where $p_{small}(L') = \sum_{i\in L':p_i\le\frac{t}{n\varepsilon}} p_i$ and $d_{small}(L') = \sum_{i\in L':p_i\le\frac{t}{n\varepsilon}} 1$

where (197) is by $L=\{i\in[d]: x_i=0\}$ and by probability mass function of Poisson random variables, and (198) is by choosing $t\ge 1$ such that $t\cdot e^{-t}\le\frac{1}{\ln d}$. Specifically, we can choose $t=\min\{1, 2\ln\ln d\}$. □

F Deferred Proofs for DP Per-Instance Lower Bound Theorem 3.1

For ease of presentation and understanding, we break up the neighborhood $N_{\le\frac{t}{n\varepsilon}}(p)$ for a given small $t\ge 1$ into two sub-neighborhoods: one $N_{\le\frac{1}{n}}\subseteq N_{\le\frac{t}{n\varepsilon}}(p)$ with the same small perturbation for every symbol, and the other $N_{\le\frac{t}{n\varepsilon}}(p)$ with larger perturbations for small symbols, defined as follows.

We then prove per-instance lower bounds under the sub-neighborhoods $N_{\le\frac{t}{n\varepsilon}}(p)$ and $N_{\le\frac{t}{n\varepsilon}}(p)$ respectively. By definition (2), their average is a lower bound for the per-instance lower bound under $N_{\le\frac{t}{n\varepsilon}}(p)$, i.e., $\text{lower}(p,n,N_{\le\frac{t}{n\varepsilon}})\ge\frac{1}{2}\cdot\text{lower}(p,n,N_{\le\frac{1}{n\varepsilon}})+\frac{1}{2}\cdot\text{lower}(p,n,N_{\le\frac{t}{n\varepsilon}})$. (This is because one can construct a distribution over hard instances, choosing the hard instance(s) in $N_{\le\frac{t}{n}}(p)$ and $N_{stat}(p)$ with $1/2$ probability respectively.)

Our results for DP per-instance lower bounds under neighborhoods $N_{\le\frac{t}{n\varepsilon}}, t\ge 1$ are summarized in Table 5.

F.1 DP Lower Bound: $N_{\frac{1}{n\varepsilon}}$ neighborhood

In this section, we prove the per-instance DP estimation lower bound in Table 5 under additive neighborhood. We first prove a lower bound under add-$\frac{1}{n\varepsilon}$ neighborhood.

Theorem F.1 (Lower Bound - $N_{\le\frac{1}{n\varepsilon}}$ Neighborhood)**.** *Let $d\in\mathbb{N}$, $\varepsilon\ge 0$ and $0\le\delta\le 1$, $p\in\Delta(d)$ be fixed. Let $N_{\le\frac{1}{n\varepsilon}}(p)$ be the additive neighborhood defined in* (6) *for $t=1$ as follows.*

$$N_{\le\frac{1}{n\varepsilon}}(p)=\left\{q: |q_i-p_i|\le\frac{1}{n\varepsilon}\text{ for any } i\in[d] \text{ and } \sum_{i:p_i\le\frac{1}{n\varepsilon}} q_i\le\max\left\{\frac{1}{n\varepsilon}, \sum_{i:p_i\le\frac{1}{n\varepsilon}} p_i\right\}\right\} \tag{199}$$

If $\delta\le\varepsilon$ and $d\ge 2$, then the expected KL error of any (ε,δ)-DP estimator $\mathcal{A}$ satisfies

$$\max_{q\in N_{\le\frac{1}{n\varepsilon}}(p)}\mathop{\mathbb{E}}_{x\sim Poi(n,q)}[KL(q,\mathcal{A}(x))]\ge\Omega\left(\sum_i\min\left\{p_i,\frac{1}{p_i}\cdot\frac{1}{n^2\varepsilon^2}\right\}\right) \tag{200}$$

Proof. We will apply Theorem A.4 to prove the lower bound. Below we first construct the packing set. Without loss of generality, assume that that $d \mod 2 = 0$. (Otherwise, sort the symbols to satisfy $\min\left\{p_1,\frac{1}{p_1}\cdot\frac{1}{n^2\varepsilon^2}\right\}\ge\cdots\ge\min\left\{p_d,\frac{1}{p_d}\cdot\frac{1}{n^2\varepsilon^2}\right\}$ and ignore the last symbol in the below constructions.) For brevity, denote

$$w_k=p_{2k-1}+p_{2k}\quad\text{and}\quad\Delta_k=\frac{1}{160}\min\left\{p_{2k},\frac{1}{n\varepsilon}\right\}\text{ for } k=1,\cdots,\frac{d}{2} \tag{201}$$

Let $\mathcal{P}$ be the following packing set of distributions.

$$\mathcal{P} = \left\{ q \coloneqq \sum_{k=1}^{\frac{d}{2}} w_k \cdot q^k : q_k \in \mathcal{P}_k \coloneqq \left\{ q_+^k, q_-^k \right\} \right\} \tag{202}$$

where

$$q_-^k(j) = \begin{cases} \frac{1}{w_k} \cdot p_{2k-1} & j = 2k-1 \\ \frac{1}{w_k} \cdot p_{2k} & j = 2k \\ 0 & j \in [d] \setminus \{2k-1, 2k\} \end{cases} \quad \text{and} \quad q_+^k(j) = \begin{cases} \frac{1}{w_k} \cdot (p_{2k-1} + \Delta_k) & j = 2k-1 \\ \frac{1}{w_k} \cdot (p_{2k} - \Delta_k) & j = 2k \\ 0 & j \in [d] \setminus \{2k-1, 2k\} \end{cases}$$

One can verify that distributions in the packing set $\mathcal{P}$ are normalized and lie in the additive neighborhood $N_{\leq \frac{1}{n\varepsilon}}(p)$ by observing that $\frac{1}{n\varepsilon} \geq \min\left\{p_{i-1}, \frac{1}{n\varepsilon}\right\} \geq \min\left\{p_i, \frac{1}{n\varepsilon}\right\}$ for any i. By applying Lemma A.5 (under setting $a = \frac{p_{2k-1}}{w_k}$ and $\Delta = \frac{\Delta_k}{w_k}$), we further prove that for any $q \in \Delta(d)$,

$$\frac{1}{|\mathcal{P}_k|} \sum_{q^k \in \mathcal{P}_k} \mathbf{1}_{KL(q^k, q) \geq f(\mathcal{P}_k)} \geq \frac{1}{2} \text{ where } f(\mathcal{P}_k) = \frac{\left(\frac{\Delta_k}{w_k}\right)^2}{8 \cdot \frac{p_{2k}}{w_k}} = \frac{1}{8 \cdot 160^2 \cdot w_k} \cdot \min\left\{p_{2k}, \frac{1}{p_{2k} \cdot n^2\varepsilon^2}\right\} \tag{203}$$

Thus the first condition of Theorem A.4 holds. We now analyze the second condition of Theorem A.4. For any k and any fixed $q^l \in \mathcal{P}_l, l = 1, \cdots, \frac{d}{2}$ and fixed $\bar{q}_k \in \mathcal{P}_k$. Let $x \sim \text{Poi}\left(n, \sum_{l=1}^{\frac{d}{2}} w_l \cdot q^l\right)$ and $y \sim \text{Poi}\left(n \cdot \Delta_k\right)$ and $y' \sim \text{Poi}\left(n \cdot \Delta_k\right)$ be independent Poisson random variables. Then we could construct the following coupling $(x, \bar{x})$ between distributions $\text{Poi}\left(n, \sum_{l=1}^{\frac{d}{2}} w_l \cdot q^l\right)$ and $\text{Poi}\left(n, w_k \cdot q^k + \sum_{l \neq k} w_l \cdot q^l\right)$:

$$\bar{x}_l = \begin{cases} x_l + y & l = 2k-1 \\ x_l - y' & l = 2k \\ x_l & l \in [d] \setminus \{2k-1, 2k\} \end{cases} \tag{204}$$

By definition, we compute that

$$\mathbb{E}\left[\|x - \bar{x}\|_1\right] = \mathbb{E}\left[y + y'\right] \leq \frac{1}{80\varepsilon} \coloneqq \tau \tag{205}$$

where the inequality is by definition $\Delta_k \leq \frac{1}{160n\varepsilon}$. Thus the second condition of Theorem A.4 holds. By applying Theorem A.4 under our proved conditions (203) and (205), we prove that

$$\max_{q \in \mathcal{P}} \mathbb{E}\left[KL(q, \mathcal{A}(x))\right] \geq \sum_{k=1}^{\frac{d}{2}} w_k \cdot \left(\frac{1}{10} - 4\varepsilon \cdot \tau\right) \cdot f(\mathcal{P}_k) = \frac{1}{20} \cdot \frac{1}{8 \cdot 160^2} \cdot \sum_{k=1}^{\frac{d}{2}} \min\left\{p_{2k}, \frac{1}{p_{2k} \cdot n^2\varepsilon^2}\right\} \tag{206}$$

By repeating the constructions under $\Delta_k = -\frac{1}{160} \min\left\{p_{2k-1}, \frac{1}{n\varepsilon}\right\}$ for $k = 1, \cdots, \frac{d}{2}$, we similarly prove that

$$\max_{q \in \mathcal{P}} \mathbb{E}\left[KL(q, \mathcal{A}(x))\right] \geq \frac{1}{20} \cdot \frac{1}{8 \cdot 160^2} \cdot \sum_{k=1}^{\frac{d}{2}} \min\left\{p_{2k-1}, \frac{1}{p_{2k-1} \cdot n^2\varepsilon^2}\right\} \tag{207}$$

By combining (206) and (207), we prove the bound (200) in the statement. □

F.2 DP Lower Bound: $N_{\frac{t}{n\varepsilon}}$ neighborhood

In this section, we prove the per-instance DP estimation lower bound in Table 5 for low probability symbols.

Theorem F.2 (Lower Bound - low probability symbols)**.** *Let $p \in \Delta(d)$ be fixed. For any $t > 0$, let $N_{\frac{t}{n\varepsilon}}$ be the additive neighborhood defined as follows.*

$$N_{\leq \frac{t}{n\varepsilon}}(p,n) = \left\{ q : |q_i - p_i| \leq \frac{t}{n\varepsilon} \text{ for any } i \in [d] \text{ and } \sum_{i:p_i \leq \frac{t}{n\varepsilon}} q_i \leq \max\left\{ \frac{t}{n\varepsilon}, \sum_{i:p_i \leq \frac{t}{n\varepsilon}} p_i \right\} \right\} \tag{208}$$

Then if $\delta \leq \varepsilon$, $n\varepsilon \geq 1$, $t \geq 1$ and $d \geq 2$, then for any $L' \subseteq [d]$, the expected KL error of any (ε, δ)-DP estimator $\mathcal{A}$ satisfies

$$\max_{q \in N_{\leq \frac{t}{n\varepsilon}}(p,n)} \mathop{\mathbb{E}}_{x \sim Poi(n,q)} KL(q, \mathcal{A}(x)) \geq \Omega\left(\frac{\ln\left(1 + d_{small}(L')\right)}{n\varepsilon} + p_{small}(L') \cdot \ln\left(1 + \frac{d_{small}(L')}{n\varepsilon \cdot p_{small}(L')}\right)\right) \tag{209}$$

where $p_{small}(L') = \sum_{i \in L', p_i \leq \frac{t}{n\varepsilon}} p_i$ and $d_{small}(L') = \sum_{i \in L', p_i \leq \frac{t}{n\varepsilon}} 1$.

Proof. 1. If $d_{small}(L') = 0$, then $p_{small}(L') = 0$ and thus (209) trivially holds.

2. If $1 \leq d_{small}(L') \leq 3$: Without loss of generality, assume that $\{i \in L' : p_i \leq \frac{t}{n\varepsilon}\} = \{1, \cdots, d_{small}(L')\}$. We construct a packing set of distributions $\mathcal{P} = \{p^+, p^-\}$ that contains the following two distributions.

$$p^+(j) = \begin{cases} \frac{1}{80n\varepsilon} & j = 1 \\ \left(1 - \frac{1}{80n\varepsilon}\right) \cdot \left(p_j + \frac{p_1}{d-1}\right) & j = 2, \cdots, d \end{cases} \tag{210}$$

and

$$p^-(j) = \begin{cases} \frac{1}{160n\varepsilon} & j = 1 \\ \left(1 - \frac{1}{80n\varepsilon}\right) \cdot \left(p_j + \frac{p_1}{d-1}\right) & j = 2, \cdots, d-1 \\ \left(1 - \frac{1}{80n\varepsilon}\right) \cdot \left(p_d + \frac{p_1}{d-1}\right) + \frac{1}{160n\varepsilon} & j = d \end{cases} \tag{211}$$

One can validate that the distributions in $\mathcal{P}$ are well-defined and are in the additive neighborhood $N_{\leq \frac{t}{n\varepsilon}}(p)$ (by observing that $0 \leq p_1 \leq \frac{t}{n\varepsilon}$ for $d \geq 2$ and thus $\frac{p_1}{d-1} \leq \frac{t}{n\varepsilon}$). By applying Lemma A.5 (under setting $a = \frac{1}{80n\varepsilon}$ and $\Delta = \frac{1}{160n\varepsilon}$), we further prove that for any $q' \in \Delta(d)$,

$$\frac{1}{|\mathcal{P}|} \sum_{q \in \mathcal{P}} \mathbf{1}_{KL(q,q') \geq f(\mathcal{P})} \geq \frac{1}{2} \text{ where } f(\mathcal{P}) = \frac{\Delta^2}{a} = \frac{1}{320n\varepsilon} \tag{212}$$

Thus the first condition of Theorem A.4 holds. Below we analyze the second condition of Theorem A.4. Let $\bar{q} = p^+$. For any $q \in \mathcal{P}$, let $y \sim \text{Poi}(n \min\{p^+, p^-\})$ and let $y' \sim \text{Poi}(n \max\{p^+, p^-\} - n \min\{p^+, p^-\})$. Then we could construct the following coupling $(x, \bar{x})$ between distributions Poi (n, q) and Poi $(n, \bar{q})$:

$$x_l = \begin{cases} y_1 + y_1' & l = 1 \\ y_l & l = 2, \cdots, d \end{cases} \text{ and } \bar{x}_l = \begin{cases} y_l & l = 1, \cdots, d-1 \\ y_l + y_l' & l = d \end{cases} \tag{213}$$

By definition, we compute that

$$\mathbb{E}\left[\|x - \bar{x}\|_1\right] = \mathbb{E}\left[y_1' + y_d'\right] = \frac{1}{80\varepsilon} := \tau \tag{214}$$

Thus the second condition of Theorem A.4 holds. By applying Theorem A.4 under our proved conditions (212) and (151), we finally prove that

$$\max_{q \in \mathcal{P}} \mathbb{E}\left[KL(q, \mathcal{A}(x))\right] \geq 1 \cdot \left(\frac{1}{10} - 4\varepsilon \cdot \tau\right) \cdot f(\mathcal{P}) = \frac{1}{20 \cdot 320n\varepsilon} \tag{215}$$

$$\geq \Omega\left(\frac{\ln\left(1 + d_{small}(L')\right)}{n\varepsilon} + p_{small}(L') \cdot \ln\left(1 + \frac{d_{small}(L')}{p_{small}(L') \cdot n\varepsilon}\right)\right) \tag{216}$$

where (216) is by the assumption that $d_{small}(L') \leq 3$ and by $\ln(1 + x) \leq x$ for $x > 0$.

3. If $d_{small}(L') \geq 4$: Without loss of generality, assume that $\{i \in L' : p_i \leq \frac{t}{n\varepsilon}\}$ consists of symbols $1, \cdots, d_{small}(L')$. For brevity, denote $\hat{d} = \lfloor \frac{d_{small}(L')}{2} \rfloor \geq 2$ and $\hat{p} = \sum_{i=1}^{\hat{d}} p_i$. Without loss of generality, assume that $p_1 \geq \cdots \geq p_{d_{small}(L')}$, then it follows that

$$\hat{d} \geq \frac{d_{small}(L')}{3} \text{ and } \hat{p} \geq \frac{p_{small}(L')}{3} \tag{217}$$

Let $\kappa, k \in \mathbb{N}$ be defined as follows.

$$\kappa = \begin{cases} 2 & \frac{\hat{d}}{160n\varepsilon\hat{p}} \leq 2 \\ \hat{d} & \frac{\hat{d}}{160n\varepsilon\hat{p}} > 2 \text{ and } 160n\varepsilon\hat{p} < 1 \\ \lfloor \frac{\hat{d}}{160n\varepsilon\hat{p}} \rfloor & \frac{\hat{d}}{160n\varepsilon\hat{p}} > 2 \text{ and } 160n\varepsilon\hat{p} \geq 1 \end{cases} \quad k = \begin{cases} \lfloor \frac{\hat{d}}{2} \rfloor & \frac{\hat{d}}{160n\varepsilon\hat{p}} \leq 2 \\ 1 & \frac{\hat{d}}{160n\varepsilon\hat{p}} > 2 \text{ and } 160n\varepsilon\hat{p} < 1 \\ \lfloor 160n\varepsilon\hat{p} \rfloor & \frac{\hat{d}}{160n\varepsilon\hat{p}} > 2 \text{ and } 160n\varepsilon\hat{p} \geq 1 \end{cases}. \tag{218}$$

Thus by definition,

$$\kappa \geq 2 \quad \text{and} \quad k \geq 1 \quad \text{and} \quad \kappa \cdot k \leq \hat{d} \leq \frac{d_{small}(L')}{2} \quad \text{and} \quad \frac{k}{160n\varepsilon} \leq \max\left\{\hat{p}, \frac{1}{160n\varepsilon}\right\} \tag{219}$$

For $i = 1, \cdots, k$, we construct a packing set of distributions supported on symbols $\mathcal{B}_i = \{k \cdot i - \kappa + 1, \cdots, \delta_{\kappa \cdot i}\}$ as follows.

$$\text{For } i = 1, \cdots, k \quad \mathcal{P}_i = \{\delta_{\kappa \cdot i - \kappa + 1}, \cdots, \delta_{\kappa \cdot i}\} \tag{220}$$

We construct the following set of distributions that lie in the additive neighborhood of p.

$$\mathcal{P} = \left\{ q = \left(1 - \frac{k}{160n\varepsilon}\right) \cdot q^c + \sum_{i=1}^{k} w_i \cdot q^i : w_i = \frac{1}{160n\varepsilon} \text{ and } q^i \in \mathcal{P}_i \text{ for any } i = 1, \cdots, k \right\} \tag{221}$$

where

$$q_c(j) = \begin{cases} 0 & j = 1, \cdots, \hat{d} \\ \frac{1}{1 - \frac{k}{160n\varepsilon}} \cdot \left(p_j + \frac{\max\left\{\hat{p}, \frac{1}{160n\varepsilon}\right\} - \frac{k}{160n\varepsilon}}{d_{small}(L') - \hat{d}} \right) & j = \hat{d} + 1, \cdots, d_{small}(L') \\ \frac{p_j}{1 - \frac{k}{160n\varepsilon}} \cdot \left(1 + \frac{\hat{p} - \max\left\{\hat{p}, \frac{1}{160n\varepsilon}\right\}}{1 - p_{small}(L')} \right) & j = d_{small}(L') + 1, \cdots, d \end{cases} \tag{222}$$

One can verify that the distributions in $\mathcal{P}$ are well-defined (i.e., normalized) and lie in the neighborhood $N_{\leq \frac{t}{n\varepsilon}}(p)$ defined in (6). This is by observing that $p_j \leq \frac{t}{n\varepsilon}$ for $j = 1, \cdots, \hat{d}$, and that $0 \leq \frac{\max\left\{\hat{p}, \frac{1}{160n\varepsilon}\right\} - \frac{k}{160 \cdot n\varepsilon}}{d_{small}(L') - \hat{d}} \leq \max\left\{\frac{\hat{p}}{\hat{d}}, \frac{1}{160n\varepsilon}\right\}$ due to (219) and $\hat{d} < d_{small}(L')$, and by $0 \geq \frac{\hat{p} - \max\left\{\hat{p}, \frac{1}{160n\varepsilon}\right\}}{1 - p_{small}(L')} \geq -\frac{\frac{1}{160n\varepsilon}}{1 - \frac{3}{160n\varepsilon}} > -1$ under (217) and $n\varepsilon \geq 1$, and by $0 \geq p_j \cdot \frac{\hat{p} - \max\left\{\hat{p}, \frac{1}{160n\varepsilon}\right\}}{1 - p_{small}(L')} \geq -\frac{1}{160n\varepsilon}$.

Now by applying Lemma A.6 to $\mathcal{P}_i$ for each $i = 1, \cdots, k$, we prove that for any $q' \in \Delta(d)$,

$$\frac{1}{|\mathcal{P}_i|} \sum_{q \in \mathcal{P}_i} \mathbf{1}_{KL(q,q') \geq f(\mathcal{P}_i)} \geq \frac{1}{2} \text{ where } f(\mathcal{P}_i) = \ln\left(1 + \frac{\kappa}{4}\right) \tag{223}$$

Thus the first condition of Theorem A.4 holds. Below we analyze the second condition of Theorem A.4. For each $i \in [k]$ and fixed $q^j \in \mathcal{P}_j, j = 1, \cdots, k$ and fixed $\bar{q}^i \in \mathcal{P}_i$. Let $x \sim \text{Poi}\left(n, \left(1 - \frac{k}{160n\varepsilon}\right) \cdot q^c + \sum_{j=1}^{k} w_j \cdot q^j\right)$ and $y \sim \text{Poi}\left(n, \left(1 - \frac{k}{160n\varepsilon}\right) \cdot q^c + w_i \cdot \bar{q}^i + \sum_{j \neq i} w_j \cdot q^j\right)$ be independent Poisson

random variables. Then we could construct the following coupling $(x, \bar{x})$ between distribution $\text{Poi}\left(n, \left(1 - \frac{k}{160n\varepsilon}\right) \cdot q^c + \sum_{j=1}^{k} w_j \cdot q^j\right)$ and distribution $\text{Poi}\left(n, \left(1 - \frac{k}{160n\varepsilon}\right) \cdot q^c + w_i \cdot \bar{q}^i + \sum_{j \neq i} w_j \cdot q^j\right)$:

$$\bar{x}_l = \begin{cases} y_l & l = \kappa \cdot i - \kappa + 1, \cdots, \kappa \cdot i \\ x_l & l \in [d] \setminus \{\kappa \cdot i - \kappa + 1, \cdots, \kappa \cdot i\} \end{cases} \tag{224}$$

By definition, we compute that

$$\mathbb{E}\left[\|x - \bar{x}\|_1\right] = \sum_{l=\kappa \cdot i - \kappa + 1}^{\kappa \cdot i} \mathbb{E}\left[|x_l - y_l|\right] \leq \sum_{l=\kappa \cdot i - \kappa + 1}^{\kappa \cdot i} \mathbb{E}\left[x_l + y_l\right] = \frac{1}{80\varepsilon} \coloneqq \tau \tag{225}$$

where the inequality is by triangle inequality for ℓ_1 distance. By applying Theorem A.4 under our proved conditions (223) and (225), we finally prove that

$$\max_{q \in \mathcal{P}} \mathbb{E}\left[KL(q, \mathcal{A}(x))\right] \geq \sum_{i=1}^{k} w_i \cdot \left(\frac{1}{10} - 4\varepsilon \cdot \tau\right) \cdot f(\mathcal{P}_i) = \frac{k}{160 \cdot n\varepsilon} \cdot \frac{1}{20} \cdot \ln\left(1 + \frac{\kappa}{4}\right) \tag{226}$$

$$= \begin{cases} \frac{1}{160 \cdot n\varepsilon} \lfloor \frac{\hat{d}}{2} \rfloor \cdot \ln\left(1 + \frac{1}{2}\right) & \frac{\hat{d}}{160n\varepsilon\hat{p}} \leq 2 \\ \frac{1}{160 \cdot n\varepsilon} \cdot \ln\left(1 + \frac{1}{4}\hat{d}\right) & \frac{\hat{d}}{160n\varepsilon\hat{p}} > 2 \text{ and } 160n\varepsilon\hat{p} < 1 \\ \frac{1}{160 \cdot n\varepsilon} \cdot \lfloor 160n\varepsilon\hat{p} \rfloor \cdot \ln\left(1 + \frac{1}{4}\lfloor \frac{\hat{d}}{160n\varepsilon\hat{p}} \rfloor\right) & \frac{\hat{d}}{160n\varepsilon\hat{p}} > 2 \text{ and } 160n\varepsilon\hat{p} \geq 1 \end{cases} \tag{227}$$

$$\geq \begin{cases} \Omega\left(\frac{\hat{d}}{n\varepsilon}\right) & \frac{\hat{d}}{160n\varepsilon\hat{p}} \leq 2 \\ \Omega\left(\frac{1}{n\varepsilon} \cdot \ln\left(1 + \hat{d}\right)\right) & \frac{\hat{d}}{160n\varepsilon\hat{p}} > 2 \text{ and } 160n\varepsilon\hat{p} < 1 \\ \Omega\left(\hat{p} \cdot \ln\left(1 + \frac{\hat{d}}{n\varepsilon\hat{p}}\right)\right) & \frac{\hat{d}}{160n\varepsilon\hat{p}} > 2 \text{ and } 160n\varepsilon\hat{p} \geq 1 \end{cases} \tag{228}$$

$$\geq \Omega\left(\frac{\ln(1 + \hat{d})}{n\varepsilon} + \hat{p} \cdot \ln\left(1 + \frac{\hat{d}}{n\varepsilon \cdot \hat{p}}\right)\right) \tag{229}$$

$$\geq \Omega\left(\frac{\ln(1 + d_{small}(L'))}{n\varepsilon} + p_{small}(L') \cdot \ln\left(1 + \frac{d_{small}(L')}{n\varepsilon \cdot p_{small}(L')}\right)\right) \tag{230}$$

where (227) is by using the definitions of κ and k in (218), (228) is by $\lfloor x \rfloor \geq \frac{x}{2}$ for any $x \geq 1$, (229) is by $\lambda \ln(1 + x) \geq \ln(1 + \lambda x)$ for $\lambda \geq 1$ and $x > 0$, and by $\ln(1 + \lambda x) \leq \lambda \cdot \ln(1 + x) \leq x$ for $0 \leq \lambda \leq 1$ and $x > 0$, and the last inequality is by (217).

□

G Deferred Proofs for DP Per-Instance Upper Bounds

G.1 Upper Bound: DP Sampling Twice Algorithm

For convenience, we first prove the following lemma about the probability of small symbols going above a threshold, and the probability of large symbols going below a threshold.

Lemma G.1 (Probability of False Positive)**.** *Let $m \in \mathbb{N}$, $p > 0$, $\tau > 0$ and $\varepsilon > 0$. Let $x \sim$ Poi(mp) and $z \sim$ Lap $\left(\frac{1}{\varepsilon}\right)$. If $p \leq \frac{1}{m \min\{\varepsilon, 1\}}$, then*

$$\Pr\left[x + z \geq \frac{\tau}{\min\{\varepsilon, 1\}}\right] \leq O\left(e^{-\frac{\tau}{2}}\right) \tag{231}$$

Proof. By applying Lemma B.2 under setting $a = mp_i$, $b = \frac{1}{\varepsilon}$ and $c = \frac{\tau}{\min\{\varepsilon,1\}}$, we prove that

$$\Pr\left[x + z \geq \frac{\tau}{\min\{\varepsilon,1\}}\right] \leq \frac{4}{3} e^{\frac{mp_i \min\{\varepsilon,1\}}{2} - \frac{\tau}{2}} \leq O\left(e^{-\frac{\tau}{2}}\right) \tag{232}$$

where the last inequality is by $p \leq \frac{1}{m \min\{\varepsilon,1\}}$.

□

Lemma G.2 (Probability of False Negative)**.** *Let $m \in \mathbb{N}$, $p > 0$, $\tau > 0$, and $\varepsilon > 0$. Let $x \sim$ Poi(mp) and $z \sim$ Lap $\left(\frac{1}{\varepsilon}\right)$. If $p \geq \frac{t}{m \min\{\varepsilon,1\}}$ for $t = 6\tau \geq 1$, then for any constant $\gamma \geq 0$,*

$$\Pr\left[x + z \leq \frac{\tau}{\min\{\varepsilon,1\}}\right] \leq O\left(e^{-\frac{\tau}{2}} \cdot \frac{1}{m^\gamma \min\{1,\varepsilon\}^\gamma p_i^\gamma}\right) \tag{233}$$

Proof. By applying Lemma B.2 under setting $a = mp_i$, $b = \frac{1}{\varepsilon}$ and $c = \frac{\tau}{\min\{\varepsilon,1\}}$, we prove that

$$\begin{aligned}\Pr\left[x + z \leq \frac{\tau}{\min\{\varepsilon,1\}}\right] \leq& \frac{4}{3} e^{\frac{\tau}{2} - \frac{mp_i \min\{\varepsilon,1\}}{3}} \leq O\left(e^{-\frac{\tau}{2} - \frac{mp_i \min\{1,\varepsilon\}}{6}}\right) &(234)\\ \leq& O\left(e^{-\frac{\tau}{2}} \frac{1}{m^\gamma \min\{1,\varepsilon\}^\gamma p_i^\gamma}\right) &(235)\end{aligned}$$

where (234) is by $\frac{\tau}{2} - \frac{mp_i \min\{1,\varepsilon\}}{3} < -\frac{\tau}{2} - \frac{mp_i \min\{1,\varepsilon\}}{6}$ for $p_i > \frac{t}{m \min\{1,\varepsilon\}}$ undersetting $t = 6\tau$, and the last inequality is by $e^{-\frac{x}{6}} \leq O\left(\frac{1}{x^\gamma}\right)$ for $x = mp_i \min\{1,\varepsilon\} \geq t \geq 1$ and $\gamma \geq 0$. □

We then prove two lemmas that bounds KL estimation error on (small) symbols below threshold and (large) symbols above threshold.

Lemma G.3 (Error of Algorithm 2 on Small Symbols - reused samples)**.** *Algorithm 2 satisfies*

$$\begin{aligned}\mathbb{E}\left[\sum_{i\in L} p_i \cdot \ln\left(\frac{\frac{p_i}{\sum_{i\in L} p_i}}{\frac{\bar{x}_i}{\sum_{i\in L} \bar{x}_i}}\right)\right] \leq& O\Bigg(\mathbb{E}\left[\left(\sum_{i\in L} p_i\right) \ln\left(1 + \frac{\sum_{i\in L} 1}{\min\{1,\varepsilon\} \sum_{i\in L} \alpha n p_i}\right)\right] &(236)\\ &+ \sum_{i:p_i > \frac{1}{\alpha n \min\{\varepsilon,1\}}} \left(\frac{1}{\alpha n} + \frac{1}{\alpha^2 n^2 \min\{\varepsilon,1\}^2 p_i}\right)\Bigg) &(237)\end{aligned}$$

where we have denoted $L = \left\{i : x_i + Lap\left(0, \frac{1}{\varepsilon}\right) \leq \frac{\tau}{\min\{\varepsilon,1\}}\right\}$ as the randomized instance-dependent subset in Algorithm 2.

Proof. By definition, we have that

$$\mathbb{E}\left[\sum_{i\in L} p_i \cdot \ln\left(\frac{\frac{p_i}{\sum_{i\in L} p_i}}{\frac{\bar{x}_i}{\sum_{i\in L} \bar{x}_i}}\right)\right] = \underbrace{\mathbb{E}\left[\sum_{i\in L} p_i \cdot \ln\left(\frac{\alpha n p_i}{\bar{x}_i}\right)\right]}_{①} + \underbrace{\mathbb{E}\left[\left(\sum_{i\in L} p_i\right) \ln\left(\frac{\sum_{i\in L} \bar{x}_i}{\sum_{i\in L} \alpha n p_i}\right)\right]}_{②} \tag{238}$$

We first analyze ①. By $\alpha n p_i \leq \min\{\varepsilon,1\} \leq \bar{x}_i$ under $p_i \leq \frac{1}{\alpha n \min\{\varepsilon,1\}}$, we compute that

$$① \leq \mathbb{E}\left[\sum_{i\in L: p_i > \frac{1}{\alpha n \min\{\varepsilon,1\}}} p_i \cdot \ln\left(\frac{\alpha n p_i}{\bar{x}_i}\right)\right] \tag{239}$$

We then analyze ②. By concavity of $\ln(t)$ over $t > 0$, we have that

$$② \le \mathbb{E}_L \left[\sum_{i \in L} p_i \ln \left(1 + \frac{\sum_{i \in L} \mathbb{E}[\bar{x}_i - \alpha n p_i | i \in L]}{\sum_{i \in L} \alpha n p_i} \right) \right] \tag{240}$$

$$\le \mathbb{E}_L \left[\left(\sum_{i \in L} p_i \right) \ln \left(1 + \frac{\sum\limits_{i \in L : p_i \le \frac{1}{\alpha n \min\{\varepsilon, 1\}}} \frac{2}{\min\{1, \varepsilon\}} + \sum\limits_{i \in L : p_i > \frac{1}{\alpha n \min\{\varepsilon, 1\}}} \mathbb{E}\left[\bar{x}_i - \alpha n p_i | i \in L \right]}{\sum_{i \in L} \alpha n p_i} \right) \right] \tag{241}$$

$$\le O \left(\mathbb{E} \left[\left(\sum_{i \in L} p_i \right) \ln \left(1 + \frac{\sum\limits_{i \in L} 1}{\min\{1, \varepsilon\} \sum_{i \in L} \alpha n p_i} \right) + \sum_{i \in L : p_i > \frac{1}{\alpha n \min\{\varepsilon, 1\}}} \frac{\bar{x}_i - \alpha n p_i}{\alpha n} \right] \right) \tag{242}$$

where we have denoted $L = \left\{ i : x_i + Lap\left(0, \frac{1}{\varepsilon}\right) \le \frac{\tau}{\min\{\varepsilon, 1\}} \right\}$ as the randomized instance-dependent subset in Algorithm 2. (241) is by applying Corollary B.5; (242) is by $\ln(1 + x + y) \le \ln(1 + x) + y$ for any $x \ge 0$ and any y.

By plugging (239) and (242) into (238), we prove that

$$\mathbb{E} \left[\sum_{i \in L} p_i \cdot \ln \left(\frac{\frac{p_i}{\sum_{i \in L} p_i}}{\frac{\bar{x}_i}{\sum_{i \in L} \bar{x}_i}} \right) \right] \tag{243}$$

$$\le O \left(\mathbb{E} \left[\left(\sum_{i \in L} p_i \right) \ln \left(1 + \frac{\sum\limits_{i \in L} 1}{\min\{1, \varepsilon\} \sum_{i \in L} \alpha n p_i} \right) + \sum_{i \in L : p_i > \frac{1}{\alpha n \min\{\varepsilon, 1\}}} \left(p_i \ln \left(\frac{\alpha n p_i}{\bar{x}_i} \right) + \frac{\bar{x}_i - \alpha n p_i}{\alpha n} \right) \right] \right)$$

$$\le O \left(\mathbb{E} \left[\left(\sum_{i \in L} p_i \right) \ln \left(1 + \frac{\sum\limits_{i \in L} 1}{\min\{1, \varepsilon\} \sum_{i \in L} \alpha n p_i} \right) \right] + \sum_{i : p_i > \frac{1}{\alpha n \min\{\varepsilon, 1\}}} \left(\frac{1}{\alpha n} + \frac{1}{\alpha^2 n^2 \min\{\varepsilon, 1\}^2 p_i} \right) \right) \tag{244}$$

where (244) is by applying Lemma B.10 under setting $m = \alpha n$, $c_1 = -\infty$ and $c_2 = \frac{\tau}{\min\{\varepsilon, 1\}}$, $b = \frac{1}{\varepsilon}$ and $c = \frac{1}{\min\{\varepsilon, 1\}}$ for $p_i > \frac{1}{\alpha n \min\{1, \varepsilon\}}$.

□

Lemma G.4 (Error of Algorithm 2 on Large Symbols - reused samples)**.** *Algorithm 2 satisfies*

$$\mathbb{E} \left[\sum_{i \notin L} \left(p_i \ln \left(\frac{\alpha n p_i}{\max\left\{ \tilde{x}_i, \frac{1}{\min\{\varepsilon, 1\}} \right\}} \right) + \frac{\max\left\{ \tilde{x}_i, \frac{1}{\min\{\varepsilon, 1\}} \right\} - \alpha n p_i}{\alpha n} \right) \right] \tag{245}$$

$$\le O \left(d \tau e^{-\frac{\tau}{2}} \cdot \frac{\mathbf{1}_{\min_i p_i \le \frac{1}{\alpha n \min\{\varepsilon, 1\}}}}{\alpha n \min\{\varepsilon, 1\}} + \sum_{i : p_i \ge \frac{1}{\alpha n \min\{\varepsilon, 1\}}} \left(\frac{1}{\alpha n} + \frac{1}{p_i} \cdot \frac{1}{\alpha^2 n^2 \min\{\varepsilon, 1\}^2} \right) \right) \tag{246}$$

where we have denoted $L = \left\{ i : x_i + Lap\left(0, \frac{1}{\varepsilon}\right) \le \frac{\tau}{\min\{\varepsilon, 1\}} \right\}$ *as the randomized instance-dependent subset in Algorithm 2.*

Proof. By definition and by $\tau \geq 1$, we compute that

$$\mathbb{E}\left[\sum_{i\notin L}\left(p_i \ln\left(\frac{\alpha n p_i}{\max\left\{\tilde{x}_i, \frac{1}{\min\{\varepsilon,1\}}\right\}}\right) + \frac{\max\left\{\tilde{x}_i, \frac{1}{\min\{\varepsilon,1\}}\right\} - \alpha n p_i}{\alpha n}\right)\right] \tag{247}$$

$$= \underbrace{\mathbb{E}\left[\sum_{i\notin L: p_i \geq \frac{1}{\alpha n \min\{1,\varepsilon\}}}\left(p_i \ln\left(\frac{\alpha n p_i}{\tilde{x}_i}\right) + \frac{\tilde{x}_i - \alpha n p_i}{\alpha n}\right)\right]}_{①} + \underbrace{\mathbb{E}\left[\sum_{i\notin L: p_i < \frac{1}{\alpha n \min\{1,\varepsilon\}}}\left(p_i \ln\left(\frac{\alpha n p_i}{\tilde{x}_i}\right) + \frac{\tilde{x}_i - \alpha n p_i}{\alpha n}\right)\right]}_{②} \tag{248}$$

We now analyze ①. By applying Lemma B.10 under setting $m = \alpha n$, $c_1 = \frac{\tau}{\min\{\varepsilon,1\}}$, $c_2 = +\infty$, $b = \frac{1}{\varepsilon}$ and $c = \frac{1}{\min\{1,\varepsilon\}}$, we prove that

$$① \leq O\left(\sum_{i: p_i \geq \frac{1}{\alpha n \min\{\varepsilon,1\}}}\left(\frac{1}{\alpha n} + \frac{1}{p_i}\cdot\frac{1}{\alpha^2 n^2 \min\{\varepsilon,1\}^2}\right)\right) \tag{249}$$

We finally analyze ②. By $\tilde{x}_i \geq \alpha n p_i$ for $i \notin L$ and $p_i < \frac{1}{\alpha n \min\{1,\varepsilon\}}$, we prove that

$$② \leq \underbrace{\sum_{i: p_i < \frac{1}{\alpha n \min\{\varepsilon,1\}}} \Pr[i \notin L]\cdot\frac{\mathbb{E}\left[\tilde{x}_i | i \notin L\right] - \alpha n p_i}{\alpha n}}_{\text{Error on FP}} \leq O\left(e^{-\frac{\tau}{2}}\cdot\frac{\frac{\tau}{\min\{\varepsilon,1\}} + \frac{1}{\varepsilon}}{\alpha n}\cdot d\cdot \mathbf{1}_{\min_i p_i \leq \frac{1}{\alpha n \min\{\varepsilon,1\}}}\right) \tag{250}$$

$$= O\left(d\tau e^{-\frac{\tau}{2}}\cdot\frac{\mathbf{1}_{\min_i p_i \leq \frac{1}{\alpha n \min\{\varepsilon,1\}}}}{\alpha n \min\{\varepsilon,1\}}\right) \tag{251}$$

where the second inequality in (250) is by applying Lemma G.1 under setting $m = \alpha n$, and by applying Lemma B.7 under setting $\lambda = \alpha n p_i$, $b = \frac{1}{\varepsilon}$ and $c = \frac{\tau}{\min\{\varepsilon,1\}}$. By plugging (249) and (251) into (248), we obtain the bound in the statement. □

Lemma G.5 (Error of Algorithm 2 on Large Symbols - fresh samples)**.** *Algorithm 2 satisfies*

$$\mathbb{E}\left[\sum_{i\notin L} p_i \ln\left(\frac{(1-\alpha) n p_i}{\max\left\{\tilde{x}'_i, \frac{1}{\min\{\varepsilon,1\}}\right\}}\right) + \frac{\max\left\{\tilde{x}'_i, \frac{1}{\min\{\varepsilon,1\}}\right\} - (1-\alpha) n p_i}{(1-\alpha)n}\right] \tag{252}$$

$$\leq O\left(\sum_{i: p_i \geq \frac{1}{\alpha n \min\{\varepsilon,1\}}}\left(\frac{1}{(1-\alpha)n} + \frac{1}{p_i}\cdot\frac{1}{(1-\alpha)^2 n^2 \min\{\varepsilon,1\}^2}\right) + d e^{-\frac{\tau}{2}}\cdot\frac{\mathbf{1}_{\min_i p_i \leq \frac{1}{\alpha n \min\{1,\varepsilon\}}}}{(1-\alpha)n\min\{\varepsilon,1\}}\right) \tag{253}$$

where we have denoted $L = \left\{i : x_i + Lap\left(0, \frac{1}{\varepsilon}\right) \leq \frac{\tau}{\min\{\varepsilon,1\}}\right\}$ *as the randomized instance-dependent subset in Algorithm 2.*

Proof. By the independence between set L and noisy estimate $\tilde{x}'_i$ (as they are computed on independently sampled datasets x and x' respectively), we compute that

$$(252) = \sum_i \Pr[i \notin L] \cdot \mathbb{E}\left[p_i \ln\left(\frac{(1-\alpha)np_i}{\max\left\{\tilde{x}'_i, \frac{1}{\min\{\varepsilon,1\}}\right\}}\right) + \frac{\max\left\{\tilde{x}'_i, \frac{1}{\min\{\varepsilon,1\}}\right\} - (1-\alpha)np_i}{(1-\alpha)n}\right] \tag{254}$$

$$\leq O\Bigg(\underbrace{\sum_{i:p_i \geq \frac{1}{\alpha n \min\{\varepsilon,1\}}} \left(\frac{1}{(1-\alpha)n} + \frac{1}{p_i} \cdot \frac{1}{(1-\alpha)^2 n^2 \min\{\varepsilon,1\}^2}\right)}_{\text{Error on TP}} + \underbrace{\frac{1}{(1-\alpha)n\min\{1,\varepsilon\}} \sum_{i:p_i < \frac{1}{\alpha n \min\{\varepsilon,1\}}} e^{-\frac{\tau}{2}}}_{\text{Error on FP}}\Bigg) \tag{255}$$

$$\leq O\left(\sum_{i:p_i \geq \frac{1}{\alpha n \min\{\varepsilon,1\}}} \left(\frac{1}{(1-\alpha)n} + \frac{1}{p_i} \cdot \frac{1}{(1-\alpha)^2 n^2 \min\{\varepsilon,1\}^2}\right) + de^{-\frac{\tau}{2}} \cdot \frac{\mathbf{1}_{\min_i p_i \leq \frac{1}{\alpha n \min\{1,\varepsilon\}}}}{(1-\alpha)n\min\{\varepsilon,1\}}\right) \tag{256}$$

where the first term in (255) is by $\Pr[i \notin L] \leq 1$ for $p_i \geq \frac{1}{\alpha n \min\{\varepsilon,1\}}$ (by definition) and by applying Lemma B.10 under $m = (1-\alpha)n$, $b = \frac{1}{\varepsilon}$, $c = \frac{1}{\min\{1,\varepsilon\}}$, $c_1 = -\infty$ and $c_2 = +\infty$; the second term in (255) is by applying Lemma G.1 under $m = \alpha n$; (256) is by definition. □

We are now ready to prove the per-instance upper bound for Algorithm 2.

Theorem G.6 (DP "Sampling Twice" Algorithm)**.** *The estimator $\mathcal{A}$ given by Algorithm 2 is ε-DP and satisfies the following error bound for any fixed $p \in \Delta(d)$.*

$$\mathop{\mathbb{E}}_{x \sim Poi(n,p)}\Big[KL(p\|\mathcal{A}(x))\Big] \leq O\Bigg(\mathbb{E}\left[\left(\sum_{i\in L} p_i\right)\ln\left(1 + \frac{\sum_{i\in L} 1}{\min\{1,\varepsilon\}\sum_{i\in L} np_i}\right) + \frac{\mathbf{1}_{L\neq\emptyset}}{n\min\{1,\varepsilon\}}\right] \tag{257}$$

$$+ \sum_{i:p_i \geq \frac{1}{n\min\{1,\varepsilon\}}} \frac{1}{p_i} \cdot \frac{1}{n^2\min\{\varepsilon,1\}^2}\Bigg) \tag{258}$$

where $L = \left\{i : x_i + Lap\left(0, \frac{1}{\varepsilon}\right) \leq \frac{\tau}{\min\{\varepsilon,1\}}\right\}$ is as defined in Algorithm 2.

Proof. By the definition of Algorithm 2, we compute that

$$\mathbb{E}\left[KL(p, \mathcal{A}(x))\right] \tag{259}$$

$$= \mathbb{E}\left[\sum_{i\in L} p_i \ln\left(\frac{p_i/\sum_{j\in L} p_j}{\bar{x}_i/\sum_{j\in L}\bar{x}_j}\right) + \left(\sum_{i\in L} p_i\right)\ln\left(\frac{\sum_{i\in L} p_i}{\frac{\tilde{c}}{(1-\alpha)n}}\right) + \sum_{i\notin L} p_i \ln\left(\frac{p_i}{\frac{\bar{x}_i}{(1-\alpha)n}}\right) + \ln\left(\frac{\tilde{c} + \sum_{i\notin L}\bar{x}_i}{(1-\alpha)n}\right)\right]$$

$$\leq \underbrace{\mathbb{E}\left[\sum_{i\in L} p_i \ln\left(\frac{p_i/\sum_{j\in L} p_j}{\bar{x}_i/\sum_{j\in L}\bar{x}_j}\right)\right]}_{①} + \underbrace{\mathbb{E}\left[\left(\sum_{i\in L} p_i\right)\ln\left(\frac{\sum_{i\in L}(1-\alpha)np_i}{\tilde{c}}\right) + \frac{\tilde{c} - \sum_{i\in L}(1-\alpha)np_i}{(1-\alpha)n}\right]}_{②} \tag{260}$$

$$+ \underbrace{\mathbb{E}\left[\sum_{i\notin L}\left(p_i \ln\left(\frac{(1-\alpha)np_i}{\bar{x}_i}\right) + \frac{\bar{x}_i - (1-\alpha)np_i}{(1-\alpha)n}\right)\right]}_{③} \tag{261}$$

where the last inequality is by $\ln(t) \le t-1$ for any $t>0$. We first analyze (1). By applying Lemma G.3, we prove that

$$(1) \le O\left(\mathbb{E}\left[\left(\sum_{i\in L} p_i\right)\ln\left(1+\frac{\sum_{i\in L}1}{\min\{1,\varepsilon\}\sum_{i\in L}\alpha n p_i}\right)\right]+\sum_{i:p_i>\frac{1}{\alpha n\min\{\varepsilon,1\}}}\left(\frac{1}{\alpha n}+\frac{1}{\alpha^2 n^2\varepsilon^2 p_i}\right)\right) \tag{262}$$

$$\le O\left(\mathbb{E}\left[\left(\sum_{i\in L} p_i\right)\ln\left(1+\frac{\sum_{i\in L}1}{\min\{1,\varepsilon\}\sum_{i\in L} n p_i}\right)\right]+\sum_{i:p_i>\frac{1}{n\min\{\varepsilon,1\}}}\left(\frac{1}{n}+\frac{1}{n^2\varepsilon^2 p_i}\right)\right) \tag{263}$$

where (263) is by $\alpha=0.5$ in Algorithm 2. We then analyze (2). By the independence between L and $\tilde{c}$ in Algorithm 2 (due to independent sampling of datasets x and x'), conditioned on fixed L, we apply Lemma B.10 under setting $m=(1-\alpha)n$, $p=\sum_{i\in L}p_i$, $c_1=-\infty$, $c_2=+\infty$, $b=\frac{1}{\varepsilon}$ and $c=\frac{1}{\min\{1,\varepsilon\}}$ and prove that

$$(2) \le O\left(\mathbb{E}_L\left[\frac{\mathbf{1}_{L\neq\emptyset}}{(1-\alpha)n\min\{1,\varepsilon\}}\right]\right) \le \mathbb{E}\left[\frac{\mathbf{1}_{L\neq\emptyset}}{n\min\{1,\varepsilon\}}\right] \tag{264}$$

where (264) is by setting $\alpha=0.5$ in Algorithm 2. We finally analyze (3). By definition of $\bar{x}_i$ for $i\notin L$, we compute that

$$(3) = \mathbb{E}\Bigg[\sum_{i\notin L}\Bigg(p_i\ln\left(\frac{np_i}{\max\left\{\tilde{x}_i,\frac{1}{\min\{\varepsilon,1\}}\right\}+\max\left\{\tilde{x}'_i,\frac{1}{\min\{\varepsilon,1\}}\right\}}\right) \tag{265}$$

$$+\frac{\max\left\{\tilde{x}_i,\frac{1}{\min\{\varepsilon,1\}}\right\}+\max\left\{\tilde{x}'_i,\frac{1}{\min\{\varepsilon,1\}}\right\}-np_i}{n}\Bigg)\Bigg] \tag{266}$$

$$\le \alpha\cdot\mathbb{E}\left[\sum_{i\notin L}\left(p_i\ln\left(\frac{\alpha np_i}{\max\left\{\tilde{x}_i,\frac{1}{\min\{\varepsilon,1\}}\right\}}\right)+\frac{\max\left\{\tilde{x}_i,\frac{1}{\min\{\varepsilon,1\}}\right\}-\alpha np_i}{\alpha n}\right)\right]$$

$$+(1-\alpha)\cdot\mathbb{E}\left[\sum_{i\notin L}p_i\ln\left(\frac{(1-\alpha)np_i}{\max\left\{\tilde{x}'_i,\frac{1}{\min\{\varepsilon,1\}}\right\}}\right)+\frac{\max\left\{\tilde{x}'_i,\frac{1}{\min\{\varepsilon,1\}}\right\}-(1-\alpha)np_i}{(1-\alpha)n}\right] \tag{267}$$

$$\le O\left(\sum_{i:p_i\ge\frac{1}{n\min\{\varepsilon,1\}}}\frac{1}{p_i}\cdot\frac{1}{n^2\min\{\varepsilon,1\}^2}+\frac{\mathbf{1}_{\min_i p_i\le\frac{2}{n\min\{1,\varepsilon\}}}}{n\min\{\varepsilon,1\}}\right) \tag{268}$$

$$\le O\left(\sum_{i:p_i\ge\frac{2}{n\min\{\varepsilon,1\}}}\frac{1}{p_i}\cdot\frac{1}{n^2\min\{\varepsilon,1\}^2}+\frac{\Pr[L\neq\emptyset]}{n\min\{\varepsilon,1\}}\right) \tag{269}$$

where (267) is by the joint convexity of the function $x\ln\left(\frac{x}{y}\right)$ with regard to arguments $x,y\ge 0$; (268) is by applying Lemma G.4 and Lemma G.5 under setting $\alpha=0.5$ and $\tau=4\ln d$ as in Algorithm 2; and (269) is by $\Pr[i\in L]\ge\Omega(1)$ for $p_i<\frac{2}{n\min\{1,\varepsilon\}}$ under $\alpha=0.5$ and $\tau=4\ln d$ in Algorithm 2 (by applying Lemma B.2 under setting $a=\alpha np_i$, $b=\frac{1}{\varepsilon}$ and $c=\frac{\tau}{n\min\{1,\varepsilon\}}$). By plugging our proved bound (263), (264) and (269) for (1), (2), and (3) into (261), we prove the bound in the statement. □

G.2 Proof for Matching Lower and Upper Bound

Corollary G.7. *Let $\mathcal{A}$ be the estimator given by Algorithm 2. Let N_{stat}, $N_{\leq \frac{t}{n}}$, $N_{\frac{1}{n\varepsilon}}$, $N_{\leq \frac{t}{n\varepsilon}}$ be the additive neighborhoods defined in* (137), (136), (6) *respectively. Then for any n and any $p \in \Delta(d)$,*

$$\mathop{\mathbb{E}}_{x\sim Poi(n,p)}\Big[KL(p\|\mathcal{A}(x))\Big] \leq O\Bigg(\underbrace{lower\left(p,n,N_{stat}\right)+lower\left(p,n,N_{\leq\frac{t}{n}}\right)}_{\text{Non-DP Per-instance Lower Bound}} + \underbrace{lower_{\varepsilon,\delta}\left(p,n,N_{\frac{1}{n\varepsilon}}\right)+lower_{\varepsilon,\delta}\left(p,n,N_{\leq\frac{t}{n\varepsilon}}\right)}_{\text{DP Per-instance Lower Bound}}\Bigg) \tag{270}$$

under choosing neighborhood size $t = 6\tau$ for $\tau = 4\ln d$.

Proof. We will use the upper bound given by Theorem G.6. Observe that for any $t > 0$, $\frac{\sum_{i\in L}1}{\min\{1,\varepsilon\}\sum_{i\in L}np_i} \leq \frac{d_{small}(L)}{\min\{1,\varepsilon\}np_{small}(L)}$, where $d_{small}(L) = \sum_{i\in L:p_i\leq\frac{t}{\alpha n\min\{\varepsilon,1\}}} 1$ and $p_{small}(L) = \sum_{i\in L:p_i\leq\frac{t}{\alpha n\min\{\varepsilon,1\}}} p_i$. Thus the first term in Theorem G.6 satisfies

$$\mathbb{E}\left[\left(\sum_{i\in L}p_i\right)\ln\left(1+\frac{\sum_{i\in L}1}{\min\{1,\varepsilon\}\sum_{i\in L}np_i}\right)+\frac{\mathbf{1}_{L\neq\emptyset}}{n\min\{1,\varepsilon\}}\right] \tag{271}$$

$$\leq\mathbb{E}\left[p_{small}(L)\cdot\ln\left(1+\frac{d_{small}(L)}{\min\{1,\varepsilon\}np_{small}(L)}\right)\right]+\sum_{i:p_i>\frac{t}{\alpha n\min\{\varepsilon,1\}}}p_i\cdot\ln\left(1+\frac{d}{t}\right)\cdot\Pr[i\in L] \tag{272}$$

$$+\mathbb{E}\left[\frac{\mathbf{1}_{\min_{i\in L}p_i\leq\frac{t}{\alpha n\min\{\varepsilon,1\}}}}{n\min\{\varepsilon,1\}}\right]+\sum_{i:p_i\geq\frac{t}{\alpha n\min\{\varepsilon,1\}}}\frac{\Pr[i\in L]}{n\min\{\varepsilon,1\}} \tag{273}$$

$$\leq\mathbb{E}\left[p_{small}(L)\ln\left(\frac{d_{small}(L)}{p_{small}(L)\cdot n\min\{1,\varepsilon\}}\right)+\frac{\ln\left(1+d_{small}(L)\right)}{n\min\{\varepsilon,1\}}\right]+O\left(\sum_{i:p_i>\frac{t}{n\min\{\varepsilon,1\}}}\frac{\ln de^{-\frac{\tau}{2}}+e^{-\frac{\tau}{2}}}{n^2\min\{1,\varepsilon\}^2p_i}\right) \tag{274}$$

$$\leq\max_{L'\subset[d]}\left(p_{small}(L')\ln\left(\frac{d_{small}(L')}{p_{small}(L')\cdot n\min\{1,\varepsilon\}}\right)+\frac{\ln\left(1+d_{small}(L')\right)}{n\min\{\varepsilon,1\}}\right)+O\left(\sum_{i:p_i>\frac{t}{n\min\{\varepsilon,1\}}}\frac{1}{n^2\min\{1,\varepsilon\}^2p_i}\right) \tag{275}$$

$$=O\left(\text{Theorem D.3} + \text{Theorem F.2} + \text{Theorem D.4} + \text{Theorem F.1}\right) \tag{276}$$

where (272) is by $\frac{\sum_{i\in L}1}{\min\{1,\varepsilon\}\sum_{i\in L}np_i} \leq \frac{d_{small}(L)}{\min\{1,\varepsilon\}np_{small}(L)}$, for any $t > 0$ and $d_{small}(L) = \sum_{i\in L:p_i\leq\frac{t}{n\min\{\varepsilon,1\}}} 1$ and $p_{small}(L) = \sum_{i\in L:p_i\leq\frac{t}{n\min\{\varepsilon,1\}}} p_i$; (273) is $\Pr[L \neq \emptyset] \leq \Pr\left[\min_{i\in L}p_i \leq \frac{t}{\alpha n\min\{\varepsilon,1\}}\right] + \sum_{i:p_i\geq\frac{t}{\alpha n\min\{\varepsilon,1\}}}\Pr[i \in L]$ (ensured by union bound); (274) is by $\mathbf{1}_{\min_{i\in L}p_i\leq\frac{t}{\alpha n\min\{\varepsilon,1\}}\}} \leq \ln\left(1+\sum_{i\in L:p_i\leq\frac{t}{\alpha n\min\{\varepsilon,1\}}}1\right) = \ln\left(1+d_{small}(L)\right)$, and by applying Lemma G.2 under setting $m = \alpha n$ and choosing $t = 6\tau$; and (275) is by setting $\tau = 4\ln d$ as defined in Algorithm 2, and (276) is by definition. Additionally, observe that by definition, the second term in Theorem G.6 is $\sum_{i:p_i>\frac{1}{n\min\{1,\varepsilon\}}}\frac{1}{p_i}\cdot\frac{1}{n^2\min\{\varepsilon,1\}^2} \leq O\left(\text{Theorem D.4} + \text{Theorem F.1}\right)$. Combining this with (276) suffice to prove the bound in the statement. □

G.3 Discussions on the Neighborhood Size

Lemma G.8 (Generalized Packing Argument)**.** *Let $d \geq 16 \in \mathbb{N}$ and let $\mathcal{O}$ be an output space. Let $err : \mathcal{O} \times \mathcal{O} \to \mathbb{R}$ be an error function. Given $p^1, \cdots, p^d \in \mathcal{O}$. For $i \in [d]$, denote $\mathcal{S}(p^i)$ as the distribution of histogram sampled from distribution p^i. Assume that*

1. *for any $i, j \in [d]$,*

$$\mathbb{E}_{(x,x')}\left[\|x - x'\|_1\right] \leq \frac{\ln d}{16\varepsilon} \tag{277}$$

   for a coupling (x, x') between the distributions $\mathcal{S}(p^i)$ and $\mathcal{S}(p^j)$;

2. *for any $q \in \mathcal{O}$ and any $S \subseteq [d]$ such that $|S| \geq d^{1/4}$, it holds that*

$$\frac{1}{S}\sum_{i \in S} err\left(p^i, q\right) \geq \frac{\ln d}{4} \tag{278}$$

Then for any (ε, δ)-DP algorithm $\mathcal{A}$ with $\delta < \frac{\varepsilon}{d^{1/4}\ln d}$, we have

$$\max_{i \in d} \mathbb{E}_{x \sim \mathcal{S}(p^i)}\left[err\left(p^i, \mathcal{A}(x)\right)\right] \geq \frac{\ln d}{16} \tag{279}$$

Proof. We will prove the lemma by contradiction. Consider a bipartite graph (V, E) on $V = [d] \times \mathcal{O}$, where $(i, v) \in E$ if and only if $\text{err}(p^i, q) < \frac{\ln d}{4}$. Then

$$\text{For any } v \in \mathcal{O}, \quad \text{degree}(v) < d^{1/4} \tag{280}$$

Otherwise the set of neighbors $\text{Nbr}(v) = \{i \in [d] : (i, v) \in E\}$ violates (278).

Suppose that (279) does not hold, then

$$\text{For any } i \in [d], \quad \mathbb{E}_{x \sim \mathcal{S}(p^i)}\left[\text{err}\left(p^i, \mathcal{A}(x)\right)\right] < \frac{\ln d}{16} \tag{281}$$

Denote the neighbor set $\text{Nbr}(i) = \{v \in \mathcal{O} : (i, v) \in E\}$. Then by Markov's inequality

$$\Pr_{x \sim \mathcal{S}(p^i)}\left[\mathcal{A}(x) \in \text{Nbr}(i)\right] = 1 - \Pr_{x \sim \mathcal{S}(p^i)}\left[\text{err}\left(p^i, \mathcal{A}(x)\right) \geq \frac{\ln d}{4}\right] \geq 1 - \frac{\mathbb{E}_{x \sim \mathcal{S}(p^i)}\left[\text{err}\left(p^i, \mathcal{A}(x)\right)\right]}{\frac{\ln d}{4}} \geq \frac{3}{4} \tag{282}$$

where the last equality is by (281).

On the other hand, for any $j \in [d]$, let (x, x') be the coupling between $\mathcal{S}(p^i)$ and $\mathcal{S}(p^j)$ in (277), we prove that

$$\Pr_{x \sim \mathcal{S}(p^i)}\left[\mathcal{A}(x) \in \text{Nbr}(i)\right] \leq \Pr_{(x,x')}\left[\mathcal{A}(x) \in \text{Nbr}(i) \text{ and } \|x - x'\|_1 \leq \frac{\ln d}{4\varepsilon}\right] + \Pr_{(x,x')}\left[\|x - x'\|_1 > \frac{\ln d}{4\varepsilon}\right] \tag{283}$$

$$\leq e^{\varepsilon \cdot \frac{\ln d}{4\varepsilon}} \cdot \Pr_{(x,x')}\left[\mathcal{A}(x') \in \text{Nbr}(i)\right] + \frac{\ln d}{4\varepsilon} \cdot e^{\varepsilon \cdot \frac{\ln d}{4\varepsilon}} \cdot \delta + \frac{\mathbb{E}_{(x,x')}\left[\|x - x'\|_1\right]}{\frac{\ln d}{4\varepsilon}} \tag{284}$$

$$= d^{1/4} \cdot \Pr_{x \sim \mathcal{S}(p^j)}\left[\mathcal{A}(x) \in \text{Nbr}(i)\right] + \frac{1}{2} \tag{285}$$

where (283) is by applying union bound; the first term in (284) is by recursive usage of definition of (ε, δ)-DP to datasets (x, x') with hamming distance bounded by $\frac{\ln d}{4\varepsilon}$; the second term in (284) is by applying Markov's inequality; and (285) is by applying condition (277) and $\delta \leq \frac{\varepsilon}{d^{1/4}\ln d}$.

By combining (282) and (285), it follows that

$$\Pr_{x \sim \mathcal{S}(p^j)}\left[\mathcal{A}(x) \in \text{Nbr}(i)\right] \geq \frac{1}{4d^{1/4}} \tag{286}$$

By summing (286) over all i, we prove that

$$\sum_{i=1}^{d} \Pr_{x\sim\mathcal{S}(p^j)}\left[\mathcal{A}(x) \in \text{Nbr}(i)\right] \geq \frac{d^{3/4}}{4} \tag{287}$$

Thus there exists $v \in \mathcal{O}$, such that

$$\sum_{i=1}^{d} \mathbf{1}_{v\in\text{Nbr}(i)} \geq \frac{d^{3/4}}{4} \geq d^{1/4} \tag{288}$$

where the last inequality is by $d \geq 16$. This contradicts with (280). □

Theorem G.9. *Let $n, d \geq 4 \in \mathbb{N}$, $\varepsilon = \frac{\ln(d)}{16n}$, and $0 \leq \delta \leq \frac{\varepsilon}{d^{1/4}\ln d}$. For $\gamma \leq \frac{1}{32}$, let $N_\gamma(p)$ be the local neighborhood as defined in* (6) *for $t = \gamma \ln d$.*

$$N_\gamma(p) = \left\{ q : |q_i - p_i| \leq \frac{\gamma \ln d}{n\varepsilon} \text{ for any } i \in [d] \text{ and } \sum_{i:p_i \leq \frac{\gamma\ln d}{n\varepsilon}} q_i \leq \max\left\{ \frac{\gamma \ln d}{n\varepsilon}, \sum_{i:p_i\leq\frac{\gamma\ln d}{n\varepsilon}} p_i \right\} \right\} \tag{289}$$

Then there exists a set $\mathcal{P}$ of distribution instances on $\Delta(d)$, and a per-neighborhood estimator $\mathcal{A}_{N_\gamma(p)}$ under neighborhood size γ, such that

$$\max_{p\in\mathcal{P}} \max_{q\in N_\gamma(p)} \mathop{\mathbb{E}}_{x\sim Poi(nq)} \left[KL\left(q, \mathcal{A}_{N_\gamma(q)}(x)\right)\right] \leq 48\gamma \ln d \tag{290}$$

while for any (ε, δ)-DP estimator $\mathcal{A}$, we have

$$\max_{p\in\mathcal{P}} \mathop{\mathbb{E}}_{x\sim Poi(np)} \left[KL\left(p, \mathcal{A}(x)\right)\right] \geq \frac{\ln d}{16} \tag{291}$$

Thus if $\gamma \leq o(1)$, then no (ε, δ)-DP estimator $\mathcal{A}$ could satisfy (2) *(otherwise it contradicts* (290) *and* (291)*).*

Proof. Consider the following construction of $\mathcal{P}$.

$$\mathcal{P} = \left\{ p^i \coloneqq \delta_i \text{ for } i \in [d] \right\} \tag{292}$$

Then

$$N_\gamma(p^i) = \left\{ q : q_i \geq 1 - \gamma \cdot \frac{\ln d}{n\varepsilon}, q_j \leq \gamma \cdot \frac{\ln d}{n\varepsilon} \text{ for any } j \neq i \right\} \tag{293}$$

Thus the following construction of per-neighborhood estimator satisfies (290).

$$\mathcal{A}_{N_\gamma(p^i)}(x)_j = \begin{cases} 1 - \gamma \cdot \frac{\ln d}{n\varepsilon} & j = i \\ \gamma \cdot \frac{\ln d}{n\varepsilon\cdot(d-1)} & j \neq i \end{cases} \tag{294}$$

This is because

$$\max_{q\in N_\gamma(p^i)} \mathop{\mathbb{E}}_{x\sim\text{Poi}(nq)} \left[KL(q, \mathcal{A}_{N_\gamma(p^i)})\right] \leq \ln\left(\frac{1}{1-\gamma\cdot\frac{\ln d}{n\varepsilon}}\right) + \gamma\cdot\frac{\ln d}{n\varepsilon}\cdot\ln(d-1) \tag{295}$$

$$\leq 2\gamma\cdot\frac{\ln d}{n\varepsilon} + \gamma\cdot\frac{(\ln d)^2}{n\varepsilon} \leq 48\gamma\cdot\ln d \tag{296}$$

where (296) is by $\gamma\cdot\frac{\ln d}{n\varepsilon} \leq \frac{1}{2}$ under $\gamma \leq \frac{1}{32}$ and by $\varepsilon = \frac{\ln d}{2n}$.

Below we focus on proving (291) by applying Lemma G.8. We only need to validate that the two conditions of Lemma G.8 hold.

1. The first condition (277) holds by $\varepsilon = \frac{\ln d}{16n}$.

2. The second condition of (278) holds by convexity of the function $\ln(\frac{1}{t})$ on $t > 0$, which ensures that for any $S \subseteq [d]$ with $|S| \geq d^{1/4}$ and any $q \in \Delta(d)$, we have

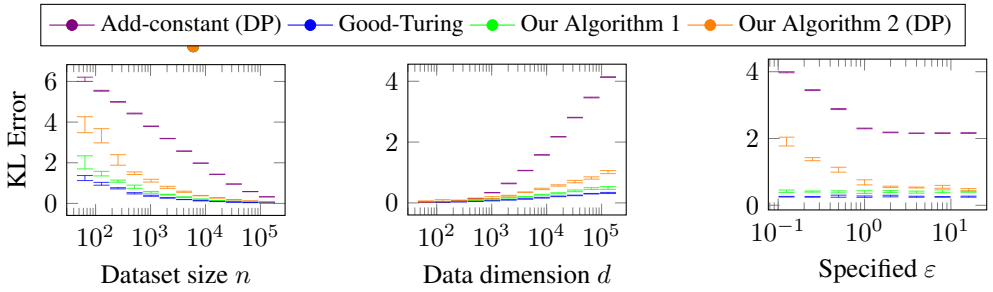
$$\frac{1}{|S|}\sum_{i\in S} KL(p^i, q) = \frac{1}{|S|}\sum_{i\in S} \ln\left(\frac{1}{q_i}\right) \geq \ln\left(\frac{1}{\frac{1}{|S|}\sum_{i\in S} q_i}\right) \geq \ln(|S|) \geq \frac{1}{4}\ln d \quad (297)$$

where the second-to-last inequality is by $\sum_{i\in S} q_i \leq 1$, and the last inequality is by $|S| \geq d^{1/4}$.

□

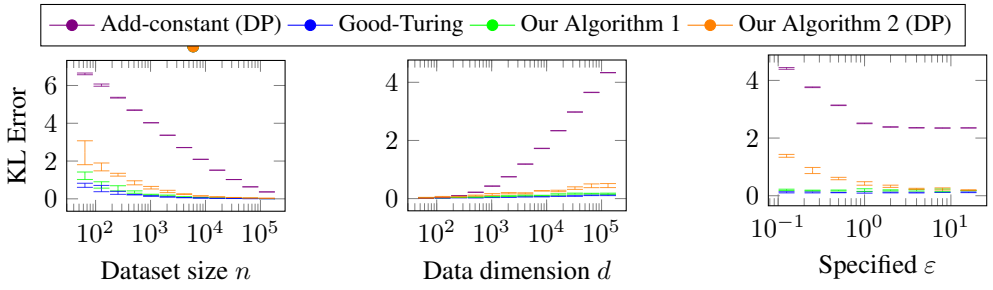
G.4 Additional Experiments on More Datasets

In this section, we further evaluate on more data distributions: power law distributions $p_i \propto \frac{1}{i^\beta}$ for $\beta = 1.5, 2$ in Figure 3, and 4; Enron-email corpus in Figure 5; and MMLU corpus in Figure 6.



(a) $d = 50000$, $\varepsilon = 1$ (b) $n = 2000$, $\varepsilon = 1$ (c) $n = 1000$, $d = 10000$

Figure 3: (Power law distribution $p_i \propto \frac{1}{i^{1.5}}$) KL error versus dataset size n, distribution dimension d, and DP guarantee ε for our methods compared with the simple minimax optimal Add-constant (DP) baseline, and the strongest non-DP baseline of prior (near) instance-optimal Good-Turing estimator.

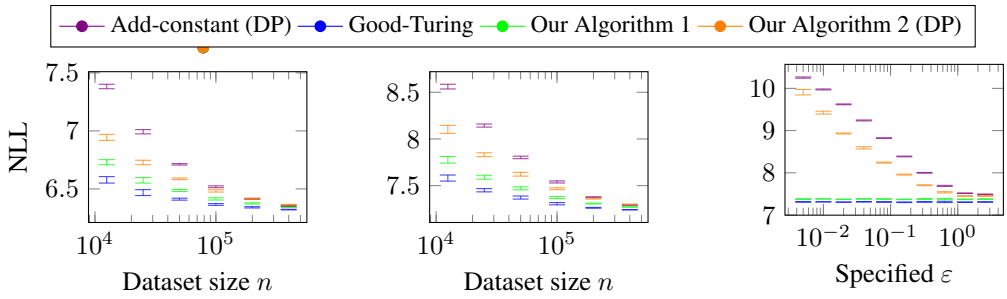


(a) $d = 50000$, $\varepsilon = 1$ (b) $n = 2000$, $\varepsilon = 1$ (c) $n = 1000$, $d = 10000$

Figure 4: (Power law distribution $p_i \propto \frac{1}{i^2}$) KL error versus dataset size n, distribution dimension d, and DP guarantee ε for our methods compared with the simple minimax optimal Add-constant (DP) baseline, and the strongest non-DP baseline of prior (near) instance-optimal Good-Turing estimator.

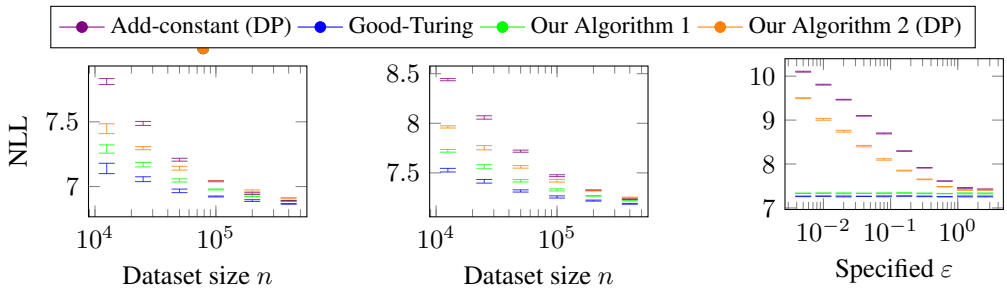
G.5 Instance Optimality of Gaussian Variant of Our Algorithm

Below we present a Gaussian variant of our algorithm, and prove that it is near instance optimal up to a $\log(1/\delta)$ factor.



(a) $d = 30522$, $\varepsilon = 1$ (Bert) (b) $d = 50257$, $\varepsilon = 1$ (GPT2) (c) $n = 10^5$, $d = 50257$ (GPT2)

Figure 5: (Enron-emails Token Distribution Estimation) KL error versus dataset size n, distribution dimension d, and DP guarantee ε for our methods compared with the simple minimax optimal Add-constant (DP) baseline, and the strongest non-DP baseline of prior (near) instance-optimal Good-Turing estimator.



(a) $d = 30522$, $\varepsilon = 1$ (Bert) (b) $d = 50257$, $\varepsilon = 1$ (GPT2) (c) $n = 10^5$, $d = 50257$ (GPT2)

Figure 6: (MMLU Token Distribution Estimation) KL error versus dataset size n, distribution dimension d, and DP guarantee ε for our methods compared with the simple minimax optimal Add-constant (DP) baseline, and the strongest non-DP baseline of prior (near) instance-optimal Good-Turing estimator.

Algorithm 3 (ε, δ)-DP "Sampling Twice" (Instance-Optimal)

Inputs: Data partition ratio $\alpha = 0.5$. Independently sampled datasets $x \sim \text{Poi}(\alpha \cdot np)$ and $x' \sim \text{Poi}((1-\alpha) \cdot np)$ s.t. $x + x' \sim \text{Poi}(np)$. Threshold $\tau = 4 \ln d$. Noise magnitude $\sigma = \frac{\sqrt{\ln(1/\delta)}}{\varepsilon}$.

$L = \emptyset$

for symbol $i = 1, \cdots, d$ **do**

Private Thresholding: If $\tilde{x}_i := x_i + z_i \leq \frac{\tau \cdot \sqrt{2\ln(1.25/\delta)}}{\min\{\varepsilon, 1\}}$ for $z_i \sim \mathcal{N}\left(0, \sigma^2\right)$, add i to L

end for

Estimate small symbols' combined mass: $\tilde{c} = \max\left\{\sum_{i \in L} x'_i + \mathcal{N}\left(0, \sigma^2\right), \frac{\sqrt{2\ln(1.25/\delta)}}{\min\{\varepsilon, 1\}}\right\}$

Estimate individual large symbols: for $i \in [d] \setminus L$, $\tilde{x}'_i = x'_i + z'_i$ for $z'_i \sim \mathcal{N}\left(0, \sigma^2\right)$

Truncation: $\bar{x}_i = \max\left\{\tilde{x}_i, \frac{\sqrt{2\ln(1.25/\delta)}}{\min\{\varepsilon, 1\}}\right\}$ for $i \in L$; $\bar{x}_i = (1-\alpha)\left(\max\left\{\tilde{x}_i, \frac{\sqrt{2\ln(1.25/\delta)}}{\min\{\varepsilon, 1\}}\right\} + \max\left\{\tilde{x}'_i, \frac{\sqrt{2\ln(1.25/\delta)}}{\min\{\varepsilon, 1\}}\right\}\right)$ for $i \in [d] \setminus L$

Return $\mathcal{A}(x)$ with $\mathcal{A}(x)_i = \begin{cases} \frac{1}{N} \cdot \tilde{c} \cdot \frac{\bar{x}'_i}{\sum_{i \in L} \bar{x}'_i} & i \in L \\ \frac{1}{N} \cdot \bar{x}'_i & i \notin L \end{cases}$ where $N = \tilde{c} + \sum_{i \notin L} \bar{x}_i$

Theorem G.10. *The estimator $\mathcal{A}$ given by Algorithm 3 is (ε, δ)-DP and satisfies the following error bound for any fixed $p \in \Delta(d)$.*

$$\mathop{\mathbb{E}}_{x\sim Poi(n,p)}\Big[KL(p\|\mathcal{A}(x))\Big] \leq 2\ln(1.25/\delta)\cdot O\Bigg(\mathbb{E}\left[\left(\sum_{i\in L} p_i\right)\ln\left(1+\frac{\sum_{i\in L}1}{\min\{1,\varepsilon\}\sum_{i\in L} np_i}\right)+\frac{\mathbf{1}_{L\neq\emptyset}}{n\min\{1,\varepsilon\}}\right] \tag{298}$$

$$+\sum_{i:p_i\geq\frac{1}{n\min\{1,\varepsilon\}}}\frac{1}{p_i}\cdot\frac{1}{n^2\min\{\varepsilon,1\}^2}\Bigg) \tag{299}$$

where L is the randomized set as defined in Algorithm 3.

Proof. The (ε, δ)-DP guarantee follows by observing that the ℓ_2-sensitivity of vector release $(x_1, \cdots, x_d, x'_1, \cdots, x'_d)$ is one, and by applying the (ε, δ)-DP guarantee for Gaussian mechanism in [27, Theorem A.1]. The KL error bound follows identically as Appendix G.1, after replacing all Laplace Tail bounds under $\text{Lap}(1/\varepsilon)$ with Gaussian Tail bounds under $\mathcal{N}(0, \sigma^2)$ for $\sigma = \frac{\sqrt{2\ln(1.25/\delta)}}{\varepsilon}$. □