Sublinear Algorithms for Estimating Wasserstein and TV Distances: Applications to Fairness and Privacy Auditing

Anonymous authors
Paper under double-blind review

Abstract

Resource-efficiently computing representations of probability distributions and the distances between them while only having access to the samples is a fundamental and useful problem across mathematical sciences. In this paper, we propose a generic framework to learn the probability and cumulative distribution functions (PDFs and CDFs) of a sub-Weibull, i.e. almost any light- or heavy-tailed, distribution while the samples from it arrive in a stream. The idea is to reduce these problems into estimating the frequency of an *appropriately chosen subset* of the support of a *properly discretised distribution*. We leverage this reduction to compute mergeable summaries of distributions from the stream of samples while requiring only sublinear space relative to the number of observed samples. This allows us to estimate Wasserstein and Total Variation (TV) distances between any two distributions while samples arrive in streams and from multiple sources. Our algorithms significantly improves on the existing methods for distance estimation incurring super-linear time and linear space complexities, and further extend the mergeable summaries framework to continuous distributions with possibly infinite support. Our results are tight with respect to the existing lower bounds for bounded discrete distributions. In addition, we leverage our proposed estimators of Wasserstein and TV distances to tightly audit the fairness and privacy of algorithms. We empirically demonstrate the efficiency of proposed algorithms across synthetic and real-world datasets.

1 Introduction

Computing distances between probability distributions while having access to samples is a fundamental and practical problem in statistics, computer science, information theory, and many other fields. Among the multitude of distances proposed between distributions, we focus on the Wasserstein and Total Variation distances in this work.

Wasserstein Distance. Optimal transport (Villani, 2009) has emerged as a popular and useful tool in machine learning. It involves computing Wassesrstein distance between distributions and using it further for learning. The statistical and computational aspects of Wasserstein distances have been extensively studied, refer to Panaretos & Zemel (2018); Peyré & Cuturi (2019); Chewi et al. (2024). This geometrically insightful approach has found applications in text document similarity measurement (Kusner et al., 2015), dataset distances (Alvarez-Melis & Fusi, 2020), domain adaption (Courty et al., 2016), generative adversarial networks (Arjovsky et al., 2017), data selection and valuation (Just et al., 2024; Kang et al., 2024), fair classification and regression (Jiang et al., 2020; Chzhen et al., 2020) to name a few. Thus, it is a fundamental question to compute Wasserstein distance between two distributions when we only have access to samples from it. But computing Wasserstein distance is often memory and computationally intensive until we know the exact parametric form of the data distribution (precisely, location-scatter families) (Gelbrich, 1990; Alvarez-Esteban et al., 2016; Cuesta-Albertos et al., 1996; Alvarez-Melis & Fusi, 2020). In absence of such parametric assumptions, we need linear space complexity and super-linear time complexity in terms of observed samples to estimate Wasserstein distances between the underlying data distributions (Cuturi, 2013; Panaretos & Zemel, 2018; Chizat et al., 2020; Rakotomamonjy et al., 2024; Chewi et al., 2024). Additionally, in practice, the stream might arrive from multiple sources, as in Federated Learning (Kairouz et al., 2021).

This has motivated a recent line of work to estimate Wasserstein distance in federated setting (Rakotomamonjy et al., 2024; Li et al., 2024). However, these methods incur communication cost equal to the number of samples in the stream. These gaps in literature motivate us to ask

Can we estimate the Wasserstein distance between two distributions in sublinear space, time, and communication complexity if we obtain a stream of n samples from them?

Total Variation (TV) Distance. TV (Devroye & Gyorfi, 1990) is another well-studied distance between two probability distributions. It measures the maximum gap between the probabilities of any event w.r.t. the two distributions. TV also quantifies the minimum probability that $X \neq Y$ among all couplings of (X, Y) sampled from the distributions. TV distance between the output distributions of an algorithm operated on two neighbouring datasets also central to auditing the privacy level of the algorithm (Koskela & Mohammadi, 2024). Additionally, in bandits, the hardness of a problem depends on the TV distance between the reward distributions of the available actions (Azize & Basu, 2022; Azize et al., 2023). Thus, estimating TV distance from empirical samples emerges as a fundamental problem in privacy and machine learning. Estimating TV distance has been studied extensively (Canonne, 2022). There is a long line of work focusing on estimating TV distance between discrete distributions with finite support (Kamath et al., 2015; Han et al., 2015; Feng et al., 2024; Devroye & Reddad, 2019; Bhattacharyya et al., 2023; 2024; 2025). Feigenbaum et al. (2002); Guha et al. (2006); Roy & Vasudev (2023) propose sublinear algorithm to estimate TV distance while having access to sample streams. But the question of estimating TV distance for continuous distributions with infinite supports remains wide open. This motivates the question:

Can we estimate the TV distance between two continuous distributions with infinite support in sublinear space, time, and communication complexity w.r.t. number of samples n?

We address these questions for any bounded tail distribution by *reducing them into estimation of a sublinear summary, i.e. the frequency of an appropriately chosen subset of the support of a properly discretised distribution.*

Mergeable Sublinear Representation of Distributions. We propose to learn an approximately correct sublinear summary of a distribution from the data stream can address the previous questions (Section 4.1). Because then we can store, communicate, and compute with only these sublinear and discrete summaries of distributions. In literature, we find that the sublinear summaries of histograms has been extensively studied over decades (Misra & Gries, 1982; Alon et al., 1996; Charikar et al., 2002; Cormode & Muthukrishnan, 2005). Additionally, in the continuous distributed monitoring setup (Cormode, 2013), one might want to estimate TV and Wasserstein distances when data streams arrive from multiple sources. Also, with the growth of federated and distributed learning, the question of *how these sublinear-sized summaries of histograms from multiple streams can be merged efficiently* has gained interest (Agarwal et al., 2013; Berinde et al., 2010; Anderson et al., 2017). In machine learning, we work a lot with continuous data distributions. Thus, to be practically useful, we need to extend these sublinear summarisers to continuous setting and find minimal conditions to yield theoretical guarantees. On the other hand, there is another long line of research to learn histograms from samples of a continuous distribution and control the error in this process (Scott, 1979; Freedman & Diaconis, 1981; Ioannidis, 2003; Diakonikolas et al., 2018). But it is an open question to develop algorithms and analysis to join these two streams of research, i.e.

Can we learn sublinear summaries of (possibly continuous) distributions while having a stream of n samples from it?

Our contributions address these questions affirmatively.

1. *Computing Mergeable Sublinear Summary from Multiple Streams.* We propose a generic framework to learn summaries of PDF and CDF of a continuous or discrete (possible with infinite support) distribution using sublinear space (Section 4.2). Given data arriving in a stream (or multiple streams) from an underlying distribution, we propose `SPA` and `SCA` ϵ-approximate corresponding PDF and CDF independent of stream length. To our knowledge, *we initiate the study of mergeable sublinear summaries over streams from an infinite set* and *establish theoretical error bounds for streams from an infinite set and any sub-Gaussian or sub-Weibull distribution.*

2. *Sublinearly Estimating Wasserstein distance.* We use `SCA` to propose `SWA` to PAC-estimate Wasserstein distance (Section 4.3). We first show that *mergeable* sublinear summaries learned by `SCA` are universal estimators of the true distribution in Wasserstein distance. In turn, `SWA` computes a sublinear summary of a distribution that is $\mathcal{O}\left(n^{-1/2}\right)$ close in Wasserstein distance using $\widetilde{\mathcal{O}}\left(\sqrt{n}\right)$ space for n samples[1] `SWA` operates in the distributed/federated setting with a communication cost of $\widetilde{\mathcal{O}}\left(\sqrt{n}\right)$ per round while preserving the estimation guarantee.

3. *Sublinearly Estimating TV distance.* For TV distance, we leverage `SPA` and set the parameters properly to propose `STVA` (Section 4.4). `STVA` maintains mergeable sublinear summaries of the bucketed versions of the true distributions, which have probably infinite buckets in their supports. We show that if we set the bucket width to $\widetilde{\Theta}(n^{-1/3})$, `STVA` yields a $\widetilde{\mathcal{O}}\left(n^{-1/3}\right)$ estimate of TV distance between two distributions using $\widetilde{\mathcal{O}}\left(n^{1/3}\right)$ space.

4. *Applying estimators for fairness and privacy auditing.* In Section 5.1, we demonstrate usefulness of our Wasserstein and TV distances estimators for auditing fairness and privacy of machine learning models, respectively. Experimental results demonstrate the accuracy and sublinearity of the proposed estimators for auditing models trained on real-life datasets from fairness and privacy literatures.

2 Preliminaries

We discuss the fundamentals of distributional distances and the generic algorithmic template to learn mergeable summaries of histograms, which are essential to this work.

Notations. μ, μ_n and $\mu_{b,n}$ denote the true, empirical, and bucketed empirical measure with bucket width b, respectively. $p_\mu : \mathrm{Supp}(\mu) \to [0,1]$ and $Q_\mu : \mathrm{Supp}(\mu) \to [0,1]$ denote the probability density function (PDF) and cumulative density function (CDF) of μ, respectively. $[n]$ refers to $\{1, 2, \ldots, n\}$ for $n \in \mathbb{N}$.

2.1 Distances between Probability Distributions

Now, we define the Wasserstein and TV distances between distributions. We start with defining the Wasserstein distance between distributions defined over metric spaces.

Definition 1 (p-th Wasserstein distance (Villani, 2009))**.** Given two probability measures μ, ν over a metric space $\mathcal{X}$, the p-th Wasserstein distance between them is

$$\mathcal{W}_{\mathrm{p}}(\mu,\nu) \triangleq \left(\min_{\pi\in\Pi(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} \|x-y\|^p \,\mathrm{d}\pi(x,y)\right)^{\frac{1}{p}} .$$

π is a coupling over x and y, and Π is the set of all couplings.

Here, Wasserstein distance is the cost calculated for the optimal coupling between the two probability measures over Euclidean distances. For univariate distributions, Wasserstein distance can be expressed as norms between inverse CDFs of the distributions (Peyré & Cuturi, 2019).

Lemma 2. *For univariate measures μ, ν over $\mathbb{R}$, the p-th Wasserstein distance $\mathcal{W}_{\mathrm{p}}(\mu,\nu)$ is*

$$\mathcal{W}_{\mathrm{p}}(\mu,\nu) = \left\|Q_\mu^{-1} - Q_\nu^{-1}\right\|^p_{L^p([0,1])} , \tag{1}$$

where $Q_\mu^{-1} : [0,1] \to \mathbb{R}\cup\{-\infty\}$ of a probability measure μ is the pseudoinverse function defined as $Q_\mu^{-1}(r) \triangleq \min_{x\in\mathbb{R}\cup\{-\infty\}}\{x : Q_\mu(x) \geq r\}$ for $r \in [0,1]$.

The other distance that we study is the Total Variation (TV) distance.

Definition 3 (Total Variation (TV) Distance)**.** Given measures μ, ν with support $\mathcal{X}$, TV distance between them is

$$\mathrm{TV}(\mu,\nu) \triangleq \sup_{A\subseteq\mathcal{X}} |\mu(A)-\nu(A)| = \frac{1}{2}\|\mu-\nu\|_1 .$$

[1]Soft-O notations, i.e. $\widetilde{\mathcal{O}}(\cdot)$ and $\widetilde{\Theta}(\cdot)$, ignore the polylogarithmic and other lower order terms.

2.2 Mergable Summaries of Histograms

The principal algorithmic technique that we use is that of summarising a histogram, which is a long-studied problem (Misra & Gries, 1982; Alon et al., 1996; Charikar et al., 2002; Agarwal et al., 2013; Anderson et al., 2017). We refer to Cormode & Hadjieleftheriou (2008) for an overview.

First, we formally introduce the problem. We consider data arriving in a stream ζ of length n, where each element belongs to a universe $\mathcal{U}$. The j-th element of $\mathcal{U}$, i.e. $\mathcal{U}_j$, appears f_j times in the stream ζ. Our goal is to maintain an approximate count of elements $\hat{f}_j$ such that $\left\|\hat{f}_j - f_j\right\|_p$ is small for some p. If we separately keep count of all the elements , the problem is trivial and we get $\hat{f}_j = f_j, \forall \mathcal{U}_j \in \mathcal{U}$. However, the goal is to obtain a 'good' approximation while using sublinear, i.e. $o(n)$, space. More generally, the i-th element of the stream can have a weight $\boldsymbol{w}_i \in \mathbb{N}$. Then, the task is to generate estimates $\widehat{\mathcal{F}} = \left\{\hat{f}_j | \mathcal{U}_j \in \mathcal{U}\right\}$, which are close to the true frequency $f_j = \sum_{i=1}^{n} \boldsymbol{w}_i \mathbf{1}[\zeta_i = \mathcal{U}_j]$ for all j.

Now, we extend the problem setup to aggregate from multiple streams (Agarwal et al., 2013). Given S-streams of data $\zeta_1, \zeta_2, ..., \zeta_S$, we aim to generate estimates $\widehat{\mathcal{F}}_1, \widehat{\mathcal{F}}_2, ..., \widehat{\mathcal{F}}_S$ and *combine them efficiently* to output a globally 'good' frequency estimate $\widehat{\mathcal{F}} = \texttt{merge}\left(\widehat{\mathcal{F}}_1, \widehat{\mathcal{F}}_2, ..., \widehat{\mathcal{F}}_S\right)$ for the concatenated stream $\zeta_1 \circ \zeta_2 \circ ... \circ \zeta_S$.

We build on the algorithm of Anderson et al. (2017), which extends the well-known Misra-Gries algorithm (Misra & Gries, 1982). We refer to it as the Mergeable Misra-Gries (`MMG`) algorithm and provide the details in Algorithm 5. This family of algorithms is broadly known as the *counter-based algorithms for histogram summarisation* and consists of *three main modules*. First, they operate by maintaining a sublinear number of counters, say $k = o(n)$, all initiated with 0. Then, the module $\texttt{MMG} \cdot \texttt{Update}$ updates them according to the elements in stream. Specifically, for each stream element, if our algorithm has already assigned a counter, it adds the weight to the corresponding counter. If not, it assigns an empty counter to the element if there is an unassigned counter. Otherwise, the values of all counters are reduced by the median value of counters. Second, if the streams are coming from multiple sources, the module $\texttt{MMG} \cdot \texttt{Merge}$ takes the counters in all the summaries as an input stream, and update each counter in the final summary by considering the counters as stream elements and their counts as the corresponding weights. For updating, it reuses the module $\texttt{MMG} \cdot \texttt{Update}$. Finally, given an element, the module $\texttt{MMG} \cdot \texttt{Estimate}$ returns an estimate of its frequency by returning the value stored in the counter if the element is assigned a counter, and 0 otherwise. Note that `Update`, `Estimate`, and `Merge` takes amortised $\widetilde{\mathcal{O}}(1), \widetilde{\mathcal{O}}(1)$, and $\widetilde{\mathcal{O}}(k)$ time to execute, respectively. For brevity, we defer the further details to Appendix A.

Now, we derive a rectified version of Theorem 2 of Anderson et al. (2017) that bounds the error in frequency estimation using the residuals $n^{Res(k/4)}$ (defined below).

Lemma 4 (Estimation Guarantee of `MMG`)**.** *(a) If `MMG` uses k counters, $\texttt{MMG} \cdot \texttt{Estimate}$ yields a summary $\{\hat{f}_j\}_{\mathcal{U}_j \in \mathcal{U}}$ satisfying $0 \leq f_j - \hat{f}_j \leq \frac{n^{Res(\tau)}}{(k-k^*)-\tau}$, for all $\tau \leq (k - k^*)$ and $\mathcal{U}_j \in |\mathcal{U}|$. Here, $n^{Res(\tau)}$ denotes the sum of frequency of all but τ most frequent items. (b) If we choose $k^* = \frac{k}{2}$ $\tau = k/4$, we get that $0 \leq f_j - \hat{f}_j \leq \frac{4n^{Res(k/4)}}{k}$, $\forall \mathcal{U}_j \in |\mathcal{U}|$.*

We *extend and analyse `MMG`'s design technique, originally developed for the discrete distributions with finite support points, to continuous and discrete distributions with infinite support points.*

3 Problem: Estimating Distances between Distributions from Sample Streams

Now, we formally state our problem setup. Let μ and ν be two measures on $\mathbb{R}$. We consider that the data is arriving in two streams ζ_μ and ζ_ν of size n, and each element of the two streams are independent and identically distributed (i.i.d.) samples of μ and ν, respectively. We denote by $\hat{\mu}_n$ and $\hat{\nu}_n$ the two summaries of μ and ν computed from the sample streams. Given a probability metric $\mathfrak{D}$, our objective is to compute an (ε, δ)-estimate of $\mathfrak{D}(\mu, \nu)$ while using sublinear ($o(n)$) space to store $\hat{\mu}_n$ and $\hat{\nu}_n$. Specifically, we want to

yield $\mathfrak{D}(\hat{\mu}_n, \hat{\nu}_n)$ ensuring

$$\mathbb{P}\left[|\mathfrak{D}(\mu,\nu) - \mathfrak{D}(\hat{\mu}_n, \hat{\nu}_n)| \geq \varepsilon\right] \leq \delta\,, \tag{2}$$

such that $|\text{essSup}(\hat{\mu}_n) \cup \text{essSup}(\hat{\nu}_n)| = o(n)$, and $(\varepsilon, \delta) \in (0,1) \times (0,1)$. In this work, we particularly focus on two probability metrics: Wasserstein distances $\mathfrak{D}(\mu,\nu) = \mathcal{W}_{\text{p}}(\mu,\nu)$, and TV distance $\mathfrak{D}(\mu,\nu) = \text{TV}(\mu,\nu)$. For streams coming from multiple sources, like in the federated (or continuous distributed monitoring) setting, we consider the concatenated streams $\zeta_\mu = \zeta_{\mu,1} \circ \zeta_{\mu,2} \circ \ldots \circ \zeta_{\mu,S}$ and $\zeta_\nu = \zeta_{\nu,1} \circ \zeta_{\nu,2} \circ \ldots \circ \zeta_{\nu,S}$, to be the input streams of length n each to estimate the distance $\mathfrak{D}(\mu,\nu)$.

Structural Assumptions. In this work, we assume that the data generating distributions have bounded tails (Assumption 7). For the case of $\mathfrak{D}(\mu,\nu) = \mathcal{W}_{\text{p}}(\mu,\nu)$, we assume the distribution to have $\ell_{\mathcal{D}}$ bi-Lipschitz CDF (Assumption 8). For the case of $\mathfrak{D}(\mu,\nu) = \text{TV}(\mu,\nu)$, we assume the distribution to have $\ell_{\mathcal{D}}$-Lipschitz PDF (Assumption 9).

To formalise the bounded-tail assumption, we first define the sub-Gaussian and sub-Weibull random variables.

Definition 5 (Sub-Gaussian Distributions)**.** A distribution μ is said to be σ_μ-sub-Gaussian if for any $t \geq 0$, we have

$$\mathop{\mathbb{P}}_{X\sim\mathcal{D}}\left[|X - \mathbb{E}[X]| \geq t\right] \leq 2\exp\left(-\frac{t^2}{\sigma_\mu^2}\right). \tag{3}$$

Correspondingly, the random variable X drawn from μ is a sub-Gaussian random variable (Vershynin, 2018).

Sub-Gaussians cover a wide-range of distributions including any bounded distribution, Gaussians, mixture of Gaussians etc. (Vershynin, 2018). It is standard to assume the noise and data to be sub-Gaussian in regression problems (Wainwright, 2019). Sub-Gaussianity appears in the underlying scoring mechanism for classification (Wang et al., 2018). We also consider the notion of sub-Weibull distribution that generalises the notion of sub-Gaussianity to any form of exponentially bounded tails. While sub-Gaussian distributions are considered to be light-tailed, sub-Weibull distributions are considered for heavy-tail distributions in the concentration of measure literature (Foss et al., 2011; Vladimirova et al., 2020; Bakhshizadeh et al., 2023).

Definition 6 ((τ,α)-Sub-Weibull Distributions)**.** A distribution μ is said to be (τ,α)-sub-Weibull if there exists some constant c_α for any $\tau \geq t \geq 0$, we have

$$\mathop{\Pr}_{X\sim\mu}\left[X \geq t\right] \leq c_\alpha \exp\left(-t^{1/\alpha}\right). \tag{4}$$

When $\tau \to \infty$, (τ,α)-sub-Weibull distribution reduces to the classical definition of sub-Weibull distribution (Vladimirova et al., 2020). For simplicity, we denote (∞,α)-sub-Weibull Distributions as α-sub-Weibull. Note that, for sub-Gaussians, $\alpha = 1/2$. Now, we formally state our assumptions.

Assumption 7 (Sub-Gaussian or sub-Weibull Data Generating Distributions)**.** *We consider the distributions yielding the samples are either σ-sub-Gaussian or α-sub-Weibull.*

The tail bound assumption is required because we have access to only the samples from the stream rather than the true distribution. Thus, we need to have *enough* samples such that the empirical measures available to the algorithm are *close* to the true distributions. The tail bounds provide a control of this concentration of measure phenomenon.

Assumption 8 (Bi-Lipschitz Distributions)**.** *The distribution $\mathcal{D}$ over $\mathbb{R}$ with CDF $Q_{\mathcal{D}}$ has $\ell_{\mathcal{D}}$-bi-Lipschitz CDF, if $\frac{1}{\ell_{\mathcal{D}}}|a-b| \leq |Q_{\mathcal{D}}(a) - Q_{\mathcal{D}}(b)| \leq \ell_{\mathcal{D}}|a-b|\,,\ \forall a,b \in [0,1]$.*

This essentially implies that the PDF of $\mathcal{D}$ is bounded above and below everywhere. Any bounded distribution with continuous and compact support satisfies this, which encompasses most of the common data distributions.

Assumption 9 (Lipschitz PDF Distributions)**.** *The distribution $\mathcal{D}$ with pdf $\mu_{\mathcal{D}}$ over $\mathbb{R}$ has $\ell_{\mathcal{D}}$-Lipschitz PDF, i.e. $|\mu_{\mathcal{D}}(a) - \mu_{\mathcal{D}}(b)| \leq \ell_{\mathcal{D}}|a-b|\,,\ \forall a,b \in [0,1]$.*

Any exponential family distribution with bounded parameters have Lipschitz PDF. It includes Gaussians with bounded mean and variance, exponential and Poisson distributions, and other commonly used distributions.

4 Sublinear Estimators of Wasserstein and TV Distances

In this section, we explain our methodology for mergeable summary creation of bounded-tail distributions, and estimation of Wasserstein and TV distances from sample streams. We theoretically and numerically demonstrate accuracy and sublinearity of our estimators.

4.1 From Sublinear Distance Estimation to Learning Sublinear Summaries

Let us consider the empirical measure μ_n on n points generated from the true (possibly continuous) measure μ. Thus, for any probability distance $\mathfrak{D}$, we observe that

$$|\mathfrak{D}(\mu,\nu) - \mathfrak{D}(\hat{\mu}_n,\hat{\nu}_n)| \leq \underbrace{\mathfrak{D}(\mu,\mu_n) + \mathfrak{D}(\nu,\nu_n)}_{\text{Concentration of measures}} + \underbrace{\mathfrak{D}(\mu_n,\hat{\mu}_n) + \mathfrak{D}(\nu_n,\hat{\nu}_n)}_{\text{Sublinear summaries}},$$

if all of these distances are well-defined. In that case, the distance of the empirical measure and the true one decreases due to concentration of measures. While we learn sublinear summaries for each of the empirical distributions and aim to control the distance between the empirical distributions and the sublinear summary, i.e. $\mathfrak{D}(\mu_n,\hat{\mu}_n)$, for any bounded-tail true distribution μ.

In Section 4.2, we propose algorithms to estimate PDFs and CDFs of bounded-tail distributions. This allows us to control $\mathfrak{D}(\mu_n,\hat{\mu}_n)$ and in turn, compute the desired distances if μ_n has bounded-tails. We first proof that the empirical distribution μ_n corresponding to a sub-Gaussian or sub-Weibull μ is also sub-Gaussian or sub-Weibull, respectively.

Theorem 10 (Empirical Measure is Sub-Gaussian)**.** *Let the true distribution μ be a σ_μ-sub-Gaussian with mean 0. If $n \geq \frac{c\log(\frac{1}{\delta})}{\varepsilon^2}$, the empirical measure μ_n generated by $X_1, X_2, ..., X_n$ drawn i.i.d. from μ is $(1+\varepsilon)\sigma_\mu$-sub-Gaussian with probability $1-\delta$, for any $\varepsilon, \delta \in (0,1)$ and a constant $c \geq 1$.*

Theorem 11 (Empirical Distribution is sub-Weibull)**.** *Given an α-sub-Weibull true measure μ, the empirical measure μ_n generated by $X_1, X_2, ..., X_n$ drawn i.i.d. from μ is (τ,α)-sub-Weibull with probability at least $1-\delta$ given $n \geq \frac{\exp(\sqrt[\alpha]{\tau})}{12}\log\frac{1}{\delta}$.*

We show that with high probability, the empirical distribution retains sub-Gaussianity and sub-Weibull property of the true measure μ. This result is important for learning sublinear summaries of true distribution from sample streams. The proof is in Appendix B.

4.2 Learning Mergable Sublinear Summaries

In this section, we state our results regarding sublinear summary approximation of sub-Gaussian and sub-Weibull distributions satisfying Assumption 7. We use `MMG` to learn an approximation of the PDF and CDF of distributions.

A. Learning PDF. We first turn the support of the empirical distribution into a collection of buckets of width $b > 0$, and in turn, construct an empirical bucketed distribution $\mu_{b,n}$ out of the empirical distribution μ_n.

Definition 12 (**Bucketed Empirical Distribution**)**.** Let $\mathcal{D}$ be a distribution supported on a (finite or infinite) closed interval $\mathcal{B} \subseteq \mathbb{R}$ with (discrete or continuous) measure μ. Given a reference point $x_0 \in \mathcal{B}$, bucket width b, and an index set $I \subseteq \mathbb{Z}$, we can represent $\mathcal{B} = \mathcal{B}(x_0, I, b) \triangleq \cup_{i\in I}[x_0+ib, x_0+(i+1)b]$. We define the bucketed empirical distribution, denoted $\mu_{b,n}$, to be the empirical distribution generated from n samples of μ as $\mu_{b,n}(i) \triangleq \frac{n_i}{n}$, where n_i is the number of samples falling in the bucket $\sqcup_i$.

For brevity, we denote the bucket $[x_0+ib, x_0+(i+1)b]$ by $\sqcup_i$. We note that $\mathcal{B}$ inherits the metric structure of $\mathbb{R}$. Specifically, the distance between two points $x_1 \in \sqcup_i$ and $x_2 \in \sqcup_j$ is defined as $d(x_1,x_2) = |i-j|b$.

In Algorithm 1, we first fix a number of buckets k, and treat each of the buckets as an element of the stream. The weight of a bucket is the number of elements in the stream that falls in it, which is n_i for $\sqcup_i$. We apply

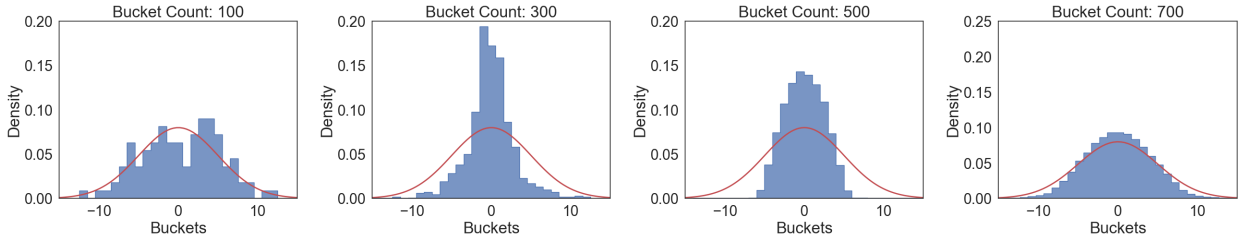

Figure 1: Evolution of summaries constructed using 10^5 samples from $\mathcal{N}(0,5)$. We set $b = 0.05$ and bucket numbers $100, 300, 500, 700$.

`MMG` over these buckets to return the estimated frequency $\hat{f}_i$ for each bucket $\sqcup_i$. Finally, we return the learned PDF $\{p_{\mu_{b,n}}(i)\}_{i\in I} = \left\{\frac{\hat{f}_i}{\max_j \hat{F}_j}\right\}_{i\in I}$ as a summary of μ_n's PDF. Now, we bound the learning error of `SPA`.

Algorithm 1 Sublinear PDF Approximator: $\texttt{SPA}(\zeta \overset{i.i.d.}{\leftarrow} \mu, k, b)$

```
Initialize
MMG · Initialize(k)
Update(ζ_i)
⊔_{ζ_i} ← ⊔ that contains ζ_i
MMG · Update(ζ_i, 1)
Merge(T_2)
MMG · Merge(T_2)
Estimate(ζ_i)
f̂_i ← MMG · Estimate(ζ_i)
return p_{μ̂_n}(i) ← f̂_i / max_j F̂_j
```

Algorithm 2 Sublinear CDF Approximator $\texttt{SCA}(\zeta \overset{i.i.d.}{\leftarrow} \mu, \varepsilon, k, b)$

```
Initialize
MMG · Initialize(k)
Update(ζ_i)
⊔_{ζ_i} ← ⊔ that contains ζ_i
MMG · Update(⊔_{ζ_i}, 1)
Merge(T_2)
MMG · Merge(T_2)
Estimate(ζ_i)
F̂_j ← MMG · EstimateCumulate(ζ_i)
return Q̂_i ← F̂_i / max_j F̂_j
```

Theorem 13 (`SPA` Learning Error)**.** *(a) If `SPA` uses k buckets and outputs $\hat{\mu}_n$, then for all $i \in I$,*

$$\left|p_{\hat{\mu}_n}(i) - p_{\mu_{b,n}}(i)\right| \leq \max\left\{\frac{4n^{Res(k/4)}}{n}\frac{f_i}{n}, \frac{f_i - \hat{f}_i}{n}\right\}.$$

(b) Further, if μ corresponding to μ_n is σ_μ-sub-Gaussian, $|\zeta| = n \geq c\log\left(\frac{1}{\delta}\right)$, and $k \geq \left\lceil \frac{8\sigma_\mu}{b}\sqrt{\log\left(\frac{1}{\varepsilon}\right)}\right\rceil$, then for all $i \in I$, $\mathbb{P}(\left|p_{\hat{\mu}_n}(i) - p_{\mu_{b,n}}(i)\right| \geq \varepsilon) \leq \delta$.

(c) Further, if μ corresponding to μ_n is α-SubWeibull, $|\zeta| = n \geq \frac{c}{\varepsilon}\log\left(\frac{1}{\delta}\right)$, and $k \geq \left\lceil \frac{c}{b}\left(\log\left(\frac{1}{\varepsilon}\right)\right)^\alpha\right\rceil$, then for all $i \in I$, $\mathbb{P}(\left|p_{\hat{\mu}_n}(i) - p_{\mu_{b,n}}(i)\right| \geq \varepsilon) \leq \delta$.

Lemma 4 shows that the error in `MMG` depends on the sum of frequency of elements except the ones with the highest frequencies. `SPA` leverages this observation, the boundedness of the tails of a distribution (Assumption 7), and the guarantee on the preservation of tail bounds for the empirical measure (Theorem 10 and 11). In `SPA`, we set bucket width b as $k \geq \tau_k(\varepsilon, b)$ such that the (possibly) uncovered tails have sufficiently small mass. This yields an (ϵ, δ)-PAC estimate of the empirical PDF. The proof is in Appendix C.

B. Learning CDF. Now, we want to learn a sublinear summary of the CDF of the empirical distribution μ_n. Thus, we start by bucketing its support with k counters. Given the corresponding PDF $p_{\mu_{b,n}}$, we define the CDF of the bucketed empirical distribution as $Q_{\mu_{b,n}}(i) \triangleq \sum_{j\leq i} p_{\mu_{b,n}}(j)$.

In general, if there is a *total order* on the stream universe $\mathcal{U}$, we can define the cumulant function $F : \mathcal{U} \to \mathbb{N}$ as $F_i \triangleq \sum_{j\leq i} f_j$. Total order exists for the real line and the bucket $\mathcal{B}$ defined on it. Lemma 14 establishes the error bound of $\texttt{MMG} \cdot \texttt{EstimateCumulate}$ for the cumulant F. The proof is in Appendix A.

Lemma 14 (Error Bound of $\texttt{MMG} \cdot \texttt{EstimateCumulate}$)**.** *If `MMG` uses k counters, $\texttt{MMG} \cdot \texttt{EstimateCumulate}$ yields a summary $\{\hat{F}_j\}_{\mathcal{U}_j \in \mathcal{U}}$ satisfying $0 \leq F_i - \hat{F}_i \leq 2n^{Res(\frac{k}{4})}$.*

Now, in the spirit of Theorem 13, we appropriately set the bucket width and collect enough samples to yield an (ϵ, δ)-PAC estimate of the empirical CDF (Appendix C).

Theorem 15 (SCA Learning Error)**.** *Given stream length $|\zeta| = n \geq \tau_n(\varepsilon, \delta)$ generated from a bounded-tail distribution μ, SCA with bucket width b sets #buckets to $k \geq \tau_k(\varepsilon, b)$ and outputs $\widehat{Q}$, such that with probability at least $1-\delta$, $\left|Q_{\mu_{b,n}}(i) - \widehat{Q}(i)\right| \leq \varepsilon$ for all $i \in I$.*

Tail Condition	$\tau_n(\varepsilon, \delta)$	$\tau_k(\varepsilon, b)$
σ_μ-sub-Gaussian	$c \log\left(\frac{1}{\delta}\right)$	$\left\lceil \frac{8\sigma_\mu}{b}\sqrt{\log\left(\frac{4}{\varepsilon}\right)} \right\rceil$
α-sub-Weibull	$\frac{c}{\varepsilon}\log\left(\frac{1}{\delta}\right)$	$\left\lceil \frac{c}{2b}\left(\log\left(\frac{4}{\varepsilon}\right)\right)^{\alpha} \right\rceil$

Note that the space complexities of both SPA and SCA are independent of streaming universe's size.

Numerical Demonstration. We take a stream of 10^5 samples from a Gaussian $\mathcal{N}(0, 5)$ and estimate its PDF with SPA. We set the bucket width to $b = 0.05$ and vary the number of buckets from 100 to 700. Figure 1 validates that increasing the number of buckets, while still being sublinear in #samples, in SPA leads to more accurate summaries of a continuous sub-Gaussian density function.

4.3 Estimating Wasserstein Distance

In this section, we introduce an algorithm, namely SWA, to estimate Wasserstein distance between distributions satisfying Assumption 7 and 8. SWA leverages SCA to learn the CDF and use the inverse CDF formulation of Wasserstein (Lemma 2) to obtain a good estimation of the Wasserstein distance between two distributions. We present the pseudocode of SWA in Algorithm 3.

Algorithm 3 Sublinear Wasserstein Estimator: $\texttt{SWA}(\zeta_\mu, \zeta_\nu \overset{i.i.d.}{\leftarrow} \mu, \nu, \varepsilon, p)$

Require: $\ell_\mu, \ell_\nu \leftarrow$ Lipschitz constant of μ and ν, $p \leftarrow p$-th Wasserstein
1: **for** $\mathcal{D} \in \{\mu, \nu\}$ **do**
2: $\quad b_{\mathcal{D}} \leftarrow \frac{\epsilon}{4}$
3: $\quad \{Q_{\hat{\mathcal{D}}_n}(i)\}_{i=1}^k \leftarrow \texttt{SCA}(\zeta_{\mathcal{D}}, b_{\mathcal{D}}, \tau_k(\varepsilon, \ell_{\mathcal{D}}, \delta), \varepsilon/2)$
4: **end for**
5: **return** $\mathcal{W}_{\mathrm{p}}(\hat{\mu}_n, \hat{\nu}_n) \leftarrow$ plug $\{Q_{\hat{\mu}_n}(i)\}_{i=1}^k$ and $\{Q_{\hat{\nu}_n}(i)\}_{i=1}^k$ in Equation (1)

Universal Learning Guarantee of SCA. We show that SCA is a universally good estimator of the CDF of any bounded-tail measure with respect to the Wasserstein distance. *This is a culmination of three results.*

First, SCA guarantees an (ϵ, δ)-PAC summary of the CDF of a bucketed empirical distribution. To leverage that guarantee in the case of Wasserstein distance for a (possibly continuous) measure μ, we need to choose the bucket width b appropriately such that $\mathcal{W}_{\mathrm{p}}(\mu_n, \mu_{b,n})$ is small as the distance of the true and the bucketed distribution is bounded by b (Lemma 33).

Second, once we have the (ϵ, δ)-PAC summary of the CDF, the rest is to show that it yields small error in computing the inverse CDF, and in turn, the Wasserstein distance (Equation (1)). We establish a guarantee on the approximation error of the pseudoinverse of a bi-Lipschitz CDF, given a guarantee on the approximation error on the CDF (Corollary 35).

Finally, we show that the empirical measure μ_n corresponding to a true measure μ with bi-Lipschitz CDF also exhibits bi-Lipschitz CDF with high probability.

Theorem 16 (Bucketed Empirical Measure is Bi-Lipschitz)**.** *Let the bucketed empirical measure $\mu_{b,n}$ generated by n i.i.d. samples coming from a distribution μ with ℓ_μ bi-Lipschitz CDF. For any $\varepsilon, \delta \in (0,1)$ and fixed constant $c > 0$, if $n \geq \frac{c}{\varepsilon^2 b^2}\log\left(\frac{2}{\delta}\right)\max\left\{\frac{1}{\ell_\mu^2}, \ell_\mu^2\right\}$, $\mu_{b,n}$ is $(1+\varepsilon)\ell_\mu$ bi-Lipschitz with probability $1-\delta$.*

Proof of Theorem 16 is in Appendix B. Finally, these results together yield Theorem 17, i.e. SCA learns a universal estimator of any measure μ. The proof is in Appendix D.

Theorem 17 (Universal Learning of SCA in $\mathcal{W}_{\mathrm{p}}(,)$)**.** *Given $|\zeta| = n \geq \tau_n(\ell_\mu, \varepsilon, \delta)$ generated from a bounded-tail distribution μ and $\varepsilon, \delta \in (0,1)$, if we set the bucket width $b = \varepsilon/2$, then SCA uses $\tau_k(\ell_\mu, \varepsilon)$ buckets (i.e. space) and outputs $\hat{\mu}_n$, such that $\mathbb{P}(\mathcal{W}_{\mathrm{p}}(\mu_n, \hat{\mu}_n) \geq \varepsilon) \leq \delta$. We omit the constants in $\tau_n(\ell_\mu, \varepsilon, \delta)$ and $\tau_k(\ell_\mu, \varepsilon)$ for simplicity.*

Tail Condition	$\tau_n(\ell_\mu, \varepsilon, \delta)$	$\tau_k(\ell_\mu, \varepsilon)$
σ_μ-sub-Gaussian	$\log(1/\delta) \max\left\{\frac{1}{\ell_\mu^2}, \ell_\mu^2\right\}$	$\frac{\sigma_\mu}{\varepsilon} \log\left(\frac{\ell_\mu}{\varepsilon}\right)$
α-sub-Weibull	$\log(1/\delta) \max\left\{\frac{1}{\varepsilon}, \frac{1}{\ell_\mu^2}, \ell_\mu^2\right\}$	$\frac{1}{\varepsilon}\left(\log\left(\frac{\ell_\mu}{\varepsilon}\right)\right)^\alpha$

PAC Gurantee of SWA. We know (Equation (5)) that controlling the distance between the empirical measure and the sublinear summary is enough to control the distance estimation error if the empirical distribution concentrates to the true one. Lemma 44 of Bhat & L.A. (2019) proposes the concentration guarantees of empirical measure in Wasserstein distance. Combining this result and Theorem 17 yields estimation guarantees of $\hat{\mu}_n$ w.r.t. true measure μ.

Theorem 18 (PAC Guarantee of SWA)**.** *Given $|\zeta| = n \geq \tau_n(\ell_\mu, \varepsilon, \delta)$ generated from a bounded-tail distribution μ, with bucket width $b = \varepsilon/2$, then SCA uses $\tau_k(\ell_\mu, \varepsilon)$ space and outputs $\hat{\mu}_n$ such that $\mathbb{P}(\mathcal{W}_1(\mu, \hat{\mu}_n) \leq \varepsilon) \geq 1 - \delta$. This further shows that for a stream of size $n = \mathcal{O}\left(\epsilon^{-2}\log(1/\delta)\right)$ from two distributions μ and ν, SWA yields*

$$|\,\mathcal{W}_1(\hat{\mu}_n, \hat{\nu}_n) - \mathcal{W}_1(\mu, \nu)\,| \leq 4\varepsilon \tag{5}$$

with probability $1 - 2\delta$. Here, $\epsilon \in (0, 1/4]$ and $\delta \in (0, 1/2]$. Here, $\sigma \triangleq \max\{\sigma_\mu, \sigma_\nu\}$, $\ell \triangleq \max\{\ell_\mu, \ell_\nu\}$. We omit the constants in $\tau_n(\ell_\mu, \varepsilon, \delta)$ and $\tau_k(\ell_\mu, \varepsilon)$ for simplicity.

Tail Condition	$\tau_n(\ell_\mu, \varepsilon, \delta)$	$\tau_k(\ell_\mu, \varepsilon)$
σ-sub-Gaussian	$\log(1/\delta) \max\left\{\frac{1}{\varepsilon^2}, \frac{1}{\ell^2}, \ell^2\right\}$	$\frac{\sigma}{\varepsilon} \log\left(\frac{\ell}{\varepsilon}\right)$
α-sub-Weibull	$\log(1/\delta) \max\left\{\frac{1}{\varepsilon^2}, \left(\log\frac{1}{\delta}\right)^{2\alpha-1}, \frac{1}{\ell^2}, \ell^2\right\}$	$\frac{1}{\varepsilon}\left(\log\left(\frac{\ell}{\varepsilon}\right)\right)^\alpha$

Remark 19 (Space, Time, and Communication Complexity of SWA.). SWA requires $\widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon}\right)$ time, space, and communication cost to learn a distribution in Wasserstein distance. Observe that the lower bound on n in Theorem 18 ensures convergence μ_n to μ in Wasserstein distance. This is the only term that has δ-dependence as the rest of the algorithm is deterministic. SWA's time, space and communication complexities are sublinear ($\widetilde{\Theta}(\sqrt{n})$) compared to the number of samples required to ensure convergence of the empirical measure while retaining the same error.

4.4 Estimating TV Distance

In this section, we introduce STVA that yields an estimate of TV distance between two bounded-tail distributions with Lipschitz PDFs. STVA leverages SPA to sublinearly learn the PDFs of the distributions, and then, estimates the TV distance as L_1 norm between them (Definition 3).

PAC Guarantee of STVA. *We establish the estimation guarantee of STVA in three steps.* First, we prove that if we choose the bucket width properly to discretise the support of μ, we obtain a control over the TV distance between true distribution and its bucketed version. Second, we leverage the concentration inequalities of Berend & Kontorovich (2013) to control the TV distance between the bucketed distribution and the empirical bucketed distribution from the streamed samples. Finally, we establish the approximation guarantee of SPA to control the TV distance between the empirical bucketed distribution and its sublinear summary.

Step 1: Setting the Bucket Width. Standard rules (Scott, 1979; Freedman & Diaconis, 1981) of choosing optimal bucket width of a histogram depends either on rule of thumbs or crucially depends on parameters

Algorithm 4 Sublinear TV Distance Estimator: $\texttt{STVA}(\zeta_\mu, \zeta_\nu \overset{i.i.d.}{\leftarrow} \mu, \nu, \varepsilon, b)$

Require: $\sigma_\mu, \sigma_\nu \leftarrow$ Subgaussian parameter of μ and ν, $b \leftarrow$ bucket width
1: **for** $\mathcal{D} \in \{\mu, \nu\}$ **do**
2: $\quad k_{\mathcal{D}} \leftarrow \tau_k(\varepsilon, \ell, \delta)$
3: $\quad b_{\mathcal{D}} \leftarrow \tau_b(\varepsilon, \ell)$
4: $\quad \{p_{\hat{\mathcal{D}}_n}(i)\}_{i=1}^k = \{\hat{p}_n(i)\}_{i=1}^k \leftarrow \texttt{SPA}(\mathcal{D}, k_{\mathcal{D}}, b_{\mathcal{D}})$
5: **end for**
6: **return** $\mathrm{TV}(\hat{\mu}_n, \hat{\nu}_n) \leftarrow 0.5 \sum_i \left| p_{\hat{\mu}_n(i)} - p_{\hat{\nu}_n(i)} \right| \mathbf{1}\left(\max\left\{p_{\hat{\mu}_n(i)}, p_{\hat{\nu}_n(i)} > 0\right\}\right)$

such as sample size n, $\int_{-\infty}^{\infty} p'(x)^2\, dx$ etc., which are difficult to estimate for an unknown distribution arriving in a stream. We propose a bucketing technique depending on the Lipschitz constant of the true PDF and the sub-Gaussian parameter of the true distribution to tune the bucket width a priori. Proof is in Appendix E

Theorem 20 (Bucket Width for TV Distance Estimation)**.** *Given a bounded-tail distribution μ with ℓ_μ-Lipschitz PDF and a corresponding bucketed measure μ_b, if we fix the bucket width $b = \tau_b(\varepsilon, \ell)$, we have* $\mathrm{TV}(\mu, \mu_b) \leq \epsilon$.

Tail Condition	σ_μ-sub-Gaussian	α-sub-Weibull
$\tau_b(\varepsilon, \ell_\mu)$	$\frac{\varepsilon}{\sigma_\mu \ell_\mu \sqrt{\log(2/\varepsilon)}}$	$\frac{\varepsilon}{\ell_\mu (\log(2c_\alpha/\varepsilon))^\alpha}$

Step 2: Concentration of Bucketed Empirical Distribution in TV. For continuous measures, assumption on the structure of the distribution is necessary to establish meaningful convergence rates in TV distance (Diakonikolas, 2016). We establish a convergence result for bucketed measure μ_b generated from a bounded-tail measure μ that possibly has infinite buckets in its support (Appendix E).

Lemma 21 (Concentration in TV over Infinite Buckets)**.** *Let μ_n be an empirical measure generated from a discrete bucketed measure μ_b corresponding to a bounded-tail distribution μ, and, $\varepsilon, \delta \in (0,1)$. Then, for $n \geq \tau_n(\varepsilon, b, \delta)$, $\mathbb{P}[\mathrm{TV}(\mu, \mu_n) \geq \varepsilon] \leq \delta$. Here, $\Gamma(\cdot)$ denotes the Gamma function (Davis, 1959).*

Tail Condition	σ_μ-sub-Gaussian	α-sub-Weibull
$\tau_n(\varepsilon, b, \delta)$	$c\varepsilon^{-2} \max\left\{\frac{4\sigma_\mu\sqrt{\pi}}{b}, \log(1/\delta)\right\}$	$c\varepsilon^{-2} \max\left\{\frac{2c_\alpha}{b}\Gamma(1+\alpha), \log(1/\delta)\right\}$

Step 3: Approximation Guarantee of SPA. Theorem 13 shows that SPA learns a pointwise approximation to the PDF of any bucketed empirical measure. However, to extend this guarantee to the TV distance, we need to show that the the sum of errors over the entire support is small. Note that bounded-tail distributions have small mass in the tails (i.e. in most of the buckets). Hence, it suffices to control the error in learning PDF of the 'heavier' buckets to control the error in estimating the TV distance. The proof is in Appendix E.

Theorem 22 (Learning Error of SPA)**.** *If SPA accesses a stream of length $|\zeta| = n \geq \tau_n(\varepsilon, \delta)$ from a bounded-tail distribution, and uses $\tau_k(\varepsilon, \delta)$ buckets to output a sublinear summary $p_{\hat{\mu}_n}$, then with probability $1-\delta$,* $\mathrm{TV}(\hat{\mu}_n, \mu_{b,n}) \leq \left\| p_{\hat{\mu}_n} - p_{\mu_{b,n}} \right\|_1 \leq \varepsilon$.

Tail Condition	$\tau_n(\varepsilon, \delta)$	$\tau_k(\varepsilon, b)$
σ_μ-sub-Gaussian	$c \log\left(\frac{1}{\delta}\right)$	$\left\lceil \frac{8\sigma_\mu}{b} \sqrt{\log\left(\frac{6}{\varepsilon}\right)} \right\rceil$
α-sub-Weibull	$\frac{c}{\varepsilon} \log \frac{1}{\delta}$	$\left\lceil \frac{c}{b} \left(\log\left(\frac{6}{\varepsilon}\right)\right)^\alpha \right\rceil$

These results together yield the PAC guarantee for STVA.

Theorem 23 (PAC Guarantee of STVA)**.** *Given $\epsilon \in (0, 1/6]$, $\delta \in (0, 1/4]$, a stream of size $n \geq \tau_n(\varepsilon, \ell, \delta)$ from two bounded-tail and ℓ-Lipschitz distributions μ and ν, and bucket width $b = \tau_b(\varepsilon, \ell, \delta)$, STVA uses*

$\tau_k(\varepsilon, \ell, \delta)$ *space and*

$$\mathbb{P}\left(|\,\mathrm{TV}\left(\hat{\mu}_n, \hat{\nu}_n\right) - \mathrm{TV}\left(\mu, \nu\right)|\geq 6\varepsilon\right) \leq 4\delta\,. \tag{6}$$

Here, $\sigma \triangleq \max\{\sigma_\mu, \sigma_\nu\}$, $\ell \triangleq \max\{\ell_\mu, \ell_\nu\}$. *We omit the constants in* $\tau_n(\ell_\mu, \varepsilon, \delta)$ *and* $\tau_k(\ell_\mu, \varepsilon)$ *for simplicity.*

Tail Condition	$\tau_n(\varepsilon, \ell, \delta)$	$\tau_k(\varepsilon, \ell)$	$\tau_b(\varepsilon, \ell)$
σ-sub-Gaussian	$\varepsilon^{-2}\max\left\{\frac{\sigma^2\ell\log(1/\varepsilon)}{\varepsilon}, \log\left(\frac{1}{\delta}\right)\right\}$	$\frac{\sigma^2\ell}{\varepsilon}\log\left(\frac{1}{\varepsilon}\right)$	$\frac{\varepsilon}{\sigma\ell\sqrt{\log(2/\varepsilon)}}$
α-sub-Weibull	$\varepsilon^{-2}\max\left\{\frac{\ell(\log(1/\varepsilon))^\alpha}{\varepsilon}\Gamma(1+\alpha), \log\left(\frac{1}{\delta}\right)\right\}$	$\frac{\ell}{\varepsilon}\left(\log\left(\frac{1}{\varepsilon}\right)\right)^{2\alpha}$	$\frac{\varepsilon}{\ell(\log(2c_\alpha/\varepsilon))^\alpha}$

Remark 24 (Space, Time, Communication Complexity of `STVA`). `STVA` requires $\widetilde{\mathcal{O}}\left(\frac{1}{\varepsilon}\right)$ time, space, and communication complexity per round to estimate the TV distance between two bounded-tail distributions[2]. Observe that the lower bound on n ensures convergence of $\hat{\mu}_n$ to μ in TV distance (Theorem 23). This is the only term with δ-dependence as the rest of the algorithm is deterministic. Thus, `STVA` achieves $\widetilde{\Theta}(n^{1/3})$ time, space, and communication complexity with respect to #samples required for convergence of empirical measure while retaining the order of error.

Note that the bounds in Theorem 23 ensures that ε is of the order $n^{-1/3}$, which ensures that the bucket width is of the order $n^{-1/3}$ and the error in TV distance is of the order $n^{-1/3}$. Our result is consistent with those of other standard histogram bucket width rules, which chooses bucket width of order $n^{-1/3}$ and ensures the integrated mean squared error to be of order $n^{2/3}$ (Scott, 1979; Freedman & Diaconis, 1981).

Remark 25 (Further Implications of Theorem 22). We observe that Theorem 22 does not only yield a guarantee on TV distance estimation error but also have two further implications, which might be of general interest.

1. ℓ_1*-estimation of True Frequencies:* A closer look into the result shows that `SPA` can also yield an ℓ_1 approximation of the true frequencies ($\{f_i\}_i$) corresponding to any probability distribution when the tails are sufficiently bounded.

2. *Estimation of any Integral Probability Metric (IPM):* As our algorithms yield pointwise estimates of both the PDF and CDF of a distribution, these methods can be used to estimate any Integral Probability Metric (IPM), such as Dudley metric, Maximum Mean Discrepancy (MMD), Kolmogorov distance (Sriperumbudur et al., 2012).

4.5 Tightness of Space Complexity

We compare the upper bounds on the space complexities of `SPA` and `SCA` to the existing lower bounds for finite universe of streams. Frequency estimation have been extensively studied in this setting. Theorem 26 provides a $\Omega\left(\varepsilon^{-1}\right)$ lower bound on space complexity of any non-trivial solution (i.e. not storing the entire stream, or counts of all elements in the universe) to the frequency estimation problem with respect to the ℓ_∞-norm.

Theorem 26 (Lower Bound of Frequency Estimation in ℓ_∞-norm (Chakrabarti, 2024))**.** *Given a stream* ζ *of length* n *coming from a universe* $\mathcal{U}$ *of size* m*, any algorithm that produces a* (εn)-ℓ_∞ *approximation of the frequency vector* $\{f_i\}_{i\in[m]}$ *as* $\left\{\hat{f}_i\right\}_{i\in[m]}$ *must use* $\Omega\left(\min\left(n, m, \varepsilon^{-1}\right)\right)$ *space.*

Our framework provides a $\widetilde{\mathcal{O}}\left(\varepsilon^{-1}\right)$ approximation of the true frequencies in ℓ_1-norm for any bounded-tail distribution with both infinite and finite supports, i.e. universe of streams (ref. Remark 25)[3]. Hence, Theorem 26 indicates that *our algorithms are tight in terms of space complexity.*

[2]Note that for sub-Gaussians $\alpha = \frac{1}{2}$.

[3]Note that while our techniques consider the data arrives from a distribution rather than arbitrarily from a universe (as in Theorem 26), it suffices if the underlying data are generated from a distribution and the order of their arrivals is arbitrary.

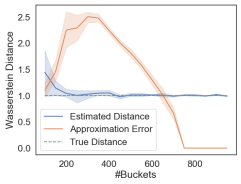

Figure 2: Performance of SWA with $\mathcal{N}(0,5)$, $\mathcal{N}(1,5)$ and $b = 0.05$

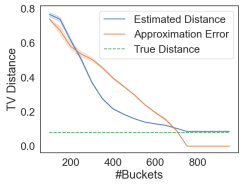

Figure 3: Performance of STVA with $\mathcal{N}(0,5)$, $\mathcal{N}(1,5)$ and $b = 0.05$

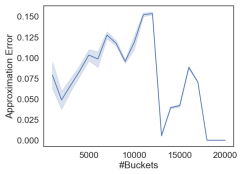

Figure 4: Auditing with SWA on regression output of ACS_Income .

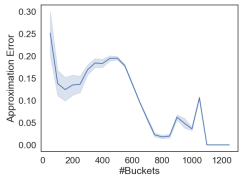

Figure 5: Auditing with SWA on classification output of ACS_Income .

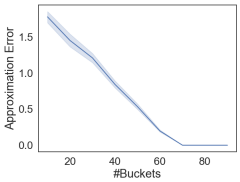

Figure 6: Privacy Auditing of logistic regression on MNIST.

5 Experimental Analysis

Our aim is to understand *whether* ***SWA*** *and* ***STVA*** *yield PAC estimates of Wasserstein and TV distances between sub-Gaussian distributions with sublinear space, i.e. number of buckets, with respect to the number of samples.*

Setup. We compute distances between two Gaussian distributions $\mathcal{N}(0,5)$ and $\mathcal{N}(1,5)$, where 10^5 samples from each of them arrive through $S = 10$ sources. True TV and Wasserstein distances between them are 0.0797 and 1.00, respectively. We set the bucket width to $b = 0.05$ and increase the number of buckets as $\{100, 150, \ldots, 1000\}$. We run each of the experiments 50 times. Finally, for SWA, we report the estimated Wasserstein distance and the learning error of SCA, i.e. $\mathcal{W}_{\mathrm{p}}(\mu_n, \hat{\mu}_n)$, in Figure 2. Similarly, for STVA, we report the estimated TV distance and the learning error of SPA, i.e. $\mathrm{TV}(\mu_n, \hat{\mu}_n)$, in Figure 3. We report the true distances in each case as the dotted horizontal lines.

Results.

1. **Accuracy and Sublinearity of SWA.** The estimation error of SWA and learning error becomes negligible as the bucket number reaches to 750 while using 10^5 samples. This is better than the theoretical upper limit of buckets, i.e. 1040, suggested by our results.

2. **Accuracy and Sublinearity of STVA.** The estimation error of STVA and learning error becomes negligible as #buckets also reaches to 750. Thus, *the results demonstrate that both* ***STVA*** *and* ***SWA*** *yield accurate estimates of TV and Wasserstein distances while using only sublinear space w.r.t. #samples.*

5.1 Applications: Auditing Fairness and Privacy

Auditing fairness (Ghosh et al., 2021; 2022; Yan & Zhang, 2022) and privacy (Nasr et al., 2023; Steinke et al., 2024; Koskela & Mohammadi, 2024; Annamalai & Cristofaro, 2024) of Machine Learning (ML) models is an important and increasingly studied question for developing trustworthy ML. Here, we deploy SWA and STVA to estimate the fairness and privacy of ML models trained on real-world data, respectively. The details are provided in Appendix G.

6 Discussions and Future Works

We propose an algorithmic framework to learn *mergeable* and *sublinear* summaries of discrete and continuous, sub-Gaussian and sub-Weibull distributions while data streams arrive from a single or multiple sources. We show that the computed mergeable sublinear summarisers are universal estimators with respect to Wasserstein distance. We establish the first sublinear time, space, and communication algorithms to estimate TV and Wasserstein distances over continuous and discrete (with infinite support) distributions. We also note that our framework can be used to estimate any ℓ_p-norm based distance between distributions with sublinear resources. However, our work is constrained to distributions over scalars. A future direction is to study these problems for multi-variate distributions.

References

Pankaj K. Agarwal, Graham Cormode, Zengfeng Huang, Jeff M. Phillips, Zhewei Wei, and Ke Yi. Mergeable summaries. *ACM Trans. Database Syst.*, 38(4), dec 2013. ISSN 0362-5915. doi: 10.1145/2500128. URL https://doi.org/10.1145/2500128.

Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the Twenty-Eighth Annual ACM Symposium on Theory of Computing*, STOC '96, pp. 20–29, New York, NY, USA, 1996. Association for Computing Machinery. ISBN 0897917855. doi: 10.1145/237814.237823. URL https://doi.org/10.1145/237814.237823.

Pedro C. Alvarez-Esteban, E. del Barrio, J.A. Cuesta-Albertos, and C. Matran. A fixed-point approach to barycenters in wasserstein space. *Journal of Mathematical Analysis and Applications*, 441(2):744–762, 2016. ISSN 0022-247X. doi: https://doi.org/10.1016/j.jmaa.2016.04.045. URL https://www.sciencedirect.com/science/article/pii/S0022247X16300907.

David Alvarez-Melis and Nicolo Fusi. Geometric dataset distances via optimal transport. *Advances in Neural Information Processing Systems*, 33:21428–21439, 2020.

Daniel Anderson, Pryce Bevan, Kevin Lang, Edo Liberty, Lee Rhodes, and Justin Thaler. A high-performance algorithm for identifying frequent items in data streams. In *Proceedings of the 2017 Internet Measurement Conference*, IMC '17, pp. 268–282, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450351188. doi: 10.1145/3131365.3131407. URL https://doi.org/10.1145/3131365.3131407.

Meenatchi Sundaram Muthu Selva Annamalai and Emiliano De Cristofaro. Nearly tight black-box auditing of differentially private machine learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=cCDMXXiamP.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Achraf Azize and Debabrota Basu. When privacy meets partial information: A refined analysis of differentially private bandits. *Advances in Neural Information Processing Systems*, 35:32199–32210, 2022.

Achraf Azize, Marc Jourdan, Aymen Al Marjani, and Debabrota Basu. On the complexity of differentially private best-arm identification with fixed confidence. *Advances in Neural Information Processing Systems*, 2023.

Milad Bakhshizadeh, Arian Maleki, and Victor H De La Pena. Sharp concentration results for heavy-tailed distributions. *Information and Inference: A Journal of the IMA*, 12(3):1655–1685, 2023.

Borja Balle, Gilles Barthe, and Marco Gaboardi. Privacy amplification by subsampling: tight analyses via couplings and divergences. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pp. 6280–6290, Red Hook, NY, USA, 2018. Curran Associates Inc.

Daniel Berend and Aryeh Kontorovich. A sharp estimate of the binomial mean absolute deviation with applications. *Statistics and Probability Letters*, 83(4):1254–1259, 2013. ISSN 0167-7152. doi: https://doi.org/10.1016/j.spl.2013.01.023. URL https://www.sciencedirect.com/science/article/pii/S0167715213000242.

Radu Berinde, Piotr Indyk, Graham Cormode, and Martin J. Strauss. Space-optimal heavy hitters with strong error bounds. *ACM Trans. Database Syst.*, 35(4), oct 2010. ISSN 0362-5915. doi: 10.1145/1862919.1862923. URL https://doi.org/10.1145/1862919.1862923.

Sanjay P. Bhat and Prashanth L.A. Concentration of risk measures: A wasserstein distance approach. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/091bc5440296cc0e41dd60ce22fbaf88-Paper.pdf.

Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, Dimitrios Myrisiotis, A. Pavan, and N. V. Vinodchandran. On approximating total variation distance. In Edith Elkind (ed.), *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pp. 3479–3487. International Joint Conferences on Artificial Intelligence Organization, 8 2023. doi: 10.24963/ijcai.2023/387. URL https://doi.org/10.24963/ijcai.2023/387. Main Track.

Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, Dimitrios Myrisiotis, A. Pavan, and N. V. Vinodchandran. Computational explorations of total variation distance, 2024. URL https://arxiv.org/abs/2412.10370.

Arnab Bhattacharyya, Sutanu Gayen, Kuldeep S. Meel, Dimitrios Myrisiotis, A. Pavan, and N. V. Vinodchandran. Total variation distance meets probabilistic inference. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2025.

Clément L. Canonne. Topics and techniques in distribution testing: A biased but representative sample. *Foundations and Trends® in Communications and Information Theory*, 19(6):1032–1198, 2022. ISSN 1567-2190. doi: 10.1561/0100000114. URL http://dx.doi.org/10.1561/0100000114.

Amit Chakrabarti. Data stream algorithms lecture notes - computer science, Mar 2024. URL https://www.cs.dartmouth.edu/~ac/Teach/data-streams-lecnotes.pdf.

Moses Charikar, Kevin C. Chen, and Martin Farach-Colton. Finding frequent items in data streams. In Peter Widmayer, Francisco Triguero Ruiz, Rafael Morales Bueno, Matthew Hennessy, Stephan J. Eidenbenz, and Ricardo Conejo (eds.), *Automata, Languages and Programming, 29th International Colloquium, ICALP 2002, Malaga, Spain, July 8-13, 2002, Proceedings*, volume 2380 of *Lecture Notes in Computer Science*, pp. 693–703. Springer, 2002. doi: 10.1007/3-540-45465-9_59. URL https://doi.org/10.1007/3-540-45465-9_59.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. Statistical optimal transport, 2024. URL https://arxiv.org/abs/2407.18163.

Lénaïc Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster wasserstein distance estimation with the sinkhorn divergence. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Fair regression with wasserstein barycenters. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/51cdbd2611e844ece5d80878eb770436-Abstract.html.

Doron Cohen, Aryeh Kontorovich, and Geoffrey Wolfer. Learning discrete distributions with infinite support. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Graham Cormode. The continuous distributed monitoring model. *ACM SIGMOD Record*, 42(1):5–14, 2013.

Graham Cormode and Marios Hadjieleftheriou. Finding frequent items in data streams. *Proc. VLDB Endow.*, 1(2):1530–1541, aug 2008. ISSN 2150-8097. doi: 10.14778/1454159.1454225. URL https://doi.org/10.14778/1454159.1454225.

Graham Cormode and S. Muthukrishnan. An improved data stream summary: the count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005. ISSN 0196-6774. doi: https://doi.org/10.1016/j.jalgor.2003.12.001. URL https://www.sciencedirect.com/science/article/pii/S0196677403001913.

Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

Juan Antonio Cuesta-Albertos, C. Matrán-Bea, and A. Tuero-Dıaz. On lower bounds for thel2-wasserstein metric in a hilbert space. *Journal of Theoretical Probability*, 9:263–283, 1996. URL https://api.semanticscholar.org/CorpusID:121128838.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pp. 2292–2300, Red Hook, NY, USA, 2013. Curran Associates Inc.

Philip J. Davis. Leonhard euler's integral: A historical profile of the gamma function: In memoriam: Milton abramowitz. *The American Mathematical Monthly*, 66(10):849–869, 1959. ISSN 00029890, 19300972. URL http://www.jstor.org/stable/2309786.

Luc Devroye and Laszlo Gyorfi. No Empirical Probability Measure can Converge in the Total Variation Sense for all Distributions. *The Annals of Statistics*, 18(3):1496 – 1499, 1990. doi: 10.1214/aos/1176347765. URL https://doi.org/10.1214/aos/1176347765.

Luc Devroye and Tommy Reddad. Discrete minimax estimation with trees. *Electronic Journal of Statistics*, 13(2):2595 – 2623, 2019. doi: 10.1214/19-EJS1586. URL https://doi.org/10.1214/19-EJS1586.

Ilias Diakonikolas. *Learning structured distributions*, pp. 267–288. CRC Press, 2016.

Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Fast and sample near-optimal algorithms for learning multidimensional histograms. In *Conference On Learning Theory*, pp. 819–842. PMLR, 2018.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: new datasets for fair machine learning. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA, 2021. Curran Associates Inc. ISBN 9781713845393.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *The Annals of Mathematical Statistics*, 27(3):642 – 669, 1956. doi: 10.1214/aoms/1177728174. URL https://doi.org/10.1214/aoms/1177728174.

Cynthia Dwork. Differential privacy. In *International colloquium on automata, languages, and programming*, pp. 1–12. Springer, 2006.

Y. C. Eldar and G. Kutyniok (eds.). *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.

Joan Feigenbaum, Sampath Kannan, Martin J. Strauss, and Mahesh Viswanathan. An approximate l1 - difference algorithm for massive data streams. *SIAM Journal on Computing*, 32(1):131–151, 2002. doi: 10.1137/S0097539799361701. URL https://doi.org/10.1137/S0097539799361701.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Weiming Feng, Liqiang Liu, and Tianren Liu. On deterministically approximating total variation distance. In *Proceedings of the 2024 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 1766–1791, 2024. doi: 10.1137/1.9781611977912.70. URL https://epubs.siam.org/doi/abs/10.1137/1.9781611977912.70.

S Foss, S Zachary, and D Korshunov. An introduction to heavy-tailed and subexponential distributions. *Springer Series in Operations Research and Financial Engineering*, 38:1–120, 2011.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, Aug 2015. ISSN 1432-2064. doi: 10.1007/s00440-014-0583-7. URL https://doi.org/10.1007/s00440-014-0583-7.

David A. Freedman and Persi Diaconis. On the histogram as a density estimator:l2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 57:453–476, 1981. URL https://api.semanticscholar.org/CorpusID:14437088.

Matthias Gelbrich. On a formula for the l2 wasserstein metric between measures on euclidean and hilbert spaces. *Mathematische Nachrichten*, 147:185–203, 1990. URL https://api.semanticscholar.org/CorpusID:124656337.

Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. Justicia: A stochastic SAT approach to formally verify fairness. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pp. 7554–7563. AAAI Press, 2021. doi: 10.1609/AAAI.V35I9.16925. URL https://doi.org/10.1609/aaai.v35i9.16925.

Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S. Meel. Algorithmic fairness verification with graphical models. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pp. 9539–9548. AAAI Press, 2022. doi: 10.1609/AAAI.V36I9.21187. URL https://doi.org/10.1609/aaai.v36i9.21187.

Sudipto Guha, Andrew McGregor, and Suresh Venkatasubramanian. Streaming and sublinear approximation of entropy and information distances. In *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm*, SODA '06, pp. 733–742, USA, 2006. Society for Industrial and Applied Mathematics. ISBN 0898716055.

Yanjun Han, Jiantao Jiao, and Tsachy Weissman. Minimax estimation of discrete distributions. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 2291–2295, 2015. doi: 10.1109/ISIT.2015.7282864.

Yannis Ioannidis. The history of histograms (abridged). In *Proceedings 2003 VLDB Conference*, pp. 19–30. Elsevier, 2003.

Ray Jiang, Aldo Pacchiano, Tom Stepleton, Heinrich Jiang, and Silvia Chiappa. Wasserstein fair classification. In Ryan P. Adams and Vibhav Gogate (eds.), *Proceedings of The 35th Uncertainty in Artificial Intelligence Conference*, volume 115 of *Proceedings of Machine Learning Research*, pp. 862–872. PMLR, 22–25 Jul 2020. URL https://proceedings.mlr.press/v115/jiang20a.html.

Hoang Anh Just, Feiyang Kang, Tianhao Wang, Yi Zeng, Myeongseob Ko, Ming Jin, and Ruoxi Jia. Lava: Data valuation without pre-specified learning algorithms. In *The Eleventh International Conference on Learning Representations*, 2024.

Peter Kairouz, H Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, et al. Advances and open problems in federated learning. *Foundations and trends® in machine learning*, 14(1–2):1–210, 2021.

Sudeep Kamath, Alon Orlitsky, Dheeraj Pichapati, and Ananda Theertha Suresh. On learning distributions from their samples. In Peter Grünwald, Elad Hazan, and Satyen Kale (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 1066–1100, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Kamath15.html.

Feiyang Kang, Hoang Anh Just, Anit Kumar Sahu, and Ruoxi Jia. Performance scaling via optimal transport: Enabling data selection from partially revealed sources. *Advances in Neural Information Processing Systems*, 36, 2024.

Antti Koskela and Jafar Mohammadi. Auditing differential privacy guarantees using density estimation, 2024. URL https://arxiv.org/abs/2406.04827.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pp. 957–966. PMLR, 2015.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Wenqian Li, Haozhi Wang, Zhe Huang, and Yan Pang. Private wasserstein distance with random noises, 2024. URL https://arxiv.org/abs/2404.06787.

J. Misra and David Gries. Finding repeated elements. *Science of Computer Programming*, 2(2):143–152, 1982. ISSN 0167-6423. doi: https://doi.org/10.1016/0167-6423(82)90012-0. URL https://www.sciencedirect.com/science/article/pii/0167642382900120.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, Cambridge, 2005.

Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. Tight auditing of differentially private machine learning. In *32nd USENIX Security Symposium (USENIX Security 23)*, pp. 1631–1648, 2023.

Victor M. Panaretos and Yoav Zemel. Statistical aspects of wasserstein distances. *Annual Review of Statistics and Its Application*, 2018. URL https://api.semanticscholar.org/CorpusID:88523547.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends in Machine Learning*, 11 (5-6):355–602, 2019. URL https://arxiv.org/abs/1803.00567.

Alain Rakotomamonjy, Kimia Nadjahi, and Liva Ralaivola. Federated wasserstein distance. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=rsg1mvUahT.

Sampriti Roy and Yadu Vasudev. Testing Properties of Distributions in the Streaming Model. In Satoru Iwata and Naonori Kakimura (eds.), *34th International Symposium on Algorithms and Computation (ISAAC 2023)*, volume 283 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pp. 56:1–56:17, Dagstuhl, Germany, 2023. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. ISBN 978-3-95977-289-1. doi: 10.4230/LIPIcs.ISAAC.2023.56. URL https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.ISAAC.2023.56.

David W. Scott. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 12 1979. ISSN 0006-3444. doi: 10.1093/biomet/66.3.605. URL https://doi.org/10.1093/biomet/66.3.605.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6(none):1550 – 1599, 2012. doi: 10.1214/12-EJS722. URL https://doi.org/10.1214/12-EJS722.

Matthew Staib, Sebastian Claici, Justin Solomon, and Stefanie Jegelka. Parallel streaming wasserstein barycenters. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 2647–2658, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/253f7b5d921338af34da817c00f42753-Abstract.html.

Thomas Steinke, Milad Nasr, and Matthew Jagielski. Privacy auditing with one (1) training run. *Advances in Neural Information Processing Systems*, 36, 2024.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.

Cédric Villani. *Optimal Transport*. Springer Berlin Heidelberg, 2009. ISBN 9783540710509. doi: 10.1007/978-3-540-71050-9. URL http://dx.doi.org/10.1007/978-3-540-71050-9.

Mariia Vladimirova, Stéphane Girard, Hien Nguyen, and Julyan Arbel. Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat*, 9(1):e318, 2020. doi: https://doi.org/10.1002/sta4.318. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.318. e318 sta4.318.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018.

Tom Yan and Chicheng Zhang. Active fairness auditing. In *International Conference on Machine Learning*, pp. 24929–24962. PMLR, 2022.

Appendix

Table of Contents

A Mergeable Misra Gries Algorithm

Algorithm 5 `MMG` (Anderson et al., 2017)

Initialize(k)
$T \leftarrow \emptyset$ {T is set of items assigned a counter}
$k \leftarrow k$
$k^* \leftarrow \frac{k}{2}$

Update(i, Δ)
if $\zeta_i \in T$ **then**
  $c(\zeta_i) \leftarrow c(\zeta_i) + \Delta$
else if $|T| < k$ **then**
  $T = T \cup \{\zeta_i\}$
  $c(\zeta_i) \leftarrow c(\zeta_i) + \Delta$
else
  DecrementCounters()
end if
if $\Delta \geq c_{k^*}$ **then**
  $T = T \cup \{\zeta_i\}$ // $|T| \leq k^* + 1 \leq k$ after this line
  $c(\zeta_i) \leftarrow c(\zeta_i) + \Delta - c_{k^*}$
end if

DecrementCounters()
// Notation: Let c_{k^*} be the k^*-th largest value, counting multiplicity,
// in the multiset $\{c(\zeta_j) \colon \zeta_j \in T\}$.
for $\zeta_j \in T$: **do**
  $c(\zeta_j) = c(\zeta_j) - c_{k^*}$
  if $c(\zeta_j) \leq 0$:
  $T = T \setminus \{\zeta_j\}$
end for

Estimate(i)
if $\zeta_i \in T$ **then**
  return $c(\zeta_i)$
else
  return 0
end if

EstimateCumulate(ζ_i)
$c\,(\zeta_i) \leftarrow$ **Estimate** (ζ_i)
return $\hat{F}_i \leftarrow \sum_{\zeta_j \leq \zeta_i} C_j$

Merge(T_2)
for $\zeta_i \in T_2$ **do**
  $\text{Update}_{T_1}(\zeta_i, c(\zeta_i))$
  return T_1
end for

First, we outline the proof for the Lemma 4, restated here for ease of reading.

Lemma 4 (Estimation Guarantee of `MMG`)**.** *(a) If `MMG` uses k counters, `MMG` $\cdot$ `Estimate` yields a summary $\{\hat{f}_j\}_{\mathcal{U}_j \in \mathcal{U}}$ satisfying $0 \leq f_j - \hat{f}_j \leq \frac{n^{Res(\tau)}}{(k-k^*)-\tau}$, for all $\tau \leq (k - k^*)$ and $\mathcal{U}_j \in |\mathcal{U}|$. Here, $n^{Res(\tau)}$ denotes*

the sum of frequency of all but τ most frequent items. (b) If we choose $k^* = \frac{k}{2}$ $\tau = k/4$, we get that $0 \leq f_j - \hat{f}_j \leq \frac{4n^{Res(k/4)}}{k}$, $\forall \mathcal{U}_j \in |\mathcal{U}|$.

The proof of the result follows that of Theorem 2 in Anderson et al. (2017). The change is in the Lemma 4. We provide the proof for the updated statement here. The notation is kept same as that of the original work for ease of reading.

As in the original paper, we define $N_l = \sum_{i=1}^{l} \boldsymbol{w}_j$, $C_l = \sum_{i=1}^{k} c_i$, and $E_l = \sum_{i \in [m]} f_{i,l} - \hat{f}_{i,l}$

Lemma 27. $E_n \leq (N_n - C_n) / (k - k^*)$

Proof. As in the original proof, we proceed by proof by induction, the case for $l = 0$ is true as $N_o = C_0 = E_0 = 0$.

Suppose the hypothesis holds for some $l - 1$, i.e.:

$$E_{l-1} \leq (N_{l-1} - C_{l-1}) / (k - k^*)$$

If the l-th element in the stream does not cause `DecrementCounters` to be called, we have $E_l = E_{l-1}$, $C_l = C_{l-1} + \boldsymbol{w}_l$, $N_l = N_{l-1} + \boldsymbol{w}_l$. Hence ,we have:

$$E_l = E_{l-1} \leq (N_{l-1} - C_{l-1}) / (k - k^*) = (N_l - C_l) / (k - k^*)$$

The case when `DecrementCounters` is called is more interesting, we have

$$E_l = E_{l-1} + c_{k^*} \tag{7}$$

Here, we know that at $l - 1$-th step, $(k - k^*)$ counters had value $\geq c_{k^*}$. Therefore, they retain some non-negative value after being reduced by c_{k^*}, while the remaining k^* counters have value 0. Hence, we have $C_l \leq C_{l-1} - (k - k^*) c_{k^*}$.

$$\begin{aligned} & N_l - C_l \\ =& N_{l-1} + \boldsymbol{w}_l - C_l \\ \geq& N_{l-1} + \boldsymbol{w}_l - C_{l-1} + (k - k^*) c_{k^*} \\ \geq& N_{l-1} - C_{l-1} + (k - k^*) c_{k^*} \end{aligned} \tag{8}$$

Now, combining Equations (7) and (8), we have:

$$E_l = E_{l-1} + c_{k^*} \leq \frac{N_{l-1} - C_{l-1}}{(k - k^*)} + c_{k^*} = \frac{N_{l-1} - C_{l-1} + (k - k^*) c_{k^*}}{(k - k^*)} \leq \frac{N_l - C_l}{(k - k^*)}$$

□

The second part of the proof of Theorem 2 in the original paper follows similarly under the updated statement of Lemma 27. Giving us the final result of Lemma 4.

To establish the estimation guarantee of `EstimateCumulate`, we need a simple corollary of the Lemma 4:

Corollary 28 (`MMG` Estimation Guarantees)**.** *For $j \in \{1, 2\}$, given streams ζ_j being stored in two separate sets of counters of size k, the merged summary satisfies:*

$$0 \leq f_i - \hat{f}_i \leq \frac{4n^{Res(\frac{k}{4})}}{k}$$

for all $j \leq k^*$, where $n^{Res(j)}$ denotes the sum frequency of all but j of the most frequent items.

Proof. We fix $k^* = \frac{k}{2}$, and $j = \frac{k}{4}$. These values satisfy the criteria of Theorem 4, i.e. $\frac{k}{2} = \Omega(k)$ and $j < k^*$. Plugging in these values gives the result. □

We now prove the Lemma 14:

Proof. We denote by Top_k the set of k elements with highest true frequency f_i. Then, we have:

$$\begin{aligned} F_i - \hat{F}_i &= \sum_{j \le i} f_j - \hat{f}_j \\ &= \sum_{\substack{j \le i \\ j \in Top_{\frac{k}{4}}}} f_j - \hat{f}_j + \sum_{\substack{j \le i \\ j \notin Top_{\frac{k}{4}}}} f_j \\ &\le \sum_{j \in Top_{\frac{k}{4}}} \frac{4n^{Res(\frac{k}{4})}}{k} + \sum_{j \notin Top_{\frac{k}{4}}} f_j \\ &\le 2n^{Res(\frac{k}{4})} \end{aligned}$$

Where the first inequality uses Corollary 28, and the last inequality follows from the definition of $n^{Res(k)}$. □

B Properties of Empirical Measure

In this section we provide the proofs for Theorems 10, 11, and 16. We restate the theorems for easy of reading. To establish the Theorem 10, we state the following equivalent definitions of a subgaussian distribution(Proposition 2.5.2 in Vershynin (2018)).

Lemma 29 (Bounded Second Moment Condition of sub-Gaussian distribution)**.** *A distribution $\mathcal{D}$ is $\sigma_{\mathcal{D}}$-subgaussian if and only if $\mathbb{E}_{X \sim \mathcal{D}} X^2 \le c_\sigma \sigma_{\mathcal{D}}^2$ where c_σ is a fixed constant.*

We will also require the following result on concentration of sum of squared sub-Gaussian random variables (Eldar & Kutyniok, 2012).

Lemma 30 (Concentration of Sum of Squares of i.i.d. sub-Gaussian r.v.)**.** *Let $X_1, X_2, ..., X_n$ be a sequence of i.i.d. sub-Gaussian random variables with parameter σ, then we have*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i\in[n]} X_i^2 - \mathbb{E}\left[X^2\right]\right| \ge t\right] \le 2\exp\left(-c\min\left(\frac{nt^2}{4\sigma^4}, \frac{nt}{2\sigma^2}\right)\right).$$

Theorem 10 (Empirical Measure is Sub-Gaussian)**.** *Let the true distribution μ be a σ_μ-sub-Gaussian with mean 0. If $n \ge \frac{c\log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$, the empirical measure μ_n generated by $X_1, X_2, ..., X_n$ drawn i.i.d. from μ is $(1+\varepsilon)\sigma_\mu$-sub-Gaussian with probability $1-\delta$, for any $\varepsilon, \delta \in (0,1)$ and a constant $c \ge 1$.*

Proof. Recall that by Lemma 29, to show that a distribution is sub-Gaussian , it suffices to show that it has bounded second moment. Let $X_1, X_2, ..., X_n$ be the i.i.d. samples generated from μ that constitutes μ_n. By Lemma 29, and Lemma 30, we have:

$$\mathbb{P}\left[\frac{1}{n}\sum_{i\in[n]} X_i^2 \ge t + c_\sigma \sigma_\mu^2\right] \le 2\exp\left(-c\min\left(\frac{nt^2}{\sigma^4}, \frac{nt}{\sigma^2}\right)\right).$$

Now, we fix $t = \frac{\sigma_\mu^2\sqrt{\log\left(\frac{2}{\delta}\right)}}{\sqrt{cn}}$. Under the assumption $n \geq \frac{\log\left(\frac{2}{\delta}\right)}{4c}$, we then have $\frac{nt^2}{4\sigma_\mu^4} \leq \frac{nt}{2\sigma_\mu^2}$ for $c \geq 1$. Hence, we have

$$\begin{aligned}
&\mathbb{P}\left[\frac{1}{n}\sum_{i\in[n]} X_i^2 \geq (1+\varepsilon)c_\sigma\sigma_\mu^2\right] \\
\leq &\mathbb{P}\left[\frac{1}{n}\sum_{i\in[n]} X_i^2 \geq \frac{\sigma_\mu^2\sqrt{\log\left(\frac{2}{\delta}\right)}}{\sqrt{cn}} + c_\sigma\sigma_\mu^2\right] \\
\leq &\,\delta\,.
\end{aligned}$$

The first inequality follows from the fact that assuming $n \geq \frac{c'\log\left(\frac{1}{\delta}\right)}{\varepsilon^2}$ and setting appropriate constant c' that ensures $c_\sigma\varepsilon\sigma_\mu^2 \geq \frac{c\sigma_\mu^2\sqrt{\log\left(\frac{2}{\delta}\right)}}{\sqrt{n}}$, and $\frac{c'\log\left(\frac{1}{\delta}\right)}{\varepsilon^2} \geq \frac{\log\left(\frac{2}{\delta}\right)}{4c}$. The second inequality follows from Lemma 30. Thus, By Lemma 29 ensures that the empirical measure μ_n is $(1+\varepsilon)\sigma_\mu$-sub-Gaussian with probability $1-\delta$. □

Next, we prove Theorem 11. We introduce the multiplicative chernoff bound that will be relevant to our analysis.

Lemma 31 (Multiplicative Chernoff Bound(Mitzenmacher & Upfal, 2005))**.** *Given i.i.d. random variables $X_1, X_2, ..., X_n$ where $\Pr[X_i = 1] \leq p$ and, define $X = \frac{1}{n}\sum_{i\in[n]} X_i$. Then, we have:*

$$\Pr[X \geq (1+\varepsilon)p] \leq \exp\left(-\frac{n\varepsilon^2 p}{3}\right) \qquad 0 \leq \varepsilon < 1 \tag{9}$$

$$\tag{10}$$

Now, we prove Theorem 11, restated here for ease of reading:

Theorem 11 (Empirical Distribution is sub-Weibull)**.** *Given an α-sub-Weibull true measure μ, the empirical measure μ_n generated by $X_1, X_2, ..., X_n$ drawn i.i.d. from μ is (τ, α)-sub-Weibull with probability at least $1-\delta$ given $n \geq \frac{\exp\left(\sqrt[\alpha]{\tau}\right)}{12}\log\frac{1}{\delta}$.*

Proof. To capture the tail behaviour w.r.t. some $t \leq \tau$ for each sample, we define random variables $Z_1, Z_2, ..., Z_n$ as:

$$Z_i = \begin{cases} 1 & X_i \geq t \\ 0 & X_i < t \end{cases}$$

We denote by $Z = \frac{1}{n}\sum_{i\in[n]} Z_i$. Then, we have for the empirical measure μ_n,

$$\Pr_{X\sim\mu_n}[X \geq t] = Z$$

Furthermore, we have:

$$\Pr\left[Z_i = 1\right] \leq c_\alpha \exp\left(-t^{1/\alpha}\right)$$

Then, by Lemma 31, we have:

$$\Pr\left[Z \geq 1.5c_\alpha \exp\left(t^{1/\alpha}\right)\right] \leq \exp\left(-\frac{nc_\alpha \exp\left(-t^{1/\alpha}\right)}{12}\right) \leq \exp\left(-\frac{nc_\alpha \exp\left(-\sqrt[\alpha]{\tau}\right)}{12}\right) \leq \delta$$

□

Next, we prove the Theorem 16. For that purpose, we need the following lemma.

Lemma 32 (DKW Inequality (Dvoretzky et al., 1956)). *Given a sequence of i.i.d. random variables $X_1, X_2, ..., X_n$ generated from a distribution μ with true cdf Q_μ. Let Q_{μ_n} be the cdf of the empirical measure μ_n generated from the samples $X_1, X_2, ..., X_n$. Then, for all $\varepsilon \geq 0$, given $n \geq \frac{\ln(2)}{2\varepsilon^2}$, we have:*

$$\mathbb{P}\left(\sup_{x\in\mathbb{R}} |Q_\mu(x) - Q_{\mu_n}(x)| \geq \varepsilon\right) \leq 2\exp\left(-2n\varepsilon^2\right)$$

Theorem 16 (Bucketed Empirical Measure is Bi-Lipschitz). *Let the bucketed empirical measure $\mu_{b,n}$ generated by n i.i.d. samples coming from a distribution μ with ℓ_μ bi-Lipschitz CDF. For any $\varepsilon, \delta \in (0,1)$ and fixed constant $c > 0$, if $n \geq \frac{c}{\varepsilon^2 b^2}\log\left(\frac{2}{\delta}\right)\max\left\{\frac{1}{\ell_\mu^2}, \ell_\mu^2\right\}$, $\mu_{b,n}$ is $(1+\varepsilon)\ell_\mu$ bi-Lipschitz with probability $1-\delta$.*

Proof. For $n \geq \log\left(\frac{2}{\delta}\right)\max\left(\frac{2}{\varepsilon^2 b^2 \ell_\mu^2}, \frac{8\ell_\mu^2}{\varepsilon^2 b^2}\right)$, we have by Lemma 32:

$$\mathbb{P}\left[\sup_{x\in\mathcal{B}} \left|Q_\mu(x) - Q_{\mu_{b,n}}(x)\right| \geq \min\left(\frac{\varepsilon b \ell_\mu}{2}, \frac{\varepsilon b}{4\ell_\mu}\right)\right] \leq \delta$$

Let us denote $\varepsilon_{DKW} = \min\left(\frac{\varepsilon b\ell_\mu}{2}, \frac{\varepsilon b}{4\ell_\mu}\right)$. Then, we have with probability $1-\delta$,

$$\sup_{x\in\mathcal{B}} \left|Q_\mu(x) - Q_{\mu_{b,n}}(x)\right| \leq \varepsilon_{DKW}\,. \tag{11}$$

Now, we have $\forall a, b \in \mathcal{B}$, with probability $1-\delta$:

$$\begin{aligned}
&\left|Q_{\mu_{b,n}}(a) - Q_{\mu_{b,n}}(b)\right| \\
\leq &|Q_\mu(a) - Q_\mu(b)| + 2\varepsilon_{DKW} \\
\leq &\ell_\mu |a-b| + \varepsilon\ell_\mu |a-b| \\
\leq &\ell_\mu(1+\varepsilon)|a-b|
\end{aligned}$$

Where the first inequality follows from Equation (11) and triangle inequality, and the second inequality follows from the fact that $\varepsilon_{DKW} \leq \frac{\varepsilon b\ell_\mu}{2}$ and $b \leq |a-b|, \forall a, b \in \mathcal{B}$. For the other side of the inequality,

$$\begin{aligned}
&\left|Q_{\mu_{b,n}}(a) - Q_{\mu_{b,n}}(b)\right| \\
\geq &|Q_\mu(a) - Q_\mu(b)| - 2\varepsilon_{DKW} \\
\geq &\frac{1}{\ell_\mu}||a-b|| - \frac{\varepsilon|a-b|}{2\ell_\mu} \\
\geq &\frac{1}{\ell_\mu(1+\varepsilon)}|a-b|
\end{aligned}$$

Where the first inequality follows from Equation (11) and triangle inequality, and the second inequality follows from the fact that $\varepsilon_{DKW} \leq \frac{\varepsilon b}{4\ell_\mu}$ and $b \leq |a-b|, \forall a, b \in \mathcal{B}$, and the third inequality follows from the fact that for all $\varepsilon \in (0,1)$, $1 - \frac{\varepsilon}{2} \geq \frac{1}{1+\varepsilon}$. □

C Estimation Guarantees of SPA and SCA

In this section, we provide the proofs for Theorem 13 and 15. We restate the theorems here for ease of reading:

Theorem 13 (SPA Learning Error). *(a) If SPA uses k buckets and outputs $\hat{\mu}_n$, then for all $i \in I$,*

$$\left|p_{\hat{\mu}_n}(i) - p_{\mu_{b,n}}(i)\right| \leq \max\left\{\frac{4n^{Res(k/4)}}{n}\frac{f_i}{n}, \frac{f_i - \hat{f}_i}{n}\right\}.$$

(b) Further, if μ corresponding to μ_n is σ_μ-sub-Gaussian, $|\zeta| = n \geq c \log\left(\frac{1}{\delta}\right)$, and $k \geq \left\lceil \frac{8\sigma_\mu}{b}\sqrt{\log\left(\frac{1}{\varepsilon}\right)}\right\rceil$, then for all $i \in I$, $\mathbb{P}(\left|p_{\hat{\mu}_n}(i) - p_{\mu_{b,n}}(i)\right| \geq \varepsilon) \leq \delta$.

(c) Further, if μ corresponding to μ_n is α-SubWeibull, $|\zeta| = n \geq \frac{c}{\varepsilon}\log\left(\frac{1}{\delta}\right)$, and $k \geq \left\lceil \frac{c}{b}\left(\log\left(\frac{1}{\varepsilon}\right)\right)^\alpha\right\rceil$, then for all $i \in I$, $\mathbb{P}(\left|p_{\hat{\mu}_n}(i) - p_{\mu_{b,n}}(i)\right| \geq \varepsilon) \leq \delta$.

Proof of Theorem 13. **Proof of (a):** We Start with the first equation:

$$\begin{aligned}
\hat{p}(i) - p_{\mu_{b,n}}(i) = \frac{\hat{f}_i}{\max_j \hat{F}_j} - \frac{f_i}{n} \leq & \frac{\hat{f}_i}{n - 2n^{Res(k/4)}} - \frac{f_i}{n} \\
\leq & \frac{f_i}{n - 2n^{Res(k/4)}} - \frac{f_i}{n} \\
= & f_i\left(\frac{1}{n - 2n^{Res(k/4)}} - \frac{1}{n}\right) \\
= & \frac{f_i}{n}\left(\frac{2n^{Res(k/4)}}{n - 2n^{Res(k/4)}}\right) \\
\leq & \frac{f_i}{n}\ \frac{4n^{Res(k/4)}}{n}
\end{aligned}$$

For the second part,

$$p_{\mu_{b,n}}(i) - \hat{p}(i) = \frac{f_i}{n} - \frac{\hat{f}_i}{\max_j \hat{F}_j} \leq \frac{f_i}{n} - \frac{\hat{f}_i}{n} \leq \frac{f_i - \hat{f}_i}{n}$$

Finally, we have:

$$\left|p_{\mu_{b,n}}(i) - \hat{p}(i)\right| = \max\left(p_{\mu_{b,n}}(i) - \hat{p}(i), \hat{p}(i) - p_{\mu_{b,n}}(i)\right) \leq \max\left(\frac{4n^{Res(k/4)}}{n}\frac{f_i}{n}, \frac{f_i - \hat{f}_i}{n}\right)$$

Now, we also have by Corollary 28 and the fact that $f_i \leq n$,

$$\left|p_{\mu_{b,n}}(i) - \hat{p}(i)\right| \leq \frac{4n^{Res(k/4)}}{n}$$

Proof of (b): If μ is σ_μ-sub-Gaussian, by Lemma 10, we know that μ_n generated by μ is $2\sigma_\mu$ sub-Gaussian given $n \geq c\log\left(\frac{1}{\delta}\right)$. Hence, by the property of sub-Gaussian distributions 7 and the fact that each bucket of size b,

$$\frac{n^{Res(k/4)}}{n} = \mathop{\mathbb{P}}_{\mu_n}\left[X \geq \left\lceil 2\sigma_\mu\sqrt{\log\left(\frac{1}{\varepsilon}\right)}\right\rceil\right] \leq \varepsilon\,.$$

Proof of (c): For α-sub-Weibull distributions, by Lemma 11, we know that μ_n generated by μ is τ, α-sub-Weibull given $n \geq \frac{\exp\left(\sqrt[\alpha]{\tau}\right)}{12}\log\frac{1}{\delta}$. Here, we can fix $\tau = \left(\log\left(\frac{4}{\varepsilon}\right)\right)^\alpha$, and thus $n \geq \frac{1}{3\varepsilon}\log\frac{1}{\delta}$ suffices. Hence, by the property of sub-Weibull distributions (Definition 6) and the fact that each bucket of size b,

$$\frac{n^{Res(k/4)}}{n} = \mathop{\mathbb{P}}_{\mu_n}\left[X \geq \left\lceil c\left(\log\left(\frac{1}{\varepsilon}\right)\right)^\alpha\right\rceil\right] \leq \varepsilon$$

□

Theorem 15 (SCA Learning Error)**.** *Given stream length $|\zeta| = n \geq \tau_n(\varepsilon, \delta)$ generated from a bounded-tail distribution μ, SCA with bucket width b sets #buckets to $k \geq \tau_k(\varepsilon, b)$ and outputs $\widehat{Q}$, such that with probability at least $1-\delta$, $\left|Q_{\mu_{b,n}}(i) - \widehat{Q}(i)\right| \leq \varepsilon$ for all $i \in I$.*

Tail Condition	$\tau_n(\varepsilon, \delta)$	$\tau_k(\varepsilon, b)$
σ_μ-sub-Gaussian	$c \log\left(\frac{1}{\delta}\right)$	$\left\lceil \frac{8\sigma_\mu}{b}\sqrt{\log\left(\frac{4}{\varepsilon}\right)} \right\rceil$
α-sub-Weibull	$\frac{c}{\varepsilon}\log\left(\frac{1}{\delta}\right)$	$\left\lceil \frac{c}{2b}\left(\log\left(\frac{4}{\varepsilon}\right)\right)^{\alpha} \right\rceil$

Proof of Theorem 15. **Step 1: Bounding the residuals in frequency estimation under tail conditions.** For sub-Gaussian distributions, by Lemma 10, we know that μ_n generated by μ is $2\sigma_\mu$ sub-Gaussian given $n \geq c \log\left(\frac{1}{\delta}\right)$. Hence, by the property of sub-Gaussian distributions 7 and the fact that each bucket of size b,

$$\frac{n^{Res(k/4)}}{n} = \mathop{\mathbb{P}}_{\mu_n}\left[X \geq \left\lceil 2\sigma_\mu \sqrt{\log\left(\frac{4}{\varepsilon}\right)} \right\rceil\right] \leq \frac{\varepsilon}{4} \tag{12}$$

For α-sub-Weibull distributions, by Lemma 11, we know that μ_n generated by μ is (τ, α)-sub-Weibull given $n \geq \frac{\exp\left(\sqrt[\alpha]{\tau}\right)}{12}\log\frac{1}{\delta}$. Here, we can fix $\tau = \left(\log\left(\frac{4}{\varepsilon}\right)\right)^{\alpha}$, and thus $n \geq \frac{1}{3\varepsilon}\log\frac{1}{\delta}$ suffices. Hence, by the property of sub-Weibull distributions (Definition 6) and the fact that each bucket of size b,

$$\frac{n^{Res(k/4)}}{n} = \mathop{\mathbb{P}}_{\mu_n}\left[X \geq \left\lceil c\left(\log\left(\frac{4}{\varepsilon}\right)\right)^{\alpha} \right\rceil\right] \leq \frac{\varepsilon}{4} \tag{13}$$

Step 2: Bounding the error in CDF estimation.

Part i. Now, we have:

$$\begin{aligned}
Q_{\mu_{b,n}}(i) - \widehat{Q}(i) &= \frac{F_i}{N} - \frac{\hat{F}_i}{\max_j \hat{F}_j} \\
&\leq \frac{F_i - \hat{F}_i}{n} && \text{since } n \geq \max_j \hat{F}_j \\
&= \frac{2n^{Res(k/4)}}{n} && \text{By Lemma 14} \\
&\leq \varepsilon && \text{By Equation (12) and (13)}
\end{aligned}$$

Part ii. Now, we look into the other side of the error. First, we observe that:

$$\frac{n^{Res(k/4)}}{n - 2n^{Res(k/4)}} \leq \frac{2n^{Res(k/4)}}{n} \leq \varepsilon$$

Here, the first inequality is given by the fact $2n^{Res(k/4)} \leq n$, and the second inequality follows from (12).

Thus,

$$\begin{aligned}
\widehat{Q}(i) - Q_{\mu_{b,n}}(i) &= \frac{\hat{F}_i}{\max_j \hat{F}_j} - \frac{F_i}{N} \\
&\leq \frac{F_i}{\max_j \hat{F}_j} - \frac{F_i}{N}, && \text{since } F_i \geq \hat{F}_i \\
&= F_i \left(\frac{1}{\max_j \hat{F}_j} - \frac{1}{n} \right) \\
&\leq F_i \left(\frac{1}{n - 2n^{Res(k/4)}} - \frac{1}{n} \right), && \text{since } \max_j \hat{F}_j \geq \max_j F_j - 2n^{Res(k/4)} = n - n^{Res(k/4)} \\
&= F_i \left(\frac{2n^{Res(k/4)}}{n\left(n - 2n^{Res(k/4)}\right)} \right) \\
&\leq \frac{2n^{Res(k/4)}}{n - 2n^{Res(k/4)}} && F_i \leq n \\
&\leq \frac{4n^{Res(k/4)}}{n} && n \geq 4n^{Res(k/4)} \\
&\leq \varepsilon && \text{By Equation (12) and (13)}
\end{aligned}$$

□

D Estimation Guarantee of SWA

In this section, we state the proofs for Theorem 17. We start with Lemma 33 which shows that a bucketed empirical measure $\mu_{b,n}$ is close to the underlying empirical measure μ_n in Wasserstein distance given the bucket size is sufficiently small.

Lemma 33. *Given a distribution μ and corresponding μ_b^{Disc} with bucket size b, we have:*

$$\mathcal{W}_{\mathrm{p}}\left(\mu, \mu_b^{Disc}\right) \leq b$$

Proof. For any i-th bucket corresponding to μ_b^{Disc}, we have:

$$Q_\mu(\mathcal{B}_i + \frac{b}{2}) = Q_{\mu_b^{Disc}}(\mathcal{B}_i + \frac{b}{2})$$

$$\left| Q_\mu^{-1}(x) - Q_{\mu_b^{Disc}}^{-1}(x) \right| \leq b$$

Hence, we have:

$$\mathcal{W}_{\mathrm{p}}\left(\mu, \mu_b^{Disc}\right) = \left(\int_0^1 \left| Q_\mu^{-1}(x) - Q_{\mu_b^{Disc}}^{-1}(x) \right|^p \, dx \right)^{\frac{1}{p}} \leq \left(\int_0^1 b^p \, dx \right)^{\frac{1}{p}} = b$$

This concludes the proof. □

A general version of Lemma 33 is given in Staib et al. (2017, Theorem 4.1). However, we prove our version here for simplicity and ease of access.

Next, we prove the following lemma that characterizes the approximation of a function based on an approximation guarantee on its inverse.

Lemma 34 (Inverse Approximation of inverse ℓ-Lipschitz functions)**.** *Let f be a ℓ bi-Lipschitz function. Given an estimate $\hat{f}$ of f such that $\hat{f}(x) \in [f(x) - \varepsilon, f(x) + \varepsilon]$, we can construct an estimate $\hat{f}^{-1}$ of the pseudoinverse function f^{-1} such that $\forall x \in \text{Range}(f)$:*

$$\hat{f}^{-1}(x) \in [f^{-1}(x) - \ell\varepsilon, f^{-1}(x) + \ell\varepsilon]$$

Proof. We construct $\hat{f}^{-1}$ as $\hat{f}^{-1}(y) = \min_x \left\{x \in \mathbb{R} \cup \{-\infty\} : \hat{f}(x) = y\right\}$. Let us consider $x \in \text{Range}(f)$, and $f^{-1}(x) = y_1$ and $\hat{f}^{-1}(x) = y_2$, and w.l.o.g. assume $y_2 \geq y_1$. Then, we have:

$$\begin{aligned} f(y_1) &= \hat{f}(y_2) \\ \hat{f}(y_1) + \varepsilon &\geq \hat{f}(y_2) \\ \varepsilon &\geq \hat{f}(y_2) - \hat{f}(y_1) \end{aligned}$$

Next, we use the fact that the function f is ℓ bi-Lipschitz to show that $y_2 - y_1$ is bounded by $\ell\varepsilon$:

$$\begin{aligned} & y_2 - y_1 \\ \leq & \hat{f}^{-1}(\hat{f}(y_2)) - \hat{f}^{-1}(\hat{f}(y_1)) && \text{By definition of } \hat{f}^{-1} \\ \leq & \ell(\hat{f}(y_2) - \hat{f}(y_1)) \leq \ell\varepsilon \end{aligned}$$

Hence, for any $x \in \text{Range}(f)$, we have $|\hat{f}^{-1}(x) - f^{-1}(x)| \leq \ell\varepsilon$, completing our proof. □

The following corollary follows directly.

Corollary 35 (From CDF to Inverse CDF)**.** *For an ℓ bi-Lipschitz distribution μ with CDF Q_μ and pseudoinverse CDF Q_μ^{-1}, given an approximate CDF $\widehat{Q}_\mu$ satisfying $|\widehat{Q}_\mu(x) - Q_\mu(x)| \leq \varepsilon, \forall x \in \mathbb{R}$, we construct an approximation of pseudoinverse CDF $\widehat{Q_\mu^{-1}}$, such that $\forall x \in [0,1]$, $|\widehat{Q_\mu^{-1}}(x) - Q_\mu^{-1}(x)| \leq \ell\varepsilon$.*

Now, we prove the Theorem 17. We restate the theorems here for ease of reading:

Theorem 17 (Universal Learning of `SCA` in $\mathcal{W}_\text{p}(,)$)**.** *Given $|\zeta| = n \geq \tau_n(\ell_\mu, \varepsilon, \delta)$ generated from a bounded-tail distribution μ and $\varepsilon, \delta \in (0,1)$, if we set the bucket width $b = \varepsilon/2$, then `SCA` uses $\tau_k(\ell_\mu, \varepsilon)$ buckets (i.e. space) and outputs $\hat{\mu}_n$, such that $\mathbb{P}(\mathcal{W}_\text{p}(\mu_n, \hat{\mu}_n) \geq \varepsilon) \leq \delta$. We omit the constants in $\tau_n(\ell_\mu, \varepsilon, \delta)$ and $\tau_k(\ell_\mu, \varepsilon)$ for simplicity.*

Tail Condition	$\tau_n(\ell_\mu, \varepsilon, \delta)$	$\tau_k(\ell_\mu, \varepsilon)$
σ_μ-sub-Gaussian	$\log(1/\delta) \max\left\{\frac{1}{\ell_\mu^2}, \ell_\mu^2\right\}$	$\frac{\sigma_\mu}{\varepsilon} \log\left(\frac{\ell_\mu}{\varepsilon}\right)$
α-sub-Weibull	$\log(1/\delta) \max\left\{\frac{1}{\varepsilon}, \frac{1}{\ell_\mu^2}, \ell_\mu^2\right\}$	$\frac{1}{\varepsilon}\left(\log\left(\frac{\ell_\mu}{\varepsilon}\right)\right)^\alpha$

Proof of Theorem 17. By Theorem 16, 15, and Lemma 34; and the parameters we have used, we have with probability $1 - \delta$:

$$\left|Q^{-1}_{\mu_{\varepsilon/2,n}} - \widehat{Q^{-1}}\right| \leq 2\ell_\mu \left|Q_{\mu_{\varepsilon/2,n}} - \widehat{Q}\right| \leq \varepsilon/2$$

Hence, by Lemma 2, we have with probability at least $1 - \delta$:

$$\begin{aligned} \mathcal{W}_\text{p}(\mu_n, \hat{\mu}_n) &\leq \mathcal{W}_\text{p}\left(\mu_n, \mu_{\varepsilon/2,n}\right) + \mathcal{W}_\text{p}\left(\mu_{\varepsilon/2,n}, \hat{\mu}_n\right) \\ &\leq \varepsilon/2 + \left(\int_0^1 |Q^{-1}_{\mu_{\varepsilon/2,n}}(r) - \widehat{Q^{-1}}(r)|^p \, dr\right)^{1/p} \\ &\leq \varepsilon/2 + \left(\int_0^1 |\varepsilon/2|^p \, dr\right)^{1/p} = \varepsilon \end{aligned}$$

The bound on number of buckets can be obtained by plugging in the values in the line 3 of Algorithm 3 in Theorem 15. □

We now proof the Theorem 18, restated here for ease of reading:

Theorem 18 (PAC Guarantee of `SWA`)**.** *Given $|\zeta| = n \geq \tau_n(\ell_\mu, \varepsilon, \delta)$ generated from a bounded-tail distribution μ, with bucket width $b = \varepsilon/2$, then `SCA` uses $\tau_k(\ell_\mu, \varepsilon)$ space and outputs $\hat{\mu}_n$ such that $\mathbb{P}(\mathcal{W}_1(\mu, \hat{\mu}_n) \leq \varepsilon) \geq 1 - \delta$. This further shows that for a stream of size $n = \mathcal{O}\left(\epsilon^{-2}\log(1/\delta)\right)$ from two distributions μ and ν, `SWA` yields*

$$|\,\mathcal{W}_1(\hat{\mu}_n, \hat{\nu}_n) - \mathcal{W}_1(\mu, \nu)\,| \leq 4\varepsilon \tag{5}$$

with probability $1 - 2\delta$. Here, $\epsilon \in (0, 1/4]$ and $\delta \in (0, 1/2]$. Here, $\sigma \triangleq \max\{\sigma_\mu, \sigma_\nu\}$, $\ell \triangleq \max\{\ell_\mu, \ell_\nu\}$. We omit the constants in $\tau_n(\ell_\mu, \varepsilon, \delta)$ and $\tau_k(\ell_\mu, \varepsilon)$ for simplicity.

Tail Condition	$\tau_n(\ell_\mu, \varepsilon, \delta)$	$\tau_k(\ell_\mu, \varepsilon)$
σ-sub-Gaussian	$\log(1/\delta)\max\left\{\frac{1}{\varepsilon^2}, \frac{1}{\ell^2}, \ell^2\right\}$	$\frac{\sigma}{\varepsilon}\log\left(\frac{\ell}{\varepsilon}\right)$
α-sub-Weibull	$\log(1/\delta)\max\left\{\frac{1}{\varepsilon^2}, \left(\log\frac{1}{\delta}\right)^{2\alpha-1}, \frac{1}{\ell^2}, \ell^2\right\}$	$\frac{1}{\varepsilon}\left(\log\left(\frac{\ell}{\varepsilon}\right)\right)^{\alpha}$

Proof. We combine the results of Corollary 45 , 46, and Theorem 18 to establish this result. By Corollary 45 for sub-Gaussian distributions, we have with probability $1 - 2\delta$:

$$|\mathcal{W}_1(\mu, \nu) - \mathcal{W}_1(\mu_n, \nu_n)| \leq 2\varepsilon \qquad \text{By Triangle Inequality} \tag{14}$$

Similarly, by Corollary 46 for α-sub-Weibull distributions, we have with probability $1 - 2\delta$:

$$|\mathcal{W}_1(\mu, \nu) - \mathcal{W}_1(\mu_n, \nu_n)| \leq 2\varepsilon \qquad \text{By Triangle Inequality} \tag{15}$$

By Theorem 18, we have with probability $1 - 2\delta$:

$$|\mathcal{W}_1(\hat{\mu}_n, \hat{\nu}_n) - \mathcal{W}_1(\mu_n, \nu_n)| \leq 2\varepsilon \qquad \text{By Triangle Inequality} \tag{16}$$

A union bound argument and triangle inequality over Equations (14) or (15) and (16) completes the proof. □

E Estimation Guarantee of `STVA`

In this section, we prove the theorems concerning `STVA`, i.e. Theorem 20, Lemma 21, and Theorems 22 and 23.

E.1 Proof of Theorem 20

In this section, we detail the proof of Theorem 20.

Theorem 20 (Bucket Width for TV Distance Estimation)**.** *Given a bounded-tail distribution μ with ℓ_μ-Lipschitz PDF and a corresponding bucketed measure μ_b, if we fix the bucket width $b = \tau_b(\varepsilon, \ell)$, we have $\mathrm{TV}(\mu, \mu_b) \leq \epsilon$.*

Tail Condition	σ_μ-sub-Gaussian	α-sub-Weibull
$\tau_b(\varepsilon, \ell_\mu)$	$\frac{\varepsilon}{\sigma_\mu \ell_\mu \sqrt{\log(2/\varepsilon)}}$	$\frac{\varepsilon}{\ell_\mu(\log(2c_\alpha/\varepsilon))^{\alpha}}$

Proof. **Initial Bounds:** The distribution μ has a ℓ_μ-Lipschitz PDF. Consider any bucket $\sqcup$ of size b, then $\min_{y\in\sqcup}\mu(y) - \max_{y\in\sqcup}\mu(y) \leq b\ell_\mu$. As $\min_{y\in\sqcup}\mu(y) \leq \mu_b(x) \leq \max_{y\in\sqcup}\mu(y), \forall x \in \sqcup$, we have for any bucket $\sqcup$:

$$\begin{aligned} |\mu(x) - \mu_b(x)| &\leq b\ell_\mu \qquad && \forall x \in \sqcup \\ \int_{\sqcup} |\mu(x) - \mu_b(x)|\,\mathrm{d}x &\leq b^2\ell_\mu \end{aligned} \tag{17}$$

The main idea of our proof is that we bound the error due to bucketing in buckets with high frequency and use the fact that the remaining buckets has sufficiently small frequency to achieve our bound. To denote the

buckets with high frequency, recall that Top_k denotes the set of k buckets with highest frequency, b denotes the length of each bucket.

$$
\begin{aligned}
&\mathrm{TV}\left(\mu, \mu_b\right) \\
=&\frac{1}{2}\int_{-\infty}^{\infty}\left|\mu(x)-\mu_b(x)\right|\mathrm{d}x \qquad (18)\\
=&\frac{1}{2}\sum_{\sqcup\in\mathcal{B}}\int_{\sqcup}\left|\mu(x)-\mu_b(x)\right|\mathrm{d}x \\
=&\frac{1}{2}\sum_{\sqcup\in Top_k}\int_{\sqcup}\left|\mu(x)-\mu_b(x)\right|\mathrm{d}x+\frac{1}{2}\sum_{\sqcup\notin Top_k}\int_{\sqcup}\left|\mu(x)-\mu_b(x)\right|\mathrm{d}x \\
\leq&\frac{1}{2}\sum_{\sqcup\in Top_k} b^2\ell_\mu+\frac{1}{2}\sum_{\sqcup\notin Top_k}\left[\int_{\sqcup}\mu_b(x)\mathrm{d}x+\int_{\sqcup}\mu(x)\mathrm{d}x\right] \qquad \text{By Equation (17)} \\
\leq& kb^2\ell_\mu/2+\sum_{\sqcup\notin Top_k}\int_{\sqcup}\mu(x)\mathrm{d}x \qquad (19)
\end{aligned}
$$

Sub-Weibull Distributions: For α-sub-Weibull distributions, we fix $k=\frac{\ell_\mu(\log(2c_\alpha/\varepsilon))^{2\alpha}}{\varepsilon}$ and bucket size $b=\frac{\varepsilon}{\ell_\mu(\log(2c_\alpha/\varepsilon))^{\alpha}}$ and bound the terms in Equation (19). For the first term, we have:

$$
kb^2\frac{\ell_\mu}{2}=\frac{\ell_\mu\left(\log(2c_\alpha/\varepsilon)\right)^{2\alpha}}{\varepsilon}\left(\frac{\varepsilon}{\ell_\mu\left(\log\left(2c_\alpha/\varepsilon\right)\right)^{\alpha}}\right)^2\frac{\ell_\mu}{2}\leq\varepsilon/2 \qquad (20)
$$

For the second term:

$$
\begin{aligned}
\sum_{\sqcup\notin Top_{\frac{\ell_\mu(\log(2/\varepsilon))^{2\alpha}}{\varepsilon}}}\int_{\sqcup}\mu(x)\mathrm{d}x&=\mathop{\mathbb{P}}_{X\sim\mu}\left[X\in\cup_{\sqcup\notin Top_{\frac{\ell_\mu(\log(2/\varepsilon))^{2\alpha}}{\varepsilon}}}\sqcup\right] \\
&\leq\mathop{\mathbb{P}}_{X\sim\mu}\left[X\geq\left(\log\frac{2c_\alpha}{\varepsilon}\right)^{\alpha}\right]\leq\varepsilon/2 \qquad (21)
\end{aligned}
$$

Combining Equations (20) and (21) with the bound from Equation (19), we obtain the result.

Sub-Gaussian Distributions: For sub-Gaussian distributions, we fix $k=\frac{\sigma_\mu^2\ell_\mu\log(2/\varepsilon)}{\varepsilon}$ and bucket size $b=\frac{\varepsilon}{\sigma_\mu\ell_\mu\sqrt{\log(2/\varepsilon)}}$ and bound the terms in Equation (19). For the first term, we have

$$
kb^2\frac{\ell_\mu}{2}=\frac{\sigma_\mu^2\ell_\mu\log(2/\varepsilon)}{\varepsilon}\left(\frac{\varepsilon}{\sigma_\mu\ell_\mu\sqrt{\log\left(2/\varepsilon\right)}}\right)^2\frac{\ell_\mu}{2}\leq\varepsilon/2 \qquad (22)
$$

For the second term:

$$
\begin{aligned}
\sum_{\sqcup\notin Top_{\frac{\sigma_\mu^2\ell_\mu\log^2(2/\varepsilon)}{\varepsilon}}}\int_{\sqcup}\mu(x)\mathrm{d}x&=\mathop{\mathbb{P}}_{X\sim\mu}\left[X\in\cup_{\sqcup\notin Top_{\frac{\sigma_\mu^2\ell_\mu\log(2/\varepsilon)}{\varepsilon}}}\sqcup\right] \\
&\leq\mathop{\mathbb{P}}_{X\sim\mu}\left[X\geq\sigma_\mu\sqrt{\log\left(2/\varepsilon\right)}\right] \\
&\leq\varepsilon/2 \qquad (23)
\end{aligned}
$$

Combining Equations (22) and (23) with the bound from Equation (19), we obtain the result. □

E.2 Proof of Lemma 21

Consider the two possible ways of looking at a histogram. Given a histogram defined over $\mathbb{R}$, one can think of it as defining a (not necessarily probability) measure over $\mathbb{R}$ defined so that the measure at each point is equal to the frequency of the bucket the point lies in. Alternatively, one can define a measure over the buckets with the measure at each bucket is equal to the frequency of the bucket. In this section, we formalize these notions and establish the relation between them in term of TV and Wasserstein distances.

Definition 36 (Bucketed Continuous Distribution)**.** Let $\mathcal{D}$ be a distribution supported on a (finite or infinite) closed interval $I \subseteq \mathbb{R}$ with (discrete or continuous) measure μ. Given x_0 as a reference point in I and b as the bucket width, we further represent I as $I(x_0, J, b) \triangleq \cup_{j\in J}[x_0 + jb, x_0 + (j+1)b]$. Here, index set $J \subseteq \mathbb{Z}$. We define the corresponding bucketed continuous distribution $\mathcal{D}_b^{Cont}$ with measure μ_b^{Cont} defined over $I \subseteq \mathbb{R}$ as:

$$\begin{aligned}\mu_b^{Cont}(x) &= \frac{1}{b}\int_{x_0+jb}^{x_0+(j+1)b} \mu(x)\mathrm{d}x \\ &= \frac{1}{b}\mu\left([x_0 + jb, x_0 + (j+1)b]\right) ,\end{aligned}$$

where $x \in [x_0 + jb, x_0 + (j+1)b]$.

Definition 37 (Bucketed Discrete Distribution)**.** Let $\mathcal{D}$ be a distribution supported on a (finite or infinite) closed interval $I \subseteq \mathbb{R}$ with (discrete or continuous) measure μ. Given x_0 as a reference point in I and b as the bucket width, we further represent I as $I(x_0, J, b) \triangleq \cup_{j\in J}[x_0 + jb, x_0 + (j+1)b]$. We define the corresponding bucketed discrete distribution $\mathcal{D}_b^{Disc}$ with measure μ_b^{Disc} defined over $\cup_{j\in J}\{x_0+jb+b/2\} \subseteq \mathbb{R}$ as $\mu_b^{Disc}(j) = \int_{x_0+jb}^{x_0+(j+1)b} \mu(x)\mathrm{d}x = \mu\left([x_0 + jb, x_0 + (j+1)b]\right)$.

The support set of μ_b^{Disc} inherits the metric structure of $\mathbb{R}$. For ease of notation, we express the distance between the i-th and j-th support points of μ_b^{Disc} as $d(\sqcup_i, \sqcup_j) \triangleq |i-j|b$.

We introduce the following lemma establishing the equivalence of bucketed discrete and continuous measure with respect to the TV distance.

Lemma 38 (TV of continuous and discrete bucketed distributions)**.** *Given two measures μ, and ν, we have:*

$$\mathrm{TV}\left(\mu_b^{Cont}, \nu_b^{Cont}\right) = \mathrm{TV}\left(\mu_b^{Disc}, \nu_b^{Disc}\right)$$

Proof.

$$\begin{aligned}\mathrm{TV}\left(\mu_b^{Cont}, \nu_b^{Cont}\right) =& \frac{1}{2}\int \left|\mu_b^{Cont}(x) - \nu_b^{Cont}(x)\right| \mathrm{d}x \\ =& \frac{1}{2}\sum_{j\in J}\frac{1}{b}\int_{x_0+jb}^{x_0+(j+1)b}\left|\int_{x_0+jb}^{x_0+(j+1)b}\mu(x)\mathrm{d}x - \int_{x_0+jb}^{x_0+(j+1)b}\nu(x)\mathrm{d}x\right|\mathrm{d}x \\ =& \frac{1}{2}\sum_{j\in J}\left|\int_{x_0+jb}^{x_0+(j+1)b}\mu(x)\mathrm{d}x - \int_{x_0+jb}^{x_0+(j+1)b}\nu(x)\mathrm{d}x\right| \\ =& \mathrm{TV}\left(\mu_b^{Disc}, \nu_b^{Disc}\right)\end{aligned}$$

□

We now establish a TV distance concentration result on bucketed empirical distribution corresponding to a sub-Gaussian distribution.

Lemma 21 (Concentration in TV over Infinite Buckets)**.** *Let μ_n be an empirical measure generated from a discrete bucketed measure μ_b corresponding to a bounded-tail distribution μ, and, $\varepsilon, \delta \in (0,1)$. Then, for $n \geq \tau_n(\varepsilon, b, \delta)$, $\mathbb{P}\left[\mathrm{TV}\left(\mu, \mu_n\right) \geq \varepsilon\right] \leq \delta$. Here, $\Gamma(\cdot)$ denotes the Gamma function (Davis, 1959).*

Tail Condition	σ_μ-sub-Gaussian	α-sub-Weibull
$\tau_n(\varepsilon, b, \delta)$	$c\varepsilon^{-2}\max\left\{\frac{4\sigma_\mu\sqrt{\pi}}{b}, \log(1/\delta)\right\}$	$c\varepsilon^{-2}\max\left\{\frac{2c_\alpha}{b}\Gamma(1+\alpha), \log(1/\delta)\right\}$

Proof. Let us consider the standard bijection from $\mathbb{Z}$ to $\mathbb{N}$ as:

$$f(x) = \begin{cases} 2\text{x} & \text{if } x > 0 \\ \text{-2x+1} & \text{if } x \leq 0 \end{cases}$$

σ_μ-Sub-Gaussian Distributions: For sub-Gaussian distributions,If i is even, we have:

$$p(i) = \mu_b^{Disc}(\sqcup_{i/2}) \leq \mu_b^{Disc}\left(\cup_{j\geq i/2}\sqcup_j\right) \leq \mu(\{x|x \geq bi/2\}) \leq 2\exp\left(-\frac{i^2b^2}{4\sigma_\mu^2}\right).$$

If i is odd, we have:

$$p(i) = \mu_b^{Disc}(\sqcup_{-(i-1)/2}) \leq \mu_b^{Disc}\left(\cup_{j\leq -(i-1)/2}\sqcup_j\right) \leq \mu(\{x|x \leq -b(i-1)/2\}) \leq 2\exp\left(-\frac{(i-1)^2b^2}{4\sigma_\mu^2}\right).$$

Combining these terms, we have:

$$\sum_{i\in\mathbb{N}}\sqrt{p_\mu(i)} \leq \sum_{i\in\mathbb{N}} 4\exp\left(-i^2b^2/4\sigma_\mu^2\right) \leq 4\int_0^\infty \exp(-x^2b^2/\sigma_\mu^2)\mathrm{d}x = 2\sqrt{4\pi\sigma_\mu^2/b^2} = \frac{4\sigma_\mu\sqrt{\pi}}{b}.$$

We now combine this with Lemma 43 to complete the proof.

α-SubWeibull Distributions: For α-SubWeibull distributions, if i is even, we have:

$$p(i) = \mu_b^{Disc}(\sqcup_{i/2}) \leq \mu_b^{Disc}\left(\cup_{j\geq i/2}\sqcup_j\right) \leq \mu(\{x|x \geq bi/2\}) \leq c_\alpha\exp\left(-(ib)^{1/\alpha}\right).$$

If i is odd, we have:

$$p(i) = \mu_b^{Disc}(\sqcup_{-(i-1)/2}) \leq \mu_b^{Disc}\left(\cup_{j\leq -(i-1)/2}\sqcup_j\right) \leq \mu(\{x|x \leq -b(i-1)/2\}) \leq c_\alpha\exp\left(-((i-1)b)^{1/\alpha}\right).$$

Combining these terms, we have:

$$\sum_{i\in\mathbb{N}}\sqrt{p_\mu(i)} \leq \sum_{i\in\mathbb{N}} 2c_\alpha\exp\left(-(ib)^{1/\alpha}\right) \leq 2c_\alpha\int_0^\infty \exp\left(-(xb)^{1/\alpha}\right)\mathrm{d}x = 2c_\alpha\frac{\Gamma(1+\alpha)}{b}.$$

We now combine this with Lemma 43 to complete the proof. □

E.3 Bounding Errors of SPA and STVA (Theorems 22 and 23)

Theorem 22 (Learning Error of SPA)**.** *If SPA accesses a stream of length $|\zeta| = n \geq \tau_n(\varepsilon, \delta)$ from a bounded-tail distribution, and uses $\tau_k(\varepsilon, \delta)$ buckets to output a sublinear summary $p_{\hat{\mu}_n}$, then with probability $1-\delta$, $\mathrm{TV}(\hat{\mu}_n, \mu_{b,n}) \leq \left\|p_{\hat{\mu}_n} - p_{\mu_{b,n}}\right\|_1 \leq \varepsilon$.*

Tail Condition	$\tau_n(\varepsilon, \delta)$	$\tau_k(\varepsilon, b)$
σ_μ-sub-Gaussian	$c\log\left(\frac{1}{\delta}\right)$	$\left\lceil\frac{8\sigma_\mu}{b}\sqrt{\log\left(\frac{6}{\varepsilon}\right)}\right\rceil$
α-sub-Weibull	$\frac{c}{\varepsilon}\log\frac{1}{\delta}$	$\left\lceil\frac{c}{b}\left(\log\left(\frac{6}{\varepsilon}\right)\right)^\alpha\right\rceil$

Proof. **Step 1: Bounding the residuals in frequency estimation under tail conditions.** For sub-Gaussian distributions, by Theorem 10, we know that $\mu_{b,n}$ generated by μ_b^{Disc} is $2\sigma_\mu$ sub-Gaussian given $n \geq c \log\left(\frac{1}{\delta}\right)$. Hence, by the property of sub-Gaussian distributions 7 and the fact that length of each bucket is b,

$$\frac{n^{Res(k/4)}}{n} \leq \mathop{\mathbb{P}}_{\mu_{b,n}}\left[X \geq \left\lceil 2\sigma_\mu \sqrt{\log\left(\frac{6}{\varepsilon}\right)}\right\rceil\right] \leq \frac{\varepsilon}{6} \tag{24}$$

For α-sub-Weibull distributions, by Lemma 11, we know that μ_n generated by μ is (τ, α)-sub-Weibull given $n \geq \frac{\exp(\sqrt[\alpha]{\tau})}{12} \log\frac{1}{\delta}$. Here, we can fix $\tau = \left(\log\left(\frac{6}{\varepsilon}\right)\right)^\alpha$, and thus $n \geq \frac{1}{2\varepsilon}\log\frac{1}{\delta}$ suffices. Hence, by the property of sub-Weibull distributions (Definition 6) and the fact that each bucket of size b,

$$\frac{n^{Res(k/4)}}{n} = \mathop{\mathbb{P}}_{\mu_n}\left[X \geq \left\lceil c\left(\log\left(\frac{6}{\varepsilon}\right)\right)^\alpha\right\rceil\right] \leq \frac{\varepsilon}{4} \tag{25}$$

Step 2: Bounding the error in PDF estimation. Now, we proceed to bound the TV Distance. We denote by Top_k the set of k elements with highest true frequency f_i. Then, we have:

$$\begin{aligned}
\left\|\hat{p} - p_{\mu_{b,n}}\right\|_1 &= \sum_i \left|p_{\mu_{b,n}}(i) - \hat{p}(i)\right| \\
&\leq \sum_i \max\left(\frac{4n^{Res(k/4)}}{n} \cdot \frac{f_i}{n}, \frac{f_i - \hat{f}_i}{n}\right) && \text{Theorem 13} \\
&\leq \sum_i \frac{4n^{Res(k/4)}}{n} \cdot \frac{f_i}{n} + \sum_i \frac{f_i - \hat{f}_i}{n} \\
&\leq \frac{4n^{Res(k/4)}}{n}\sum_i \frac{f_i}{n} + \sum_{i \in Top_{k/4}} \frac{f_i - \hat{f}_i}{n} + \sum_{i \notin Top_{k/4}} \frac{f_i}{n} \\
&\leq \frac{4n^{Res(k/4)}}{n} + \frac{\sum_{i \in Learning Top_{k/4}} f_i - \hat{f}_i + \sum_{i \notin Top_{k/4}} f_i}{n} \\
&\leq \frac{4n^{Res(k/4)}}{n} + \frac{2n^{Res(k/4)}}{n} && \text{Lemma 4} \\
&= \frac{6n^{Res(k/4)}}{n} \\
&\leq \varepsilon && \text{Equation (24) and (25)}
\end{aligned}$$

□

We now proof the Theorem 23, restated here for ease of reading:

Theorem 23 (PAC Guarantee of `STVA`)**.** *Given $\epsilon \in (0, 1/6]$, $\delta \in (0, 1/4]$, a stream of size $n \geq \tau_n(\varepsilon, \ell, \delta)$ from two bounded-tail and ℓ-Lipschitz distributions μ and ν, and bucket width $b = \tau_b(\varepsilon, \ell, \delta)$, **STVA** uses $\tau_k(\varepsilon, \ell, \delta)$ space and*

$$\mathbb{P}\left(\left|\operatorname{TV}(\hat{\mu}_n, \hat{\nu}_n) - \operatorname{TV}(\mu, \nu)\right| \geq 6\varepsilon\right) \leq 4\delta\,. \tag{6}$$

Here, $\sigma \triangleq \max\{\sigma_\mu, \sigma_\nu\}$, $\ell \triangleq \max\{\ell_\mu, \ell_\nu\}$. We omit the constants in $\tau_n(\ell_\mu, \varepsilon, \delta)$ and $\tau_k(\ell_\mu, \varepsilon)$ for simplicity.

Tail Condition	$\tau_n(\varepsilon, \ell, \delta)$	$\tau_k(\varepsilon, \ell)$	$\tau_b(\varepsilon, \ell)$
σ-sub-Gaussian	$\varepsilon^{-2}\max\left\{\frac{\sigma^2 \ell \log(1/\varepsilon)}{\varepsilon}, \log\left(\frac{1}{\delta}\right)\right\}$	$\frac{\sigma^2 \ell}{\varepsilon}\log\left(\frac{1}{\varepsilon}\right)$	$\frac{\varepsilon}{\sigma\ell\sqrt{\log(2/\varepsilon)}}$
α-sub-Weibull	$\varepsilon^{-2}\max\left\{\frac{\ell(\log(1/\varepsilon))^\alpha}{\varepsilon}\Gamma(1+\alpha), \log\left(\frac{1}{\delta}\right)\right\}$	$\frac{\ell}{\varepsilon}\left(\log\left(\frac{1}{\varepsilon}\right)\right)^{2\alpha}$	$\frac{\varepsilon}{\ell(\log(2c_\alpha/\varepsilon))^\alpha}$

Proof. Composing Lemma 21, Lemma 38, Theorem 20, and Theorem 22 yields the proof.

By Theorem 20, and Lemma 38, we have:

$$\begin{aligned} \left|\mathrm{TV}\left(\mu,\nu\right)-\mathrm{TV}\left(\mu_b^{Cont},\nu_b^{Cont}\right)\right| &\leq 2\varepsilon && \text{By Theorem 20 and Triangle Inequality} \\ \left|\mathrm{TV}\left(\mu,\nu\right)-\mathrm{TV}\left(\mu_b^{Disc},\nu_b^{Disc}\right)\right| &\leq 2\varepsilon && \text{By Lemma 38} \end{aligned} \tag{26}$$

From Lemma 21, and fixing the b, we have with probability $1-2\delta$:

$$\left|\mathrm{TV}\left(\mu_{b,n},\nu_{b,n}\right)-\mathrm{TV}\left(\mu_b^{Disc},\nu_b^{Disc}\right)\right| \leq 2\varepsilon \qquad \text{By Triangle Inequality} \tag{27}$$

From Theorem 22, we have with probability $1-2\delta$:

$$\left|\mathrm{TV}\left(\hat{\mu}_n,\hat{\nu}_n\right)-\mathrm{TV}\left(\mu_{b,n},\nu_{b,n}\right)\right| \leq 2\varepsilon \qquad \text{By Triangle Inequality} \tag{28}$$

A union bound argument and triangle inequality over Equation (26), (27) and (28) completes the proof. □

F Useful Technical Results

Lemma 39 (Inverse of Bi-Lipschitz function is Bi-Lipschitz)**.** *Given an invertible function $f:\mathrm{Dom}\, f\to\mathrm{Range}\, f$ that is ℓ bi-Lipschitz , the corresponding inverse function $f^{-1}:\mathrm{Range}\, f\to\mathrm{Dom}\, f$ is also bi-Lipschitz with parameter ℓ.*

The following corollary is directly implied from Lemma 39.

Corollary 40 (Bi-Lipschitz CDF implies Bi-Lipschitz Inverse CDF)**.** *A distribution satisfying Assumption 8 has bi-Lipschitz inverse CDF.*

Definition 41 (α-SubWeibull Distribution and Random Variable (Vladimirova et al., 2020))**.** A distribution μ is said to be α-SubWeibull if there exists some constant c_α for any $t\geq 0$, we have:

$$\Pr_{X\sim\mu}\left[X\geq t\right]\leq c_\alpha\exp\left(-t^{1/\alpha}\right) \tag{29}$$

Correspondingly, the random variable X drawn from μ is said to be a SubWeibull random variable.

Lemma 42 (MGF Charecterization of SubWeibull(Theorem 2.1 in Vladimirova et al. (2020)))**.** *Given X be a α-SubWeibull random variable, then the MGF of $|X|^{1/\alpha}$ satisfies:*

$$\exists c>0 \text{ such that } \mathbb{E}\left[\exp\left(\left(\gamma\left|X\right|\right)^{1/\alpha}\right)\right]\leq\exp\left(\left(\gamma c\right)^{1/\alpha}\right)$$

for all γ such that $0<\gamma\leq 1/c$.

Lemma 43 (True Measure Concentration (Cohen et al., 2020))**.** *For a measure μ defined over a simplex $\Delta_{\mathbb{N}}$ and corresponding empirical distribution μ_n generated by n samples, for any $\varepsilon,\delta\in(0,1)$, there exists a constant c such that $n\geq c\varepsilon^{-2}\max\left\{\sum_{i\in\mathbb{N}}\sqrt{p(i)},\log\left(1/\delta\right)\right\}$, we have with probability $1-\delta$,*

$$\Pr\left[\mathrm{TV}\left(\mu,\mu_n\right)\geq\varepsilon\right]\leq\delta$$

Lemma 44 (Concentration of Empirical Measure In Wasserstein Distance (Theorem 2 in Fournier & Guillin (2015)))**.** *Let μ be a distribution on $\mathbb{R}$. Then for all $p\in\mathbb{N}$,*

- *If $\exists\alpha<\frac{1}{p},\gamma>0$ such that $\mathbb{E}\left[\exp\left(\left(\gamma\left|X\right|\right)^{1/\alpha}\right)\right]<\infty$, then,*

$$\mathbb{P}\left[\mathcal{W}_1\left(\mu,\mu_n\right)\geq\varepsilon\right]\leq\exp\left(-cn\varepsilon^2\right)+\exp\left(-cn\varepsilon^{1/p\alpha}\right)$$

- *If $\exists\alpha\in\left(\frac{1}{p},\infty\right),\gamma>0$ such that $\mathbb{E}\left[\exp\left(\left(\gamma\left|X\right|\right)^{1/\alpha}\right)\right]<\infty$, then,*

$$\mathbb{P}\left[\mathcal{W}_1\left(\mu,\mu_n\right)\geq\varepsilon\right]\leq\exp\left(-cn\varepsilon^2\right)+\exp\left(-c\left(n\varepsilon\right)^{1/2p\alpha}\right)$$

The Lemma 44 implies the following two corollaries:

Corollary 45 (Concentration in Wasserstein distance of Empirical Measures over $\mathbb{R}$ (Bhat & L.A., 2019))**.** *Given an empirical measure μ_n generated by n i.i.d. samples generated from a subGaussian measure μ over $\mathbb{R}$, we have:*

$$\mathbb{P}\left[\mathcal{W}_1\left(\mu, \mu_n\right) \geq \varepsilon\right] \leq \exp\left(-cn\varepsilon^2\right)$$

Corollary 46. *Let μ be a distribution on $\mathbb{R}$ such that $\exists \alpha \in (1, \infty), \gamma > 0$ such that $\mathbb{E}\left[\exp\left(\left(\gamma |X|\right)^{1/\alpha}\right)\right] < \infty$, then,*

$$\mathbb{P}\left[\mathcal{W}_1\left(\mu, \mu_n\right) \geq \varepsilon\right] \leq \exp\left(-cn\varepsilon^2\right) + \exp\left(-c\left(n\varepsilon\right)^{1/2\alpha}\right)$$

Lemma 47 (Strong Demographic Parity and $\mathcal{W}_1\left(,\right)$ distance (Jiang et al., 2020))**.** *Let $f : \mathbb{R}^d \times [k] \to [0, 1]$ be a function where $[k]$ denotes the sensitive attribute. Let μ_s denote the output distribution of f corresponding to the sensitive attribute $s \in [k]$. Then, we have for all $s, s' \in [k]$:*

$$\mathop{\mathbb{E}}_{t \sim \mathrm{U}[0,1]}\left[\mathbb{P}\left[f(X, s) \geq t\right] - \mathbb{P}\left[f(X, s') \geq t\right]\right] = \mathcal{W}_1\left(\mu_s, \mu_{s'}\right)$$

Lemma 48 (Privacy and Hockey Stick Divergence(Balle et al., 2018))**.** *For a given $\alpha \in \mathbb{R}$, a mechanism $\mathcal{M}$ is (α, β)-differentially private if for all $\mathbf{X},\ \mathbf{X}'$ with $\mathrm{Ham}\left(\mathbf{X},\ \mathbf{X}'\right) = 1$:*

$$\mathrm{H}_{\mathrm{e}^\alpha}\left(\mathcal{M}\left(\mathbf{X}\right) || \mathcal{M}\left(\mathbf{X}'\right)\right) \leq \beta$$

where $\mathrm{H}_{\mathrm{e}^\alpha}\left(\mu || \nu\right)$ is called the Hockeystick divergence between μ and ν.

Lemma 49 (From HSD to TV Approximation (Koskela & Mohammadi, 2024))**.** *Given two measured μ, ν and their TV distance approximations $\hat{\mu}_n, \hat{\nu}_n$ satisfying $\mathrm{TV}\left(\mu, \hat{\mu}_n\right) \leq \varepsilon$, and $\mathrm{TV}\left(\hat{\mu}_n, \hat{\nu}_n\right) \leq \varepsilon$; we have for all $\alpha \in \mathbb{R}$:*

$$\mathrm{H}_{\mathrm{e}^\alpha}\left(\mu || \nu\right) \leq \mathrm{H}_{\mathrm{e}^\alpha}\left(\hat{\mu}_n || \hat{\nu}_n\right) + \left(1 + e^\alpha\right)\varepsilon$$

G Experimental Details - Fairness and Privacy Auditing

G.1 Fairness Auditing

Group fairness metrics in ML model's prediction measure the disparity in predictions of ML models across different subpopulations. The subpopulations correspond to a sensitive attribute (e.g. gender, economic status, ethnicity etc.) on which they should not be discriminated. A popular group fairness measure is demographic parity (Feldman et al., 2015).

Definition 50 (Demographic Parity)**.** Given $O \subseteq \mathbb{R}$ and sensitive attributes $[k]$, a regression function $f : \mathbb{R}^d \times [k] \to O$ satisfies demographic parity if for all $s, s' \in [k]$, we have $\sup_{t \in O} \mathbb{P}(f(X, s) \leq t) - \mathbb{P}(f(X, s') \leq t) = 0$. For classification, $S = [0, 1]$ and the definition is referred as strong demographic parity (Jiang et al., 2020).

Jiang et al. (2020) show that bounding strong demographic parity is equivalent parity is equivalent to bounding the maximum of 1-Wasserstein distance between the output distributions of any two subpopulations. This motivates us to use `SWA` to estimate demographic parity for ML models.

Experimental Setup: We test accuracy and sublinearity of `SWA` for fairness auditing on the well-known ACS Income dataset (Ding et al., 2021). We test both on linear regression ($d = 10$) for income as output and classification with logistic regression ($d = 10$) for income above and below 40000 USD as outputs. We compute Wasserstein Distance between the distribution of outputs of linear and logistic regression models for male and female data points of a model trained on ACS_Income data. We use $3:1$ train-test split for both cases. We use scikit-learn (Pedregosa et al., 2011) to train both the models, and use the 'liblinear' solver for logistic regression. For reference, we compute the distance between the bucketed versions of the

output distribution exactly. We choose the bucket size to be 10 and 0.01, for the regression and classification tasks, respectively. For the regression task, we increase the number of buckets as $\{500, 1000, \ldots, 20000\}$; and for the classification task, we increase the number of buckets as $\{50, 100, \ldots, 1250\}$. Finally we report the multiplicative approximation error of our estimates w.r.t. the true distance in both the cases with increasing number of buckets. We run each of the experiments 50 times.

Results: Sublinearity of `SWA` in Fairness Auditing: We use 416625 total samples consisting of almost half male and female samples each. The sample streams arrive via $S = 10$ sources. For the linear regression model, the approximation error drops below 0.1 when we use 12500 buckets. For the logistic regression model, the approximation error drops below 0.1 when we use 750 buckets. The difference in #buckets is due to the difference in variance and the width of buckets used in each case.

G.2 Privacy Auditing

Differential privacy (Dwork, 2006) is now considered as the gold standard for data privacy protection. It aims to keep an input datum indistinguishable while looking into outputs of an algorithm.

Definition 51 ((α, β)**- Differential Privacy** (Dwork, 2006))**.** An algorithm $f : \mathcal{X} \to \mathcal{Y}$ is (α, β)-differentially private if for any $\mathbf{X}$, $\mathbf{X}'$ with $\text{Ham}\left(\mathbf{X}, \mathbf{X}'\right) = 1$ and $\forall S \subseteq \mathcal{Y}$, we have $\mathbb{P}[f(\mathbf{x}) \in S] \leq e^{\alpha}\,\mathbb{P}[f(\mathbf{x}') \in S] + \beta$.

An equivalent representation of differential privacy is $\mathrm{H}_{\mathrm{e}^{\alpha}}\left(\mathcal{M}\left(\mathbf{X}\right) || \mathcal{M}\left(\mathbf{X}'\right)\right) \leq \beta$ (Balle et al., 2018), where Hockey Stick Divergence (HSD) is defined as $\mathrm{H}_{\tau}\left(\mu || \nu\right) \triangleq \int_{\mathcal{X}} [\mu(x) - \tau\nu(x)]_{+} \mathrm{d}x$. Recently, Koskela & Mohammadi (2024) show that estimating the HSD of two distributions is equivalent to estimating the HSD and TV distance between their empirical counterparts. They construct histograms over outputs of a black-box auditor and use this result to estimate TV distances for privacy auditing. As auditing privacy is data intensive, it motivates us to use `STVA` in this setting. We adopt the experimental setting of Annamalai & Cristofaro (2024) and compute the TV distance between the output distributions of logistic regressors trained on neighbouring datasets, say IN and OUT, sampled from MNIST (LeCun, 1998).

Experimental Setup: We study the performance of `STVA` in the case of privacy auditing. We generate losses for datasets with and without canary using the work of Annamalai & Cristofaro (2024). We run logistic regression on `MNIST` dataset with 50 epochs and $\alpha = 10$. We denote the losses with and without canary to be `lossesin` and `lossesin`, respectively. We obtain 1000 samples of losses for both the cases and take these losses as the distribution of interest. We choose the bucket size to be . The mean of `lossesin` and `lossesout` are -2.3194 and -2.3241, respectively. The standard deviation of `lossesin` and `lossesout` are 0.005654 and 0.005535, respectively. We increase the number of buckets as $\{10, 20, \ldots, 100\}$. Finally we report the multiplicative approximation error of our estimates w.r.t. the true distance in both the cases with increasing number of buckets. We run each of the experiments 50 times.

Results: Given 1000 samples, Figure 6 show that `STVA` computes the TV distance between output distributions for IN and OUT datasets. The sample streams arrive via $S = 10$ sources. The approximation error drops below 0.1 while using only 250 buckets. This shows `STVA` can conduct resource-efficient privacy auditing of large-scale datasets.

G.3 Enlarged Plots

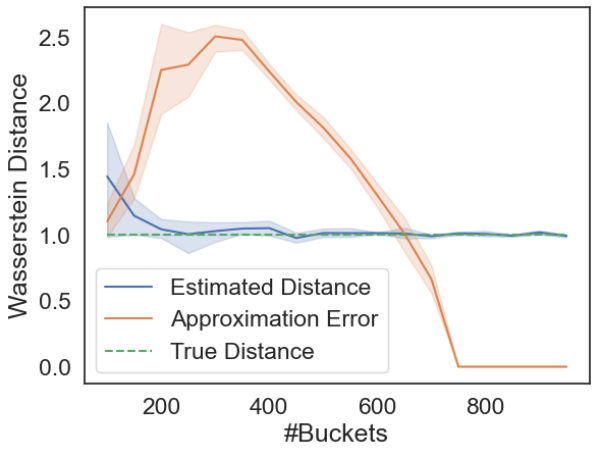

Figure 7: Performance of `SWA` with $\mathcal{N}(0,5)$, $\mathcal{N}(1,5)$ and $b = 0.05$

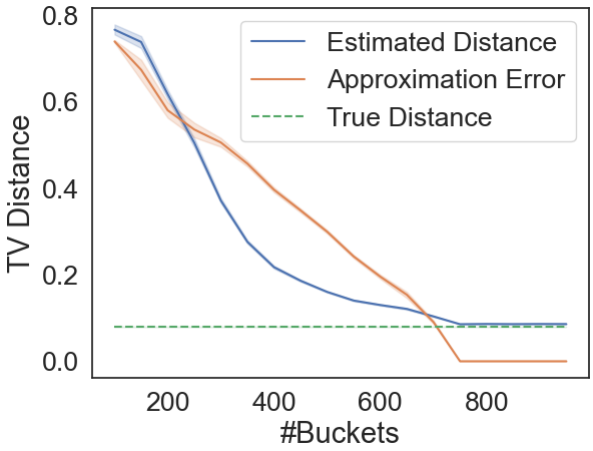

Figure 8: Performance of `STVA` with $\mathcal{N}(0,5)$, $\mathcal{N}(1,5)$ and $b = 0.05$

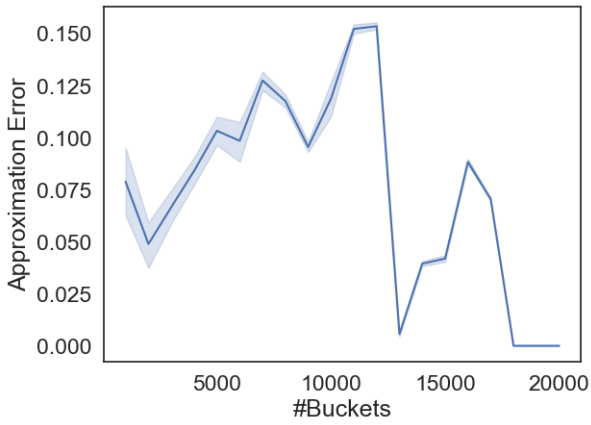

Figure 9: Auditing with `SWA` on regression output of ACS_Income

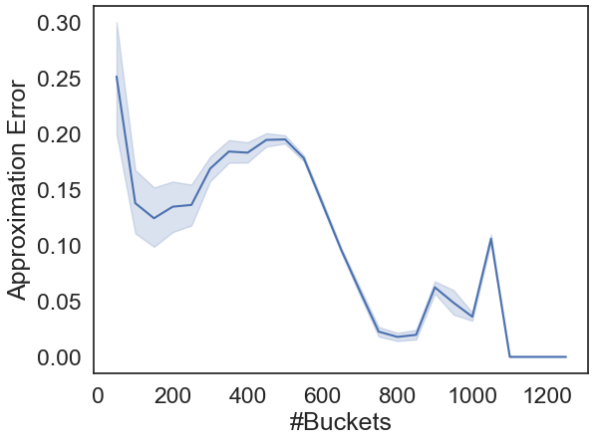

Figure 10: Auditing with `SWA` on classification output of ACS_Income

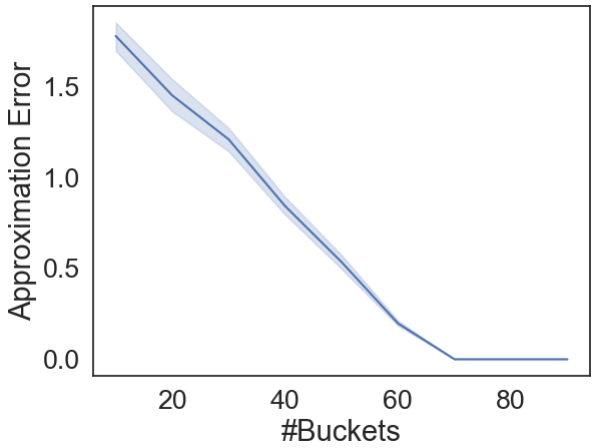

Figure 11: Privacy Auditing of logistic regression on MNIST.