An Improved Algorithm for Learning Drifting Discrete Distributions

Alessio Mazzetto
Brown University

Abstract

We present a new adaptive algorithm for learning discrete distributions under distribution drift. In this setting, we observe a sequence of independent samples from a discrete distribution that is changing over time, and the goal is to estimate the current distribution. Since we have access to only a single sample for each time step, a good estimation requires a careful choice of the number of past samples to use. To use more samples, we must resort to samples further in the past, and we incur a drift error due to the bias introduced by the change in distribution. On the other hand, if we use a small number of past samples, we incur a large statistical error as the estimation has a high variance. We present a novel adaptive algorithm that can solve this trade-off without any prior knowledge of the drift. Unlike previous adaptive results, our algorithm characterizes the statistical error using data-dependent bounds. This technicality enables us to overcome the limitations of the previous work that require a fixed finite support whose size is known in advance and that cannot change over time. Additionally, we can obtain tighter bounds depending on the complexity of the drifting distribution, and also consider distributions with infinite support.

1 INTRODUCTION

Estimating a distribution from a set of samples is a crucial challenge in data analysis and statistics (Devroye and Györfi, 1987; Silverman, 1986; Devroye and Lugosi, 2001). In this work, we focus on the classical setting of estimating the probability mass function of a discrete distribution. A long list of work characterized the error for this estimation problem given *independent* and *identically distributed* (i.i.d.) samples from the same distribution (Han et al., 2015; Kamath et al., 2015; Orlitsky and Suresh, 2015; Jiao et al., 2015; Cohen et al., 2020). The use of the total variation metric as a measure of error for this problem is a natural choice that is commonly adopted in the literature (Devroye and Lugosi, 2001). It is folklore that if a distribution has support size k, the maximum likelihood estimator with n i.i.d. samples has an expected error upper bounded by $O(\sqrt{k/n})$, which can be shown to be tight (Anthony et al., 1999).

In numerous applications where samples are collected over time, it is possible that their underlying distribution may change. In the *distribution drift* setting, we are interested in estimating the current distribution, given a sequence of past samples that could be generated by different distributions. This setting has been recently studied by Mazzetto and Upfal (2023b), but other problems have also been considered in a similar setting, for example, binary classification (e.g., Barve and Long, 1996; Long, 1998, and references therein), agnostic learning (Mohri and Muñoz Medina, 2012; Hanneke and Yang, 2019; Mazzetto and Upfal, 2023a), or crowdsourcing (Fu et al., 2020).

Concretely, let $X_1, \ldots, X_T$ denote a sequence of T *independent* samples respectively from the discrete distributions $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_T$. Equivalently, we say that the sequence of samples is generated over time by a *drifting* discrete distribution. The goal is to estimate the current distribution $\boldsymbol{\mu}_T$ given the sequence of samples. Without loss of generality, it is sufficient to consider discrete distributions over the natural numbers, and given such a distribution $\boldsymbol{\mu}$, we let $\boldsymbol{\mu}(i) = \Pr_{\boldsymbol{\mu}}(X = i)$ for any $i \in \mathbb{N}$. The total variation distance between two discrete distributions $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ is defined as $\|\boldsymbol{\mu} - \boldsymbol{\eta}\|_{\mathrm{TV}} = (1/2)\sum_{i\in\mathbb{N}} |\boldsymbol{\mu}(i) - \boldsymbol{\eta}(i)|$.

Indeed, the estimation of $\boldsymbol{\mu}_T$ is possible only if the previous distributions are related to it. Let $\Delta_1, \ldots, \Delta_T$ be a non-decreasing sequence of real numbers where

$$\Delta_r \doteq \max_{t:0\leq t<r} \|\boldsymbol{\mu}_T - \boldsymbol{\mu}_{T-t}\|_{\mathrm{TV}} \ . \tag{1}$$

Proceedings of the 27th International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238. Copyright 2024 by the author(s).

The value Δ_r is the maximum total variation distance from the current distribution $\boldsymbol{\mu}_T$ to any of the most recent r distributions $\boldsymbol{\mu}_{T-r+1}, \ldots, \boldsymbol{\mu}_T$. In past work (Mazzetto and Upfal, 2023b), it is shown that if the distributions have the same support of size k, a tight lower bound on the expected error of *any* algorithm for the estimation of the current distribution $\boldsymbol{\mu}_T$ with respect to the total variation distance is given by

$$\Omega\left(\min_{1\leq r\leq T}\left[\sqrt{\frac{k}{r}} + \Delta_r\right]\right) \quad . \tag{2}$$

This lower bound is in a minimax sense. Formally, given any (possibly adaptive) algorithm, and a sufficiently small non-decreasing sequence of values $\Delta_1, \ldots, \Delta_T$, there exists a sequence of distributions $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_T$ with shared support k such that $\max_{t:0\leq t<r}\|\boldsymbol{\mu}_{T-t} - \boldsymbol{\mu}_T\|_{\mathrm{TV}} \leq \Delta_r$ for any $1 \leq r \leq T$, and the expected estimation error of the algorithm is at least (2).

For a fixed integer r, the quantity $O(\sqrt{k/r} + \Delta_r)$ of (2) is also an upper bound to the expected error obtained by only using the most recent r samples for the estimation, and it is written as the sum of the upper bound to two errors: the *statistical error* and the *drift error*. The statistical error term $\sqrt{k/r}$ is related to the variance of the estimation, and it decreases by considering more samples; whereas the drift error term Δ_r is due to the distribution drift, and it can potentially increase by using samples further away in time. Equation (2) shows that an optimal estimation is given by the optimal solution of this trade-off. This trade-off determines an optimal number of recent samples to use as a function of the distribution drift. This is a significant difference with the i.i.d. setting, where each sample provides useful information, and the expected error goes to 0 as the number of samples goes to infinity.

The trade-off between the statistical error and the drift error is common in the literature on learning with drift (Mohri and Muñoz Medina, 2012; Mazzetto and Upfal, 2023b). However, the minimization of this trade-off is challenging as the values $\Delta_1, \ldots, \Delta_T$ of the drift error are unknown, and they cannot be estimated from the data since we only have access to a single sample from each distribution. For this reason, most of the previous work required prior knowledge of the magnitude of the drift in order to quantify and solve this trade-off. For example, a common assumption is that the magnitude of the drift is bounded by $\Delta > 0$ at each step (Bartlett, 1992), which implies $\Delta_r \leq (r-1)\Delta$ for all $r \leq T$. In this case, the optimal solution of the trade-off gives an estimation error equal to $\Theta((k \cdot \Delta)^{1/3})$ which is achieved by computing the empirical distribution induced by the most recent $\Theta((k/\Delta^2)^{1/3})$ samples.

1.1 Limitations of Existing Work

In recent work, Mazzetto and Upfal (2023a) exhibit a general learning algorithm that solves the trade-off between statistical error and drift error (up to loglog factors) based on the input samples and without any prior knowledge of the drift. In our setting, this implies that there exists an algorithm that can adaptively attain the lower bound (2), which cannot be improved in a minimax sense. Precisely, there exists an algorithm that observes the sequence $X_1, \ldots, X_T$ from a drifting distribution with a fixed support of size k, and it returns an estimate $\hat{\boldsymbol{\mu}}$ of $\boldsymbol{\mu}_T$ such that with probability at least 0.99:

$$\|\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}\|_{\mathrm{TV}} = O\left(\min_{1\leq r\leq T}\left[\sqrt{\frac{k}{r}} + \sqrt{\frac{\log\log r}{r}} + \Delta_r\right]\right) \tag{3}$$

While the above error is essentially tight according to (2), this result has several intertwined weaknesses.

First, the previous adaptive algorithm can only be applied to drifting distributions that have finite support, which cannot change over time. This constraint is in conflict with the drift setting, where the distribution, hence its support, can indeed change over time. For example, consider the drifting distribution of the items purchased over time from an online retailer: the support of this distribution can repeatedly evolve due to changes in inventory, new products, or availability. Additionally, we are also required to know the value k of the support size. Since k cannot be determined precisely using only the input samples, it is necessary to have prior knowledge of this value, which can be unfeasible in many practical applications.

Second, the aforementioned algorithm has the crucial shortcoming of using a distribution-independent upper bound $O(\sqrt{k/r})$ on the statistical error from using r samples. While this upper bound is indeed tight for distributions that are roughly uniform over a support of size k, the actual error due to the variance of the estimation can be significantly smaller for other distributions. As an example, if we consider a distribution of size k, where most of the probability mass is concentrated in $k' \ll k$ elements, we would expect a statistical error with rate $O(\sqrt{k'/r})$. Furthermore, the use of a distribution-independent upper bound on the statistical error prevents the consideration of distributions with infinite support ($k = \infty$). In the i.i.d. setting, it is possible to address this issue by using a sharp distribution-dependent upper bound on the statistical error, which also allows us to handle distributions with infinite support (Berend and Kontorovich, 2013; Cohen et al., 2020). In particular, if we assume that there is no drift, i.e. $\boldsymbol{\mu} = \boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_T$, the maximum likelihood estimator $\hat{\boldsymbol{\mu}}$ over T samples exhibits an expected error (Berend and Kontorovich, 2013):

$$\frac{1}{8}\Lambda_T(\boldsymbol{\mu}) - \frac{1}{4\sqrt{T}} \leq \mathbb{E}\|\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}\|_{\mathrm{TV}} \leq \Lambda_T(\boldsymbol{\mu}) \ , \qquad (4)$$

where

$$\Lambda_T(\boldsymbol{\mu}) \doteq \sum_{i:\boldsymbol{\mu}(i)<1/T} \boldsymbol{\mu}(i) + \frac{1}{\sqrt{T}} \sum_{i:\boldsymbol{\mu}(i)\geq 1/T} \sqrt{\boldsymbol{\mu}(i)}. \quad (5)$$

The value $\Lambda_T(\boldsymbol{\mu})$ is a measure of the *learning complexity* of $\boldsymbol{\mu}$ that provides a tight characterization of the variance of the estimation of $\boldsymbol{\mu}$ with T samples. By using the Cauchy-Schwarz inequality, it is simple to verify that if the support of $\boldsymbol{\mu}$ has size k, it holds that $\Lambda_r(\boldsymbol{\mu}) \leq \sqrt{k/r}$, recovering the aforementioned distribution-independent upper bound. We highlight that having a tight bound on the statistical error is especially important in a drift setting as it can significantly impact the quality of the estimation, since the value of this bound determines the number of samples to use in order to solve the trade-off between statistical error and drift error.

2 MAIN RESULT

In our work, we address the issues outlined in the previous section. Our main contribution is to provide an adaptive algorithm for estimating an *arbitrary* drifting discrete distribution. The result is formalized as follows.

Theorem 2.1. *Let $\delta \in (0,1)$. There exists an algorithm that given $X_1, \ldots, X_T$, it outputs a distribution $\hat{\boldsymbol{\mu}}$ such that with probability at least $1-\delta$, it holds that*

$$\|\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}\|_{\mathrm{TV}} = O\Bigg(\min_{1\leq r\leq T} \Bigg[\Lambda_r(\boldsymbol{\mu}_T) + \sqrt{\frac{\log((\log^2 r + 1)/\delta)}{r}} + \Delta_r\Bigg]\Bigg) \ ,$$

where $\Delta_r = \max_{0\leq t<r}\|\boldsymbol{\mu}_T - \boldsymbol{\mu}_{T-t}\|_{\mathrm{TV}}$ as in (1).

The above theorem shows that there exists an adaptive algorithm that can achieve (up to loglog factors) an optimal solution of the trade-off between statistical error and drift error, where the statistical error is quantified using a distribution-dependent measure of complexity. Compared to the previous adaptive result (3), our algorithm works for an *arbitrary* discrete drifting distribution, and it utilizes a sharper distribution-dependent upper bound on the statistical error that we estimate from the data. In particular, our algorithm does not require any prior knowledge of the drifting distribution, and it also works for drifting distribution with support that changes over time or with infinite support. For the special case of drifting distributions with shared support of size k, it holds that $\Lambda_r(\boldsymbol{\mu}_T) \leq \sqrt{k/r}$, and we indeed achieve the lower bound (2) up to logarithmic terms. However, we highlight that our algorithm uses a distribution-dependent measure $\Lambda_r(\boldsymbol{\mu}_T)$ to upper bound the statistical error of the estimation which could be significantly tighter than $\sqrt{k/r}$ even for distributions with support size k. This enables us to obtain tighter bounds depending on the drifting distribution's complexity.

Since $\Lambda_r(\boldsymbol{\mu}) \leq \Lambda_s(\boldsymbol{\mu})$ for any $s < r$, in the i.i.d. case, we can observe that the theorem guarantees with high-probability (e.g., $\geq 99/100$) an estimation such that

$$\|\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}\|_{\mathrm{TV}} = O\left(\Lambda_T(\boldsymbol{\mu}_T) + \sqrt{\frac{\log\log^2 T}{T}}\right) \ ,$$

retrieving up to logarithmic terms the tight characterization of the estimation error depicted in (4) for learning with T i.i.d. samples from $\boldsymbol{\mu}_T$. The additional logarithmic term is the cost of the adaptivity, since the algorithm does not know a priori whether the samples are identically distributed.

Technical contribution. It is not straightforward to extend the adaptive strategy of the previous work (Mazzetto and Upfal, 2023a) to use a data-dependent bound on the statistical error. In fact, the proof strategy of that work relies on knowing the exact rate at which the upper bound on the statistical error decreases, which is not possible for distribution-dependent upper bounds. To circumvent this issue, we develop a new analysis for learning with drift, whose proof of correctness also uses a novel result that ties the magnitude of the drift with the change in the learning complexity of the drifting distribution. We believe that our novel proof strategy for learning with drift using data-dependent upper bounds on the statistical error can also be applied to other learning problems. To the best of our knowledge, this is the first adaptive learning algorithm for discrete distributions to use data-dependent bounds in the drift setting.

3 ALGORITHM

In this section, we present the algorithm that achieves the guarantee of Theorem 2.1. First, we formally define the trade-off between statistical error and drift error obtained by using the most recent r samples. To this end, for any $1 \leq r \leq T$, we define the following distributions

$$\boldsymbol{\mu}_T^{[r]}(i) = \frac{1}{r}\sum_{t=T-r+1}^{T} \boldsymbol{\mu}_t(i) \qquad \forall i \in \mathbb{N} \ ,$$

$$\hat{\boldsymbol{\mu}}_T^{[r]}(i) = \frac{1}{r}\sum_{t=T-r+1}^{T} \mathbf{1}_{\{X_t=i\}} \qquad \forall i \in \mathbb{N} \ ,$$

which are respectively the average distribution $\boldsymbol{\mu}_T^{[r]}$ and the empirical distribution $\hat{\boldsymbol{\mu}}_T^{[r]}$ over the most recent r samples. Following the methodology of previous work, the following proposition provides an error decomposition into statistical error and drift error for the estimation of $\boldsymbol{\mu}_T$ by using the empirical distribution $\hat{\boldsymbol{\mu}}_T^{[r]}$.

Proposition 3.1. *Let $1 \le r \le T$. We have that*

$$\|\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_T^{[r]}\|_{\mathrm{TV}} \le \underbrace{\|\boldsymbol{\mu}_T^{[r]} - \hat{\boldsymbol{\mu}}_T^{[r]}\|_{\mathrm{TV}}}_{\text{Statistical Error}} + \underbrace{\Delta_r}_{\text{Drift Error}} \ .$$

Proof. By using the triangle inequality, we have that

$$\|\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_T^{[r]}\|_{\mathrm{TV}} \le \|\boldsymbol{\mu}_T^{[r]} - \hat{\boldsymbol{\mu}}_T^{[r]}\|_{\mathrm{TV}} + \|\boldsymbol{\mu}_T - \boldsymbol{\mu}_T^{[r]}\|_{\mathrm{TV}}$$

We can upper bound the second addend of the right-hand side above using the triangle inequality

$$\begin{aligned}\|\boldsymbol{\mu}_T^{[r]} - \boldsymbol{\mu}_T\|_{\mathrm{TV}} &= \left\| \frac{1}{r}\sum_{t=T-r+1}^{T} (\boldsymbol{\mu}_t - \boldsymbol{\mu}_T) \right\|_{\mathrm{TV}} \\ &\le \frac{1}{r}\sum_{t=T-r+1}^{T} \|\boldsymbol{\mu}_t - \boldsymbol{\mu}_T\|_{\mathrm{TV}} \le \Delta_r \ . \qquad (6)\end{aligned}$$

□

The first term of the right-hand side of this proposition represents the statistical error of the estimation. Our goal is to measure this error as a function of the distribution-dependent measure of complexity $\Lambda_r(\boldsymbol{\mu}_T)$ of the current distribution $\boldsymbol{\mu}_T$, which is unknown. Our algorithm relies on the input data to estimate this quantity. Following the example of previous work, we use the empirical counterpart of this measure of complexity. Given an empirical distribution $\hat{\boldsymbol{\mu}}_T^{[r]}$, we define

$$\Phi_r(\hat{\boldsymbol{\mu}}_T^{[r]}) \doteq \sqrt{\frac{\|\hat{\boldsymbol{\mu}}_T^{[r]}\|_{\frac{1}{2}}}{r}} = \frac{1}{\sqrt{r}}\sum_{i\in\mathbb{N}} \sqrt{\hat{\boldsymbol{\mu}}_T^{[r]}(i)} \ ,$$

and we observe that this quantity can be computed from the samples. The quantity $\Phi_r(\hat{\boldsymbol{\mu}}_T^{[r]})$ is an empirical measure of complexity that provides an upper bound to the statistical error with r samples. The next proposition is proven based on results of previous work (Cohen et al., 2020).

Proposition 3.2. *Let $\delta \in (0,1)$ and $1 \le r \le T$. With probability at least $1-\delta$, it both holds:*

$$\|\hat{\boldsymbol{\mu}}_T^{[r]} - \boldsymbol{\mu}_T^{[r]}\|_{\mathrm{TV}} \le \Phi_r\left(\hat{\boldsymbol{\mu}}_T^{[r]}\right) + 3\sqrt{\frac{\log(4/\delta)}{2r}}$$

and

$$\Phi_r\left(\hat{\boldsymbol{\mu}}_T^{[r]}\right) \le 4\Lambda_r\left(\boldsymbol{\mu}_T^{[r]}\right) + \sqrt{\frac{\log(4/\delta)}{r}}$$

Proof. In the appendix. □

This proposition enables us to use the empirical measure of complexity $\Phi_r\left(\hat{\boldsymbol{\mu}}_T^{[r]}\right)$ to estimate the statistical error while guaranteeing that we obtain a result that is a tight approximation to using the non-empirical measure $\Lambda_r\left(\boldsymbol{\mu}_T^{[r]}\right)$ as in (4). Proposition 3.2 can be seen as a generalization of the results of Cohen et al. (2020, Theorem 2.1 and Theorem 2.3) for independent but not identically distributed discrete distributions.

On a high level, the core of our algorithm is a condition that allows us to compare the upper bound to the estimation error induced by different choices of the number of past samples *without* explicitly estimating the drift error. For ease of notation, we let $r_j \doteq 2^j$ for any $j \ge 0$. As long as this condition is true, the algorithm iteratively considers a larger number of past samples, also referred to as *window size*, starting from $r_0 = 1$. In particular, at iteration j it considers a window size r_j (starting from $j = 0$), and it evaluates the condition by comparing the estimation obtained with r_j to the estimation obtained with $r_0, \ldots r_{j-1}$. If the condition is satisfied, we can provably maintain a solution whose upper bound to the estimation error is up to a constant factor as good as any previously considered window size (Proposition 3.4). Conversely, if the condition is violated, a non-negligible drift has occurred, thus using more samples cannot yield a significantly better estimation (Proposition 3.5). As we will see, depending on the data, it is possible that we do not need to consider all the possible window sizes $\{r_j : j \ge 0\}$ for the evaluation of this condition, and only a subset will suffice.

The correctness of our algorithm is conditioned on the event that the estimation for all window sizes $r_j = 2^j$ for $j \ge 0$ concentrates around its expectation, i.e. we want to provide an upper bound to the statistical error as in Proposition 3.2 for all those window sizes. This result is formalized in the next proposition and it is obtained by simply taking a union bound.

Proposition 3.3. *Let $\delta \in (0,1)$. With probability at*

Algorithm 1: Adaptive Learning Algorithm For A Discrete Drifting Distribution

```
$L = \{0\}$                                                              1
for $j = 1, \dots, \lfloor \log_2 T \rfloor$ do                          2
    if $\xi_{r_j} < \min_{\ell \in L} \xi_{r_\ell}$ then                 3
        for $\ell \in L$ do                                              4
            if $\|\hat{\boldsymbol{\mu}}_T^{[r_\ell]} - \hat{\boldsymbol{\mu}}_T^{[r_j]}\|_{\mathrm{TV}} \ge 3\xi_{r_\ell} + \xi_{r_j}$ then   5
                return $\hat{\boldsymbol{\mu}}_T^{[r_{\max L}]}$         6
            end                                                          7
        end                                                              8
        $L \leftarrow L \cup \{j\}$                                      9
    end                                                                  10
end                                                                      11
return $\hat{\boldsymbol{\mu}}_T^{[r_{\max L}]}$                         12
```

least $1-\delta$, for all $j \geq 0$ it holds that

$$\|\hat{\boldsymbol{\mu}}_T^{[r_j]} - \boldsymbol{\mu}_T^{[r_j]}\|_{\mathrm{TV}} \leq \Phi_{r_j}\left(\hat{\boldsymbol{\mu}}_T^{[r_j]}\right) + 3\sqrt{\frac{\log\left(\frac{c(\log^2 r_j + 1)}{\delta}\right)}{r_j}} \tag{7}$$

and

$$\Phi_{r_j}\left(\hat{\boldsymbol{\mu}}_T^{[r_j]}\right) \leq 4\Lambda_{r_j}\left(\boldsymbol{\mu}_T^{[r_j]}\right) + 3\sqrt{\frac{\log\left(\frac{c(\log^2 r_j + 1)}{\delta}\right)}{r_j}} \tag{8}$$

where c is a constant equal to $c = 4\pi^2/3$.

Proof. Let $\delta_j = \delta(6/\pi^2)/(j+1)^2$. We have that with probability at least $1-\delta_j$, the event of Proposition 3.2 holds with error probability δ_j and window size r_j. The statement follows by substituting the definition of δ_j, and taking a union bound over all possible events for $j \geq 0$, since $\sum_{j \geq 0} \delta_j = (6/\pi^2)\delta \sum_{j \geq 1} 1/j^2 = \delta$. □

The value $\delta \in (0,1)$ of the above proposition is a parameter of the algorithm, and it denotes its failure probability. Throughout this section, we assume that the event of Proposition 3.3 holds, otherwise our algorithm fails (with probability $\leq \delta$). We denote with

$$\xi_{r_j} \doteq \Phi_{r_j}\left(\hat{\boldsymbol{\mu}}_T^{[r_j]}\right) + 3\sqrt{\frac{\log(c(\log^2 r_j + 1)/\delta)}{r_j}}$$

the upper bound to the statistical error for the window size r_j given by the empirical bound (7) in Proposition 3.3. By using Proposition 3.1, we have the following upper bound to the error of estimating $\boldsymbol{\mu}_T$ by using the empirical distribution $\hat{\boldsymbol{\mu}}_T^{[r_j]}$:

$$U(r_j) \doteq \xi_{r_j} + \Delta_{r_j} \geq \|\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_T^{[r_j]}\|_{\mathrm{TV}}, \quad \forall j \geq 0. \tag{9}$$

The algorithm is looking for the window size r_j that minimizes the upper bound $U(r_j)$. We remark that this is a challenging problem due to the fact that the drift error is unknown and it cannot be estimated from the data, since we have access to a single sample from each distribution. Additionally, the sequence $(\xi_j)_{j \geq 0}$ is not necessarily strictly decreasing, and it does not have an analytical closed formula that depends only on r_j. This is a requirement of the proof strategy of the previous adaptive work Mazzetto and Upfal (2023a), which exploits the a priori knowledge of how much the statistical error is reduced by doubling the window size. Therefore, their proof strategy does not apply to our setting.

By an inspection of $U(\cdot)$, it is clear that if $\xi_{j+n} > \xi_j$ for some $n \geq 1$, then $U(r_j) < U(r_{j+n})$ since the drift error is non decreasing with respect to the window size. Thus, we can only consider a sequence of window sizes for which the upper bound to the statistical error is decreasing. Formally, we consider the following sequence of indexes $L = (\ell_1, \dots, \ell_K)$ where $K \leq 1 + \log_2(T)$ that is built iteratively as follows. We let $\ell_1 = 0$. Given $\ell_1, \dots, \ell_i$, the element ℓ_{i+1} is added (if it exists) as the first index $j > \ell_i$ such that $\xi_j < \xi_{\ell_i}$, thus we have $\xi_{\ell_{i+1}} < \xi_{\ell_i}$ for any $1 \leq i \leq K-1$. Moreover, given $j \geq 0$, we let $\gamma(j) = \ell$ be the largest value $\ell \in L$ such that $\ell \leq j$. Observe that by construction, the following relation holds:

$$U(r_{\gamma(j)}) \leq U(r_j) \ . \tag{10}$$

Therefore, we have $\min_{\ell \in L} U(r_\ell) \leq \min_{1 \leq j \leq T} U(r_j)$, and we can only consider window sizes r_ℓ with values $\ell \in L$.

The pseudo-code of the algorithm is reported in Algorithm 1. The algorithm iteratively builds the list L. Once a new element j that belongs to this list is found (Line 3), the algorithm compares the empirical distribution $\hat{\boldsymbol{\mu}}_T^{[r_j]}$ with all the empirical distributions $\hat{\boldsymbol{\mu}}_T^{[r_\ell]}$ with $\ell \in L$ such that $\ell < j$. This comparison provides the iteration condition that is the core of our algorithm. The statistical error ξ_{r_j} using r_j samples is less than ξ_{r_ℓ} for all such ℓ, however the drift error could be larger. The main idea is that if all those empirical distributions are sufficiently close (Line 5), then their distance with respect to $\boldsymbol{\mu}_T$ cannot be that large, and we can guarantee that the estimation error obtained by using $\hat{\boldsymbol{\mu}}_T^{[r_j]}$ is as good as the estimation achieved with any $\hat{\boldsymbol{\mu}}_T^{[r_\ell]}$ for all such ℓ. In this case, we can keep iterating. This intuition is formalized with the following proposition.

Proposition 3.4. *Assume that the event of Proposition 3.3 holds. Let $j \in L$. If*

$$\|\hat{\boldsymbol{\mu}}_T^{[r_\ell]} - \hat{\boldsymbol{\mu}}_T^{[r_j]}\|_{\mathrm{TV}} \leq 3\xi_{r_\ell} + \xi_{r_j} \quad \forall \ell < j : \ell \in L$$

then, we have that:

$$\|\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_T^{[r_j]}\|_{\mathrm{TV}} \le 5 \cdot \min_{\substack{\ell \in L:\\ \ell < j}} U(r_\ell) \ .$$

Proof. Let $\ell \in L$ such that $\ell < j$. By using the triangle inequality, we have that:

$$\|\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_T^{[r_j]}\|_{\mathrm{TV}} \le \|\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_T^{[r_\ell]}\|_{\mathrm{TV}} + \|\hat{\boldsymbol{\mu}}_T^{[r_\ell]} - \hat{\boldsymbol{\mu}}_T^{[r_j]}\|_{\mathrm{TV}} \ .$$

We upper bound the first term of the right-hand side using (9), and the second term by using the assumption of this proposition. We obtain:

$$\begin{aligned}\|\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_T^{[r_j]}\|_{\mathrm{TV}} &\le \xi_{r_\ell} + \Delta_{r_\ell} + \xi_{r_\ell} + 3\xi_{r_j} \\ &\le 5\xi_{r_\ell} + \Delta_{r_\ell} \\ &\le 5 \cdot U(r_\ell) \ .\end{aligned}$$

□

Conversely, we want to show that if one of the conditions in Line 5 of the algorithm is violated, then we can stop iterating. If there exists a distribution $\hat{\boldsymbol{\mu}}_T^{[r_\ell]}$ that is far enough from $\hat{\boldsymbol{\mu}}_T^{[r_j]}$ for a $\ell \in L$ such that $\ell < j$, then a significant distribution drift must have occurred. In particular, we can show a lower bound to $\Delta_{r_j} \ge \xi_{r_\ell}$, and thus $U(r_n) \ge \xi_{r_\ell}$ for any $n \ge j$ due to the drift error. Since $U(r_\ell) = \xi_{r_\ell} + \Delta_{r_\ell}$, and the drift error from using r_ℓ samples is less or equal to the drift error from using r_n samples, we are able to conclude that $U(r_n)$ cannot be significantly smaller than $U(r_\ell)$, and in particular $U(r_\ell) \le 2U(r_n)$. This provides a certificate that we can stop iterating since the window size r_ℓ provides a value of the upper bound $U(\cdot)$ that is up to constant as good as the one obtained with any window size r_n with $n \ge j$. This result is formalized with the following proposition.

Proposition 3.5. *Assume that the event of Proposition 3.3 holds. Let $j \in L$. If there exists $\ell \in L$ with $\ell < j$ such that*

$$\|\hat{\boldsymbol{\mu}}_T^{[r_\ell]} - \hat{\boldsymbol{\mu}}_T^{[r_j]}\|_{\mathrm{TV}} \ge 3\xi_{r_\ell} + \xi_{r_j} \ ,$$

then $U(r_\ell) \le 2U(r_n)$ for any $n \ge j$.

Proof. By the triangle inequality, we have that

$$\begin{aligned}&\|\hat{\boldsymbol{\mu}}_T^{[r_\ell]} - \hat{\boldsymbol{\mu}}_T^{[r_j]}\|_{\mathrm{TV}} \\ \le\ & \|\hat{\boldsymbol{\mu}}_T^{[r_\ell]} - \boldsymbol{\mu}_T^{[r_\ell]}\|_{\mathrm{TV}} + \|\hat{\boldsymbol{\mu}}_T^{[r_j]} - \boldsymbol{\mu}_T^{[r_j]}\|_{\mathrm{TV}} \\ & + \|\boldsymbol{\mu}_T^{[r_\ell]} - \boldsymbol{\mu}_T^{[r_j]}\|_{\mathrm{TV}} \\ \le\ & \xi_{r_j} + \xi_{r_\ell} + \|\boldsymbol{\mu}_T^{[r_\ell]} - \boldsymbol{\mu}_T^{[r_j]}\|_{\mathrm{TV}} \ ,\end{aligned} \tag{11}$$

where the last inequality follows from the assumption that the event of Proposition 3.3 holds. Using the triangle inequality again, we obtain

$$\begin{aligned}\|\boldsymbol{\mu}_T^{[r_\ell]} - \boldsymbol{\mu}_T^{[r_j]}\|_{\mathrm{TV}} &\le \|\boldsymbol{\mu}_T^{[r_\ell]} - \boldsymbol{\mu}_T\|_{\mathrm{TV}} + \|\boldsymbol{\mu}_T^{[r_j]} - \boldsymbol{\mu}_T\|_{\mathrm{TV}} \\ &\le \Delta_{r_\ell} + \Delta_{r_j} \le 2\Delta_{r_j} \ ,\end{aligned}$$

where in the last two inequalities we used relation (6) and the fact that the drift error is non-decreasing with the number of past samples. By combining the above upper bound with (11), we have

$$\|\hat{\boldsymbol{\mu}}_T^{[r_\ell]} - \hat{\boldsymbol{\mu}}_T^{[r_j]}\|_{\mathrm{TV}} \le \xi_{r_j} + \xi_{r_\ell} + 2\Delta_{r_j} \ .$$

We use the assumption of the proposition and obtain the following lower bound to the drift error

$$\Delta_{r_j} \ge \xi_{r_\ell} \ .$$

For any $n \ge j$, we have that

$$2U(r_n) = 2\xi_{r_n} + 2\Delta_{r_n} \ge 2\Delta_{r_j} \ge \xi_{r_\ell} + \Delta_{r_\ell} = U(r_\ell) \ .$$

□

Proposition 3.4 and Proposition 3.5 can be used to prove that our algorithm finds a window size $\hat{r} = 2^j$ for some $j \ge 0$ such that $U(\hat{r}) = \min_i U(r_i)$. We can express this upper bound using the measure of complexity $\Lambda_{\hat{r}}(\cdot)$ thanks to (8). However, this is not sufficient to prove Theorem 2.1, since we are only considering window sizes r that are powers of two, and we want to compare against any possible selection of the window size $1 \le r \le T$. For window sizes that are not power of two, Proposition 3.3 does not provide direct information on the statistical error. To prove the theorem, we need to relate the magnitude of the drift with the change in complexity of the drifting distribution. The following proposition provides this result.

Proposition 3.6. *Let $\boldsymbol{\mu}$ and $\boldsymbol{\eta}$ be two discrete distributions over $\mathbb{N}$. For any integers $1 \le r \le s$, the following two inequalities hold:*

$$\begin{aligned}|\Lambda_r(\boldsymbol{\mu}) - \Lambda_r(\boldsymbol{\eta})| &\le 2\|\boldsymbol{\mu} - \boldsymbol{\eta}\|_{\mathrm{TV}} \ , \\ \frac{\Lambda_r(\boldsymbol{\mu})}{\Lambda_s(\boldsymbol{\mu})} &\le \sqrt{s/r} \ .\end{aligned}$$

Proof. We start by proving the first inequality. We partition $\mathbb{N}$ into four sets: $S_{00} = \{i : \boldsymbol{\mu}(i) < 1/r \wedge \boldsymbol{\eta}(i) < 1/r\}$, $S_{01} = \{i : \boldsymbol{\mu}(i) < 1/r \wedge \boldsymbol{\eta}(i) \ge 1/r\}$, $S_{10} = \{i : \boldsymbol{\mu}(i) \ge 1/r \wedge \boldsymbol{\eta}(i) < 1/r\}$, and $S_{11} = \{i : \boldsymbol{\mu}(i) \ge 1/r \wedge \boldsymbol{\eta}(i) \ge 1/r\}$. We have the following decomposition

$$\begin{aligned}\Lambda_r(\boldsymbol{\mu}) - \Lambda_r(\boldsymbol{\eta}) &= \sum_{i\in S_{00}} (\boldsymbol{\mu}(i)-\boldsymbol{\eta}(i)) \\ &+ \sum_{i\in S_{01}}\left(\boldsymbol{\mu}(i) - \sqrt{\frac{\boldsymbol{\eta}(i)}{r}}\right) + \sum_{i\in S_{10}}\left(\sqrt{\frac{\boldsymbol{\mu}(i)}{r}} - \boldsymbol{\eta}(i)\right) \\ &+ \frac{1}{\sqrt{r}}\sum_{i\in S_{11}}\left(\sqrt{\boldsymbol{\mu}(i)}-\sqrt{\boldsymbol{\eta}(i)}\right) \ .\end{aligned}$$

The sum over S_{00} is upper bounded simply as

$$\sum_{i\in S_{00}} (\boldsymbol{\mu}(i)-\boldsymbol{\eta}(i)) \leq \sum_{i\in S_{00}} |\boldsymbol{\mu}(i)-\boldsymbol{\eta}(i)| \ .$$

For the sum over S_{11}, we have that

$$\begin{aligned}\sum_{i\in S_{11}} \frac{\left[\sqrt{\boldsymbol{\mu}(i)}-\sqrt{\boldsymbol{\eta}(i)}\right]}{\sqrt{r}} &= \frac{1}{\sqrt{r}}\sum_{i\in S_{11}}\frac{(\boldsymbol{\mu}(i)-\boldsymbol{\eta}(i))}{\sqrt{\boldsymbol{\mu}(i)}+\sqrt{\boldsymbol{\eta}(i)}} \\ &\leq \frac{\sqrt{r}}{2\sqrt{r}}\sum_{i\in S_{11}}|\boldsymbol{\mu}(i)-\boldsymbol{\eta}(i)| \\ &= \sum_{i\in S_{11}}\frac{|\boldsymbol{\mu}(i)-\boldsymbol{\eta}(i)|}{2} \ ,\end{aligned}$$

where in the inequality we used the fact that $\sqrt{\boldsymbol{\mu}(i)}$ and $\sqrt{\boldsymbol{\eta}(i)}$ are both at least $\sqrt{1/r}$ for any $i \in S_{11}$. Now, observe that for $i \in S_{01}$, we have that $\boldsymbol{\mu}(i) = \sqrt{\boldsymbol{\mu}(i)}\cdot\sqrt{\boldsymbol{\mu}(i)} \leq \sqrt{\boldsymbol{\mu}(i)/r}$. Using this inequality and proceeding similarly to the previous case, we have that

$$\begin{aligned}\sum_{i\in S_{01}}\left(\boldsymbol{\mu}(i) - \sqrt{\frac{\boldsymbol{\eta}(i)}{r}}\right) &\leq \sum_{i\in S_{01}}\left(\sqrt{\frac{\boldsymbol{\mu}(i)}{r}} - \sqrt{\frac{\boldsymbol{\eta}(i)}{r}}\right) \\ &\leq \sum_{i\in S_{01}}|\boldsymbol{\mu}(i)-\boldsymbol{\eta}(i)| \ .\end{aligned}$$

For $i \in S_{10}$, we observe that $\boldsymbol{\mu}(i) > 1/r$, hence $1/\sqrt{r} < \sqrt{\boldsymbol{\mu}(i)}$. Therefore, it holds that $\sqrt{\boldsymbol{\mu}(i)/r} < \sqrt{\boldsymbol{\mu}(i)}\cdot\sqrt{\boldsymbol{\mu}(i)} < \boldsymbol{\mu}(i)$, and

$$\sum_{i\in S_{10}}\left(\sqrt{\frac{\boldsymbol{\mu}(i)}{r}} - \boldsymbol{\eta}(i)\right) \leq \sum_{i\in S_{10}}(\boldsymbol{\mu}(i)-\boldsymbol{\eta}(i)) \ .$$

We can conclude:

$$\Lambda_r(\boldsymbol{\mu}) - \Lambda_r(\boldsymbol{\eta}) \leq \sum_{i\in\mathbb{N}}|\boldsymbol{\mu}(i)-\boldsymbol{\eta}(i)| \leq 2\|\boldsymbol{\mu}-\boldsymbol{\eta}\|_{\mathrm{TV}} \ .$$

To prove the second inequality, we use instead the following partition of $\mathbb{N}$ into $S_0 = \{i : \boldsymbol{\mu}(i) < 1/s\}$, $S_1 = \{i : 1/s \leq \boldsymbol{\mu}(i) < 1/r\}$, and $S_2 = \{i : 1/r \leq \boldsymbol{\mu}(i)\}$. For any $i \in S_1$, we have that $\boldsymbol{\mu}(i) \leq \sqrt{\boldsymbol{\mu}(i)/r} = \sqrt{s/r}\sqrt{\boldsymbol{\mu}(i)/s}$. We obtain the following result:

$$\begin{aligned}\Lambda_r(\boldsymbol{\mu}) &= \sum_{i\in S_0}\boldsymbol{\mu}(i) + \sum_{i\in S_1}\boldsymbol{\mu}(i) + \sum_{i\in S_2}\sqrt{\frac{\boldsymbol{\mu}(i)}{r}} \\ &\leq \sum_{i\in S_0}\boldsymbol{\mu}(i) + \sqrt{\frac{s}{r}}\sum_{i\in S_1}\sqrt{\frac{\boldsymbol{\mu}(i)}{s}} + \sqrt{\frac{s}{r}}\sum_{i\in S_2}\sqrt{\frac{\boldsymbol{\mu}(i)}{s}} \\ &\leq \sqrt{\frac{s}{r}}\Lambda_s(\boldsymbol{\mu}) \ .\end{aligned}$$

□

We can finally prove Theorem 2.1.

Proof of Theorem 2.1. We assume that the event of Proposition 3.3 holds, otherwise we say that our algorithm fails (with probability $\leq \delta$). The algorithm returns an empirical distribution $\hat{\boldsymbol{\mu}}_T^{[r_j]}$ for some $j \geq 0$, and it guarantees that

$$\|\boldsymbol{\mu}_T - \hat{\boldsymbol{\mu}}_T^{[r_j]}\|_{\mathrm{TV}} \leq U(r_j) \ .$$

Consider the function $Q(r) : \{1, \ldots, T\} \mapsto \mathbb{R}$ defined as

$$Q(r) \doteq \Lambda_r(\boldsymbol{\mu}_T) + \sqrt{\frac{\log(c(\log^2 r + 1)/\delta)}{r}} + \Delta_r \ ,$$

where c is the same constant in Proposition 3.3. Let $r^* = \mathrm{argmin}_{1\leq r\leq T} Q(r)$. In order to prove the theorem, it is sufficient to show that $U(r_j) = O(Q(r^*))$. Let γ be defined as in (10), and let $i \geq 0$ be such that $\gamma(r^*) = r_i$. We distinguish two cases: (*a*) $r_i \leq r_j$ and (*b*) $r_i > r_j$. For both cases, we will use the following result:

$$U(r_i) = O(Q(r^*)) \ . \tag{12}$$

Equation (12) is proven as follows. Let n be the largest integer such that $2^n \leq r^*$. By construction, we have $\gamma(r_n) = r_i$, thus inequality (10) shows us that $U(r_i) \leq U(r_n)$. By definition, $U(r_n) = \xi_{r_n} + \Delta_{r_n}$, and

$$\begin{aligned}\xi_{r_n} &= \Phi(\hat{\boldsymbol{\mu}}_T^{[r_n]}) + 3\sqrt{\frac{\log(c(\log^2 r_n + 1)/\delta)}{r_n}} \\ &\leq 4\Lambda_{r_n}(\boldsymbol{\mu}_T^{[r_n]}) + 12\sqrt{\frac{\log(c(\log^2 r^* + 1)/\delta)}{r^*}} \ ,\end{aligned} \tag{13}$$

where we used (8) and the fact that $r^* \leq 2r_n$. Through, Proposition 3.6 we obtain the following upper bound

$$\begin{aligned}&\Lambda_{r_n}(\boldsymbol{\mu}_T^{[r_n]}) \\ \leq\ &\Lambda_{r_n}(\boldsymbol{\mu}_T) + 2\Delta_{r_n} \\ \leq\ &\sqrt{r^*/r_n}\cdot\Lambda_{r^*}(\boldsymbol{\mu}_T) + 2\Delta_{r_j} \\ \leq\ &4\Lambda_{r^*}(\boldsymbol{\mu}_T) + 2\Delta_{r^*},\end{aligned}$$

where in the second inequality we also used (6), and in the last inequality we used the fact that the drift error is non-decreasing and that $r^* \leq 2r_n$. If we combine the above inequality with (13), we have that

$$U(r_i) \leq U(r_n) \leq 16\Lambda_{r^*}(\boldsymbol{\mu}_T) + 12\sqrt{\frac{\log\left(\frac{c(\log^2 r^* + 1)}{\delta}\right)}{r^*}} + 9\Delta_{r^*} = O(Q(r^*)) \ .$$

Equipped with (12), We will now show that $U(r_j) = O(Q(r^*))$ for both cases (a) and (b).

Case (a). Since $i \in L$, we have $U(r_j) \leq 5U(r_i)$ due to Proposition 3.5. We conclude by using equation (12).

Case (b). The algorithm chooses the window size $r_j < r_i$. This means that when the algorithm considers the next element $\ell \in L$ after j, where $\ell \leq i$, there exists an index $\ell' \in L$ with $\ell' \leq j$ such that the condition of Line 5 of the algorithm is satisfied. Proposition 3.5 applies, and we have that $U(r_{\ell'}) \leq 2U(r_n)$ for any $n \geq \ell > j$. By construction, this is also true for n equal to i, and $U(r_{\ell'}) \leq 2U(r_i)$. On the other hand, since the algorithm returned r_j, it means that all the If conditions on Line 5 must have been satisfied when considering $j \in L$, and Proposition 3.4 gives us $U(r_j) \leq 5 \min_{z \leq j : z \in L} U(r_z)$, and in particular $U(r_j) \leq U(r_{\ell'})$. Thus, $U(r_j) \leq 10U(r_i)$. We obtain the statement by using equation (12). □

4 RELATED WORK

The problem of learning with distribution drift was introduced in the context of binary classification (Helmbold and Long, 1991; Bartlett, 1992; Helmbold and Long, 1994). This line of research led to the result that if any two consecutive distributions have a bounded L_1 distance Δ, the expected error for learning a family of binary functions with VC dimension ν is $O((\nu\Delta)^{1/3})$ (Long, 1998), and the upper bound is tight (Barve and Long, 1996). Under mild assumptions, Mohri and Muñoz Medina (2012) extend the analysis of agnostic learning with drift to any family of functions. In particular, they provide an upper bound to the learning error for a given window size that uses the Rademacher complexity to quantify the statistical error, and a problem-dependent upper bound to the drift error called discrepancy that is based on previous work in domain adaptation (Mansour et al., 2009; Ben-David et al., 2010). Recent work relaxes the independence assumption and provides learning bounds under mixing condition (Hanneke and Yang, 2019). This previous work either requires a priori knowledge of the drift error to solve the trade-off between statistical error and drift error, or assumes that multiple samples can be obtained from each distribution to estimate the drift error (Mohri and Muñoz Medina, 2012; Awasthi et al., 2023). There are two work that provide an adaptive algorithm for learning a family of functions with drift: Hanneke et al. (2015) address the realizable case, and Mazzetto and Upfal (2023a) address the agnostic case.

Recent work characterizes the minimax error for the problem of learning discrete and continuous smooth distributions with distribution drift, but it assumes a prior knowledge of the drift to attain this error (Mazzetto and Upfal, 2023b). In a more specific setting, Gokcesu and Kozat (2017) provide an adaptive algorithm for learning a parametric family of exponential densities, where the parameters can slowly drift over time. We finally point out that several algorithms have been proposed for estimating a density in the online setting (Kristan et al., 2011; García-Treviño and Barria, 2012), but they do not provide an analysis of the estimation error.

5 CONCLUSION

We provide an adaptive algorithm for the problem of learning a drifting discrete distribution. Unlike previous work, our method solves this problem for *any* drifting discrete distribution, and it does not require any prior assumption on the support of the distribution. Additionally, our algorithm utilizes the input data to estimate the statistical error, and it can provide a tighter bound than existing methods depending on the complexity of the drifting distribution. To the best of our knowledge, this is the first adaptive method for learning a discrete distribution to use data-dependent bounds in a drift setting.

Acknowledgements

The authors would like to thank Eli Upfal for the helpful discussions. This material is based on research sponsored by the National Science Foundation (NSF) under award IIS-1813444, and by a Kanellakis Fellowship.

References

Anthony, M., Bartlett, P. L., Bartlett, P. L., et al. (1999). *Neural network learning: Theoretical foundations*, volume 9. Cambridge University Press.

Awasthi, P., Cortes, C., and Mohri, C. (2023). Theory and algorithm for batch distribution drift problems. In *Proc. AISTATS*, pages 9826–9851.

Bartlett, P. L. (1992). Learning with a slowly changing distribution. In *Proc. COLT*, pages 243–252.

Barve, R. D. and Long, P. M. (1996). On the complexity of learning from drifting distributions. In *Proc. COLT*, pages 122–130.

Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, 79:151–175.

Berend, D. and Kontorovich, A. (2013). A sharp estimate of the binomial mean absolute deviation with applications. *Statistics & Probability Letters*, 83(4):1254–1259.

Cohen, D., Kontorovich, A., and Wolfer, G. (2020). Learning discrete distributions with infinite support. In *Proc. NeurIPS*, pages 3942–3951.

Devroye, L. and Györfi, L. (1987). Nonparametric density estimation : the l[1] view. *Journal of the American Statistical Association*, 82:344.

Devroye, L. and Lugosi, G. (2001). *Combinatorial methods in density estimation.* Springer Science & Business Media.

Fu, D., Chen, M., Sala, F., Hooper, S., Fatahalian, K., and Ré, C. (2020). Fast and three-rious: Speeding up weak supervision with triplet methods. In *Proc. ICML*, pages 3280–3291.

García-Treviño, E. S. and Barria, J. A. (2012). Online wavelet-based density estimation for non-stationary streaming data. *Computational statistics & data analysis*, 56(2):327–344.

Gokcesu, K. and Kozat, S. S. (2017). Online density estimation of nonstationary sources using exponential family of distributions. *IEEE Transactions on Neural Networks and Learning Systems*, 29(9):4473–4478.

Han, Y., Jiao, J., and Weissman, T. (2015). Minimax estimation of discrete distributions. In *2015 IEEE International Symposium on Information Theory (ISIT)*, pages 2291–2295. IEEE.

Hanneke, S., Kanade, V., and Yang, L. (2015). Learning with a drifting target concept. In *Proc. ALT*, pages 149–164.

Hanneke, S. and Yang, L. (2019). Statistical learning under nonstationary mixing processes. In *Proc. AISTATS*, pages 1678–1686.

Helmbold, D. P. and Long, P. M. (1991). Tracking drifting concepts using random examples. In *Proc. COLT*, pages 13–23.

Helmbold, D. P. and Long, P. M. (1994). Tracking drifting concepts by minimizing disagreements. *Machine Learning*, 14:27–45.

Jiao, J., Venkat, K., Han, Y., and Weissman, T. (2015). Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885.

Kamath, S., Orlitsky, A., Pichapati, D., and Suresh, A. T. (2015). On learning distributions from their samples. In *Proc. COLT*, pages 1066–1100.

Kristan, M., Leonardis, A., and Skočaj, D. (2011). Multivariate online kernel density estimation with gaussian kernels. *Pattern recognition*, 44(10-11):2630–2642.

Long, P. M. (1998). The complexity of learning according to two models of a drifting environment. In *Proc. COLT*, pages 116–125.

Mansour, Y., Mohri, M., and Rostamizadeh, A. (2009). Domain adaptation: Learning bounds and algorithms. In *Proc. COLT*.

Mazzetto, A. and Upfal, E. (2023a). An adaptive algorithm for learning with unknown distribution drift. In *Proc. NeurIPS*.

Mazzetto, A. and Upfal, E. (2023b). Nonparametric density estimation under distribution drift. In *Proc. ICML*, pages 24251–24270.

Mohri, M. and Muñoz Medina, A. (2012). New analysis and algorithm for learning with drifting distributions. In *Proc. ALT*, pages 124–138.

Orlitsky, A. and Suresh, A. T. (2015). Competitive distribution estimation: Why is good-turing good. In *Proc. NeurIPS*.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*, volume 26. CRC press.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A DEFERRED PROOFS

Proof of Proposition 3.2. The proof of this proposition is based on previous work in the literature. The first inequality of the proposition immediately follows from the proof of Cohen et al. (2020, Theorem 2.1). Specifically, with probability at least $1-\delta/2$, it holds:

$$\|\hat{\boldsymbol{\mu}}_T^{[r]} - \boldsymbol{\mu}_T^{[r]}\|_{\mathrm{TV}} \le \Phi_r\left(\hat{\boldsymbol{\mu}}_T^{[r]}\right) + 3\sqrt{\frac{\log(4/\delta)}{2r}} \ . \tag{14}$$

The proof of the second inequality of this proposition proceeds as follows. By invoking Fubini's theorem, we have that

$$\mathbb{E}\,\Phi(\hat{\boldsymbol{\mu}}_T^{[r]}) = \frac{1}{\sqrt{r}}\,\mathbb{E}\left(\sum_{i=1}^{\infty}\sqrt{\hat{\boldsymbol{\mu}}_T^{[r]}(i)}\right) = \frac{1}{\sqrt{r}}\sum_{i=1}^{\infty}\mathbb{E}\left(\sqrt{\hat{\boldsymbol{\mu}}_T^{[r]}(i)}\right) = \frac{1}{r}\sum_{i=1}^{\infty}\mathbb{E}\left(\sqrt{r\cdot\hat{\boldsymbol{\mu}}_T^{[r]}(i)}\right) \ .$$

Consider a value $i \in \mathbb{N}$, and let $C_i = r\hat{\boldsymbol{\mu}}_T^{[r]}(i)$. The crucial observation is that

$$\mathbb{E}\,C_i = r\,\mathbb{E}\left(\frac{1}{r}\sum_{t=T-r+1}^{T}\mathbf{1}_{\{X_t=i\}}\right) = \sum_{t=T-r+1}^{T}\mathbb{E}\,\mathbf{1}_{\{X_t=i\}} = \sum_{t=T-r+1}^{T}\boldsymbol{\mu}_t(i) = r\boldsymbol{\mu}_T^{[r]}(i) \ .$$

The remaining of the proof follows the same argument in previous work (Cohen et al., 2020). Since the square root is a concave function, we have that $\mathbb{E}\sqrt{C_i} \le \sqrt{\mathbb{E}\,C_i}$ by using Jensen's inequality. Furthermore, we can exploit that $\sqrt{C_i} \le C_i$ as $C_i \in \{0,1,\dots,r\}$, to show $\mathbb{E}\sqrt{C_i} \le \mathbb{E}\,C_i$. We obtain:

$$\begin{aligned}
\mathbb{E}\,\Phi(\hat{\boldsymbol{\mu}}_T^{[r]}) &= \frac{1}{r}\sum_{i=1}^{\infty}\mathbb{E}\sqrt{C_i} \\
&\le \frac{1}{r}\sum_{i=1}^{\infty}\min\left\{r\boldsymbol{\mu}_T^{[r]}(i), \sqrt{r\boldsymbol{\mu}_T^{[r]}(i)}\right\} \\
&= \sum_{i:\boldsymbol{\mu}_T^{[r]}(i)\le 1/r}\boldsymbol{\mu}_T^{[r]}(i) + \frac{1}{\sqrt{r}}\sum_{i:\boldsymbol{\mu}_T^{[r]}(i)>1/r}\boldsymbol{\mu}_T^{[r]}(i) \\
&= \Lambda_r(\boldsymbol{\mu}_T^{[r]}) \ .
\end{aligned}$$

Since changing a single sample among $\{X_{T-r+1},\dots,X_T\}$ can change the value of $\Phi(\hat{\boldsymbol{\mu}}_T^{[r]})$ by at most $2/r$, we can use McDiarmid's inequality to show that with probability at least $1-\delta/2$, it holds that:

$$\Phi(\hat{\boldsymbol{\mu}}_T^{[r]}) \le \mathbb{E}\,\Phi(\hat{\boldsymbol{\mu}}_T^{[r]}) + \sqrt{\frac{\log(2/\delta)}{r}} \ . \tag{15}$$

We conclude by taking a union bound so that both events (14) and (15) hold with probability at least $1-\delta$. □