TURF: A Two-factor, Universal, Robust, Fast Distribution Learning Algorithm

Yi Hao [1] **Ayush Jain** [1] **Alon Orlitsky** [1] **Vaishakh Ravindrakumar** [1]

Abstract

Approximating distributions from their samples is a canonical statistical-learning problem. One of its most powerful and successful modalities approximates every distribution to an ℓ_1 distance essentially at most a constant times larger than its closest t-piece degree-d polynomial, where $t \geq 1$ and $d \geq 0$. Letting $c_{t,d}$ denote the smallest such factor, clearly $c_{1,0} = 1$, and it can be shown that $c_{t,d} \geq 2$ for all other t and d. Yet current computationally efficient algorithms show only $c_{t,1} \leq 2.25$ and the bound rises quickly to $c_{t,d} \leq 3$ for $d \geq 9$. We derive a near-linear-time and essentially sample-optimal estimator that establishes $c_{t,d} = 2$ for all $(t, d) \neq (1, 0)$. Additionally, for many practical distributions, the lowest approximation distance is achieved by polynomials with vastly varying number of pieces. We provide a method that estimates this number near-optimally, hence helps approach the best possible approximation. Experiments combining the two techniques confirm improved performance over existing methodologies.

1. Introduction

Learning distributions from samples is one of the oldest (Pearson, 1895), most natural (Silverman, 1986), and important statistical-learning paradigms (Givens & Hoeting, 2012). Its numerous applications include epidemiology (Bithell, 1990), economics (Zambom & Ronaldo, 2013), anomaly detection (Pimentel et al., 2014), language based prediction (Gerber, 2014), GANs (Goodfellow et al., 2014), and many more, as outlined in several books and surveys e.g., (Tukey, 1977; Scott, 2012; Diakonikolas, 2016).

Consider estimating an unknown, real, discrete, continuous, or mixed distribution f from n independent samples $X^n := X_1, ..., X_n$ it generates. A *distribution estimator* maps X^n to an approximating distribution f^{est} meant to approximate f. We evaluate its performance via the expected ℓ_1 *distance* $\mathbb{E}\|f^{est} - f\|_1$.

The ℓ_1 distance between two functions f_1 and f_2, $\|f_1 - f_2\|_1 := \int_{\mathbb{R}} |f_1 - f_2|$, is one of density estimation's most common distance measures (Devroye & Lugosi, 2012). Among its several desirable properties, its value remains unchanged under linear transformation of the underlying domain, and the absolute difference between the expected values of any bounded function of the observations under f_1 and f_2 is at most a constant factor larger than $\|f_1 - f_2\|_1$, as for any bounded $g : \mathbb{R} \to \mathbb{R}$, $\left|\mathbb{E}_{f_1}[g(X)] - \mathbb{E}_{f_2}[g(X)]\right| \leq \max_{x\in\mathbb{R}} g(x) \cdot \|f_1 - f_2\|_1$. Further, a small ℓ_1 distance between two distributions implies a small difference between any given Lipschitz functions of the two distributions. Therefore, learning in ℓ_1 distance implies a bound on the error of the plug-in estimator for Lipschitz functions of the underlying distribution (Hao & Orlitsky, 2019).

Ideally, we would like to learn any distribution to a small ℓ_1 distance. However, arbitrary distributions cannot be learned in ℓ_1 distance with any number of samples (Devroye & Gyorfi, 1990), as the following example shows.

Example 1. *Let u be the continuous uniform distribution over $[0, 1]$. For any number n of samples, construct a discrete distribution p by assigning probability $1/n^3$ to each of n^3 random points in $[0, 1]$. By the birthday paradox, n samples from p will be all distinct with high probability and follow the same uniform distribution as n samples from u, and hence u and p will be indistinguishable. As $\|u - p\|_1 = 2$, the triangle inequality implies that for any estimator f^{est}, $\max_{f\in\{u,p\}} \mathbb{E}\|f^{est} - f\|_1 \gtrsim 1$.*

A common remedy to this shortcoming assumes that the distribution f belongs to a structured approximation class $\mathcal{C}$, for example unimodal (Birgé, 1987), log-concave (Devroye & Lugosi, 2012) and Gaussian (Acharya et al., 2014; Ashtiani et al., 2018) distributions.

The *min-max learning rate* of $\mathcal{C}$ is the lowest worst-case expected distance achieved by any estimator,

$$\mathcal{R}_n(\mathcal{C}) \stackrel{\text{def}}{=} \min_{f^{est}} \max_{f\in\mathcal{C}} \mathop{\mathbb{E}}_{X^n\sim f} \|f^{est}_{X^n} - f\|_1.$$

[1]Electrical and Computer Engineering, University of California, San Diego. Correspondence to: Vaishakh Ravindrakumar <varavind@ucsd.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

The study of $\mathcal{R}_n(\mathcal{C})$ for various classes such as Gaussians, exponentials, and discrete distributions has been the focus of many works e.g., (Vapnik, 1999; Kamath et al., 2015; Han et al., 2015; Cohen et al., 2020).

Considering all pairs of distributions in $\mathcal{C}$, (Yatracos, 1985) defined a collection of subsets with VC dimension (Vapnik, 1999) $\mathrm{VC}(\mathcal{C})$, and applying the minimum distance estimation method (Wolfowitz, 1957), showed that

$$\mathcal{R}_n(\mathcal{C}) = \mathcal{O}(\sqrt{\mathrm{VC}(\mathcal{C})/n}).$$

However, real underlying distributions f are unlikely to fall *exactly* in any predetermined class. Hence (Yatracos, 1985) also considered approximating f nearly as well as its best approximation in $\mathcal{C}$. Letting

$$\|f - \mathcal{C}\|_1 \stackrel{\text{def}}{=} \inf_{g\in\mathcal{C}} \|f - g\|_1$$

be lowest ℓ_1 distance between f and any distribution in $\mathcal{C}$, he designed an estimator f^{Yat}, possibly outside $\mathcal{C}$, whose ℓ_1 distance from f is close to $\|f-\mathcal{C}\|_1$. For all distributions f,

$$\mathbb{E}\,\|f^{Yat} - f\|_1 \le 3\cdot\|f-\mathcal{C}\|_1 + \mathcal{O}(\sqrt{\mathrm{VC}(\mathcal{C})/n}).$$

For many natural classes, $\mathcal{R}_n(\mathcal{C}) = \Theta(\sqrt{\mathrm{VC}(\mathcal{C})/n})$. Hence, an estimator f^{est} is called a *c-factor approximation* for $\mathcal{C}$ if for any distribution f,

$$\mathbb{E}\,\|f^{est} - f\|_1 \le c\cdot\|f-\mathcal{C}\|_1 + \mathcal{O}(\mathcal{R}_n(\mathcal{C})).$$

$\|f-\mathcal{C}\|_1$ and $\mathcal{O}(\mathcal{R}_n(\mathcal{C}))$ may be thought of as the error's *bias* and *variance* components.

A small c is desirable as it upper bounds the asymptotic error when $n \nearrow \infty$ for $f \notin \mathcal{C}$, hence providing robustness guarantees when the underlying distribution does not quite follow the assumed model. It ensures robust estimation also under the Huber contamination model (Huber, 1992) where with probability $0 \le \mu \le 1$, f is perturbed by an arbitrary noise, and the error incurred by a c-factor approximation is upper bounded as $c\cdot\mu$.

One of the more important distribution classes is the collection $\mathcal{P}_{t,d}$ of t-piecewise degree-d polynomials. For simplicity, we assume that all polynomials in $\mathcal{P}_{t,d}$ are defined over a known interval $I \subseteq \mathbb{R}$, hence any $p \in \mathcal{P}_{t,d}$ consists of degree-d polynomials $p_1,\ldots,p_t$, each defined over one part in a partition $I_1,\ldots,I_t$ of I.

The significance of $\mathcal{P}_{t,d}$ stems partly from the fact that it approximates numerous important distributions even with small t and d. For example, for every distribution f in the class $\mathcal{L}$ of log-concave distributions, $\|f-\mathcal{P}_{t,1}\|_1 = \mathcal{O}(t^{-2})$ (Chan et al., 2014). Also, $\mathrm{VC}(\mathcal{P}_{t,d}) = t(d+1)$, e.g., (Acharya et al., 2017).

It follows that if f^{est} is a c-factor estimator for $\mathcal{P}_{t,1}$, then for all $f\in\mathcal{L}$, $\mathbb{E}\,\|f^{est}-f\|_1 \le c\cdot\|f-\mathcal{P}_{t,1}\|_1 + \mathcal{O}(\sqrt{t/n}) = \mathcal{O}(t^{-2}) + \mathcal{O}(\sqrt{t/n})$. Choosing $t = n^{1/5}$ to equate the bias and variance terms, f^{est} achieves an expected ℓ_1 error $\mathcal{O}(n^{-2/5})$, which is the optimal min-max learning rate of $\mathcal{L}$ (Chan et al., 2014).

Lemma 17 in Appendix A.2 shows a stronger result. If f^{est} is a c-factor approximation for $\mathcal{P}_{t,d}$ for some t and d and achieves the min-max rate of a distribution class $\mathcal{C}$, then f^{est} is also a c-factor approximation for $\mathcal{C}$. In addition to the log-concave class, this result also holds for Gaussian, and unimodal distributions, and for their mixtures.

2. Contributions

Lower Bounds: As noted above, it is beneficial to find the smallest approximation factor for $\mathcal{P}_{t,d}$. The following simple example shows that if we allow sub-distributions, even simple collections may have an approximation factor of at least 2.

Example 2. *Let class $\mathcal{C}$ consist of the uniform distribution $u(x) = 1$ and the subdistribution $z(x) = 0$, over $[0,1]$. Consider any estimator f^{est}. Let $f_u = f^{est}$ when $X^n \sim u$ as $n \nearrow \infty$. Since $\|u-\mathcal{C}\|_1 = 0$, for f^{est} to achieve finite approximation factor, we must have $\|f_u - u\|_1 = 0$. Now consider the discrete distribution p in Example 1. Since its samples are indistinguishable from those of u, $f^{est}_{X^n} = f_u$ also for $X^n \sim p$. But then $\|f_u - p\|_1 \ge \|u-p\|_1 - \|f_u - u\|_1 = 2 = 2\cdot\|p-\mathcal{C}\|_1$, so f^{est} has approximation factor ≥ 2.*

Our definition however considers only strict distributions, complicating lower bound proofs. Let $c_{t,d}$ be the lowest approximation factor for $\mathcal{P}_{t,d}$. $\mathcal{P}_{1,0}$ consists of a single distribution over a known interval, hence $c_{1,0} = 1$. (Chan et al., 2014) showed that for all $t \ge 2$ and $d \ge 0$, $c_{t,d} \ge 2$. The following lemma, proved in Appendix A.1, shows that $c_{1,d} \ge 2$ for all $d \ge 1$, and as we shall see later, establishes a precise lower bound for all t and d.

Lemma 3. *For all (t,d) except $(1,0)$, $c_{t,d} \ge 2$.*

Upper Bounds: As discussed earlier, f^{Yat} is a 3-factor approximation for $\mathcal{P}_{t,d}$. However its runtime is $n^{\mathcal{O}(t(d+1))}$. For many applications, t or d may be large, and even increase with n, for example in learning unimodal distributions, we select $t = \mathcal{O}(n^{1/3})$ (Birgé, 1987), resulting in exponential time complexity. (Chan et al., 2014) improved the runtime to polynomial in n independent of t, d, and (Acharya et al., 2017) further reduced it to near-linear $\mathcal{O}(n\log n)$. (Hao et al., 2020b) derived a $\mathcal{O}(n\log n)$ time algorithm, SURF, achieving $c_{t,1} = 2.25$, and $c_{t,d} < 3$ for $d \le 8$. They also showed that this estimator can be parallelized to run in time $\mathcal{O}(n\log n/t)$. (Bousquet et al., 2019; 2021)'s estimator for the improper learning setting (wherein f^{est} can be any distribution as we consider in this paper) achieves a bias nearly within a factor of 2, but the variance term exceeds $\mathcal{O}(\mathcal{R}_n(\mathcal{P}_{t,d}))$, hence does not satisfy the constant factor approximation definition. Moreover, like

Yatracos, they suffer a prohibitive $n^{\mathcal{O}(t(d+1))}$ runtime, that could be exponential for some applications.

Our main contribution is an estimator, TURF, *a two factor, universal, robust and fast estimator* that achieves an approximation factor that is arbitrarily close to the optimal $c_{t,d} = 2$ in near-linear $\mathcal{O}(n \log n)$ time. TURF is also simple to implement as a step on top of the existing merge routine in (Acharya et al., 2017). The construction of our estimate relies on upper bounding the maximum absolute value of polynomials (see Lemma 7) based on their ℓ_1 norm, similar to the Bernstein (Rahman et al., 2002) and Markov Brothers' (Achieser, 1992) inequalities. We show for any $p \in \mathcal{P}_{1,d}$ and $a \in [0, 1)$,

$$\|p\|_{\infty,[-a,a]} \le \frac{28(d+1)\|p\|_{1,[-1,1]}}{\sqrt{1-a^2}},$$

where $\|\cdot\|_{\cdot,I}$ indicates the respective norms over any interval $I \subseteq \mathbb{R}$. This point-wise inequality reveals a novel connection between the ℓ_∞ and ℓ_1 norms of a polynomial, which may be interesting in its own right.

Practical Estimation: For many practical distributions, the optimal parameters values of t, d in approximating with $\mathcal{P}_{t,d}$ many be unknown. While for common structured classes such as Gaussian, log-concave and unimodal, and their mixtures, it suffices to choose d to be any small value, but the optimal choice of t can vary significantly. For example, for any constant d, for a unimodal f, the optimal $t = \mathcal{O}(n^{1/3})$ pieces whereas for a smoother log-concave f, significantly lower errors are obtained with a much smaller $t = \mathcal{O}(n^{1/5})$. Given a family of $c_{t,d}$-factor approximate estimators for $\mathcal{P}_{t,d}$, $f^{est}_{t,d}$, a suitable objective is to select the number of pieces, t^{est}_d to achieve for any given degree-d,

$$\mathbb{E}\|f^{est}_{t^{est}_d} - f\|_1 \le \min_{t\ge 1}\Big(c_{t,d}\|f - \mathcal{P}_{t,d}\|_1 + \mathcal{O}\Big(\sqrt{t(d+1)/n}\Big)\Big). \tag{1}$$

Simple modifications to existing cross-validation approaches (Yatracos, 1985) partly achieve Equation (1) with the larger $c = 3c_{t,d}$ along with an additive $\mathcal{O}(\log n/\sqrt{n})$. Via a novel cross-validation technique, we obtain a t^{est}_d that satisfies Equation (1) with the factor c arbitrarily close to the optimal $c_{t,d}$ with an additive $\mathcal{O}(\sqrt{\log n/n})$. In fact, this technique removes the need to know parameters beforehand in other related settings as well, such as the *corruption level* in robust estimation that all existing works assume is known. We elaborate this in (Jain et al., 2022).

Our experiments reflect the improved errors of TURF over existing algorithms in regimes where the bias dominates.

3. Setup

3.1. Notation and Definitions

Henceforth, for brevity, we skip the X^n subscript when referring to estimators. Given samples $X^n \sim f$, the *empirical distribution* is defined via the dirac delta function $\delta(x)$ as

$$f^{emp}(x) \stackrel{\text{def}}{=} \sum_{i=1}^{n} \frac{\delta(x - X_i)}{n},$$

allotting a $1/n$ mass at each sample location.
Note that if an estimator g is partly negative but integrates to 1, then $g' \stackrel{\text{def}}{=} \max\{g, 0\}/\int_{\mathbb{R}} \max\{g, 0\}$, satisfies $d_{\text{TV}}(g', f) \le d_{\text{TV}}(g, f)$ for any distribution f, e.g., Devroye & Lugosi (2012). This allows us to estimate using any real normalized function as our estimator.

For any interval $I \subseteq \mathbb{R}$ and integrable functions $g_1, g_2 : \mathbb{R} \to \mathbb{R}$, let $\|g_1 - g_2\|_{1,I}$ denote the ℓ_1 distance evaluated over I. Similarly, for any class $\mathcal{C}$ of real functions, let $\|g - \mathcal{C}\|_{1,I}$ denote the least ℓ_1 distance between g and members of $\mathcal{C}$ over I.

The ℓ_1 distance between f and f^{est} is closely related to their TV or statistical distance as

$$1/2 \cdot \|f^{est} - f\|_1 = d_{\text{TV}}(f^{est}, f) \stackrel{\text{def}}{=} \sup_{S \in \mathbb{R}} \Big| \int_S f^{est} - f \Big|,$$

the greatest absolute difference in areas of f^{est} and f over all subsets of $\mathbb{R}$. As we argued in the introduction, a direct approach to estimate f in all possible subsets of $\mathbb{R}$ is not feasible with finitely many samples. Instead, for a given $k \ge 1$, the $\mathcal{A}_k$ distance (Devroye & Lugosi, 2012) considers the largest difference between f and f^{est} on real subsets with at most k intervals. As we show in Lemma 4, it is possible to learn any f in $\mathcal{A}_k$ distance simply by using the empirical distribution f^{emp}.

We formally define the $\mathcal{A}_k$ distance as follows. For any given $k \ge 1$ and interval $I \subseteq \mathbb{R}$, let $\mathcal{I}_k(I)$ be the set of all unions of at most k intervals contained in I. Define the $\mathcal{A}_k$ distance between g_1, g_2 as

$$\|g_1 - g_2\|_{\mathcal{A}_k, I} \stackrel{\text{def}}{=} \sup_{S \in \mathcal{I}_k(I)} |g_1(S) - g_2(S)|,$$

where $g(S)$ denotes the area of the function g on the set S. For example, if $I = [0, 1]$ and $g_1(x) = x$, $g_2(x) = 2/3$, the $\mathcal{A}_1$ distance $\|g_1 - g_2\|_{\mathcal{A}_1, I} = \int_0^{2/3} |z - 2/3| dz = 2/9$. Suppose I is the support of f. Use this to define $\|g_1 - g_2\|_{\mathcal{A}_k} \stackrel{\text{def}}{=} \|g_1 - g_2\|_{\mathcal{A}_k, I}$.

For two *distributions* g_1 and g_2, the $\mathcal{A}_k$ distance is at-most half the ℓ_1 distance and with equality achieved as $k \nearrow \infty$ since $\mathcal{I}_k(I)$ approximates all subsets of I for large k,

$$\|g_1 - g_2\|_{\mathcal{A}_k, I} \le 1/2 \cdot \|g_1 - g_2\|_1.$$

The reverse is not true, the $\mathcal{A}_k$ distance between two functions may be made arbitrarily small even for a constant ℓ_1 distance. For example the ℓ_1 distance between any distribution f and its empirical distribution f^{emp} is 2 for any $n \ge 2$. However, the $\mathcal{A}_k$ distance between f and f^{emp} goes to zero. The next lemma, which is a consequence of VC inequality (Devroye & Lugosi, 2012), gives the rate at which $\|f^{emp} - f\|_{\mathcal{A}_k}$ goes to zero.

Lemma 4. *(Devroye & Lugosi, 2012) Given $X^n \sim f$ according to any real distribution f,*

$$\mathbb{E}\,\|f^{emp} - f\|_{\mathcal{A}_k} = \mathcal{O}\Big(\sqrt{k/n}\Big).$$

Note that if f is a discrete distribution with support size k, Lemma 4 implies $\mathbb{E}\,\|f^{emp} - f\|_1 = \mathbb{E}\,\|f^{emp} - f\|_{\mathcal{A}_k} \le \mathcal{O}(\sqrt{k/n})$, matching the rate of learning discrete distributions. Since arbitrary continuous distributions can be thought of as infinite dimensional discrete distributions where $k \to \infty$, the lemma does not bound this error.

3.2. Preliminaries

The following $\mathcal{A}_k$-distance properties are helpful.

Property 5. *Given the partition I_1, I_2 of any interval $I \subseteq \mathbb{R}$ integrable functions $g_1, g_2 \in I$, and integers k_1, k_2,*

$$\|g_1 - g_2\|_{\mathcal{A}_{k_1}, I_1} + \|g_1 - g_2\|_{\mathcal{A}_{k_2}, I_2} \le \|g_1 - g_2\|_{\mathcal{A}_{k_1+k_2}, I}.$$

Property 5 follows since the interval choices with k_1 and k_2 intervals respectively that achieve the suprema of $\|g_1 - g_2\|_{\mathcal{A}_{k_1}, I_1}$ and $\|g_1 - g_2\|_{\mathcal{A}_{k_2}, I_2}$ are included in the $k_1 + k_2$ interval partition considered in the RHS.

Property 6. *Given any interval $I \subseteq \mathbb{R}$, integrable functions $g_1, g_2 \in I$, and integers $k_1 \ge k_2 > 0$,*

$$\|g_1 - g_2\|_{\mathcal{A}_{k_1}, I} \le \frac{k_1}{k_2} \cdot \|g_1 - g_2\|_{\mathcal{A}_{k_2}, I}.$$

Property 6 follows from selecting the k_2 intervals with the largest contribution to $\|g_1 - g_2\|_{\mathcal{A}_{k_1}, I}$ in the RHS expression, among the k_1 interval partition that attains $\|g_1 - g_2\|_{\mathcal{A}_{k_1}, I}$ on the LHS. In Sections 4, 5 that follow, we consider deriving the optimal rates of learning with piecewise polynomials.

4. A 2-Factor Estimator for $\mathcal{P}_{1,d}$

Our objective is to obtain a 2-factor approximation for the piecewise class, $\mathcal{P}_{t,d}$. To achieve this, we first consider the single-piece class $\mathcal{P}_{1,d}$ that for simplicity we denote by $\mathcal{P}_d = \mathcal{P}_{1,d}$, and then use the resulting estimator as a sub-routine for the multi-piece class.

4.1. Intuition and Results

It is easy to show from the triangle inequality that if an estimator is as close to *all* degree-d polynomials as their ℓ_1 distances to f, then the estimator achieves an ℓ_1 distance to f that is nearly twice that of the best degree-d polynomial.

Let $|I|$ denote the length of an interval I. The *histogram* of an integrable function g over I is $\bar{g}_I \overset{\text{def}}{=} |\int_I g|/|I|$, where we assign zero to division by zero.

Let $f^{poly} \in \mathcal{P}_d$ be a polynomial estimator of f over I, and let f^{adj} be the function obtained by adding to f^{poly} a constant to match its mass to f over I. For any $p \in \mathcal{P}_d$,

$$\begin{aligned}\|f^{adj} - p\|_{1,I} &\overset{(a)}{\le} \|f^{adj} - p - \Big(\overline{f^{adj}} - \bar{p}\Big)\|_{1,I} + \|\overline{f^{adj}} - \bar{p}\|_{1,I} \\ &\overset{(b)}{\le} \|f^{adj} - p - \Big(\overline{f^{adj}} - \bar{p}\Big)\|_{1,I} + \|f - p\|_{1,I},\end{aligned}$$

where (a) follows by the triangle inequality, and (b) follows since f^{adj} has the same mass as f by construction, it implies $\|\overline{f^{adj}} - \bar{p}\|_{1,I} = \|\bar{f} - \bar{p}\|_{1,I} \le \|f - p\|_{\mathcal{A}_1, I} \le \|f - p\|_{1,I}$. Since $f^{adj} - p \in \mathcal{P}_d$, if $\|q - \bar{q}\|_{1,I}$ is a small value $\forall q \in \mathcal{P}_d$, f^{adj} approximates f nearly as well as any degree-d polynomial. Let

$$\Delta_I(q) \overset{\text{def}}{=} \max_{x \in I} q(x) - \min_{x \in I} q(x)$$

be the difference between q's largest and smallest values. Note that $\bar{q}_I(x)$ has zero mean over I, hence must be zero on at least one point in I, implying

$$\|q - \bar{q}\|_{1,I} \le \Delta_I(q) \cdot |I|. \tag{2}$$

Thus we would like $\Delta_I(q) \cdot |I|$ to be small $\forall q \in \mathcal{P}_d$, but which may not hold for the given I. By additivity, we may partition I and perform this adjustment over each sub-interval. A *partition* $\bar{I}$ of I is a collection of disjoint intervals whose union is I. Let the *histogram of g over $\bar{I}$* be

$$\bar{g}_{\bar{I}}(x) \overset{\text{def}}{=} \frac{|\int_J g|}{|J|} = \bar{g}_J(x) \qquad x \in J \in \bar{I}, \tag{3}$$

We will construct a partition for which $\sum_{J \in \bar{I}} \Delta_J(q) \cdot |J|$ is small $\forall q \in \mathcal{P}_d$. Further, as we don't know f, we will use the empirical distribution f^{emp} that approximates the mass of f over each interval of $\bar{I}$. By Lemma 4, for any $\bar{I}$ with k intervals, the expected extra error $\mathbb{E}\,\|\bar{f}_{\bar{I}} - \overline{f^{emp}}_{\bar{I}}\|_1 = \mathcal{O}(\sqrt{k/n})$. If we want this error to be within a constant factor from the $\mathcal{O}(\sqrt{(d+1)/n})$ min-max rate of $\mathcal{P}_d$, we need to take $k = \mathcal{O}(d+1)$.

We use the bound in Lemma 7 to construct a partition $\bar{I}$ whose widths decrease towards the extremes, while ensuring $k = \mathcal{O}(d+1)$. In Section 4.2 we show that for this universal partition, $\sum_{I \in \bar{I}} \Delta_I(q) \cdot |I|$ decreases at the rate $\mathcal{O}((d+1)/k)$ for *all* $q \in \mathcal{P}_d$ that we conjecture is optimal in d and k.

In Section 4.3 we formally define the construction of f^{adj} by modifying f^{poly} over $\bar{I}$ using f^{emp}. We show in Lemma 11 that it suffices to select f^{poly} to be the polynomial estimator f^{adls} in (Acharya et al., 2017) or f^{surf} in (Hao et al., 2020b) to obtain Theorem 12 which shows that f^{adj} is a 2-factor approximation for $\mathcal{P}_d$.

4.2. Polynomial Histogram Approximation

We would first like to bound $\Delta_I(q)$ for any $q \in \mathcal{P}_d$ in terms of its ℓ_1 norm. From the Markov Brothers' inequality (Achieser, 1992), for any $q \in \mathcal{P}_d$,

$$\Delta_I(q) = \mathcal{O}(d+1)^2 \cdot \|q\|_{1,I},$$

and is achieved by the Chebyshev polynomial of degree-d. Instead, the next lemma shows that the bound can be improved for the interior of I. Its proof in Appendix B.1 carefully applies Markov Brothers' inequality over select sub-intervals of I based on the Bernstein's inequality. For simplicity, consider $I = [-1, 1]$.

Lemma 7. *For any $a \in [0, 1)$ and $q \in \mathcal{P}_d$,*

$$\Delta_{[-a,a]}(q) \le \int_{-a}^{a} |q'(x)|dx \le \frac{28(d+1)}{\sqrt{1-a^2}} \|q\|_{1,[-1,1]}.$$

We use the lemma and Equation 2 to construct a partition $\overline{[-1,1]}^{d,k}$ of $[-1, 1]$ such that $\forall J \in \overline{[-1,1]}^{d,k}$, $\Delta_J(p)$ is bounded by a small value. Note that the lemma's bound is weaker when a is close to the boundary of $(-1, 1)$, hence the parts of $\overline{[-1,1]}^{d,k}$ decrease roughly geometrically towards the boundary, ensuring $\Delta_J(p)$ is small over each. The geometric partition ensures that the number of intervals is still upper bounded by k as we show in Lemma 8.

Consider the positive half $[0, 1]$ of $[-1, 1]$. Given $\ell \ge 1$, let $m = \lceil \log_2(\ell(d+1)^2) \rceil$. For $1 \le i \le m$ define the intervals $I_i^+ = [1 - 1/2^{i-1}, 1 - 1/2^i)$, that together span $[0, 1 - 1/2^m)$, and let $E_m^+ \stackrel{\text{def}}{=} [1 - 1/2^m, 1]$ complete the partition of $[0, 1]$. Note that $|E_m^+| = 1/2^m \le 1/(\ell(d+1)^2)$. For each $1 \le i \le m$ further partition I_i^+ into $\lceil \ell(d+1)/2^{i/4} \rceil$ intervals of equal width, and denote this partition by $\bar{I}_i^+$. Clearly $\bar{I}^+ \stackrel{\text{def}}{=} (\bar{I}_1^+, \ldots, \bar{I}_m^+, E_m^+)$ partitions $[0, 1]$.

Define the mirror-image partition $\bar{I}^-$ of $[-1, 0]$, where, for example, we mirror the interval $[c, d)$ in $\bar{I}^+$ to $(-d, -c]$. The following lemma upper bounds the number of intervals in the combination of $\bar{I}^-$ and $\bar{I}^+$ and is proven in Appendix B.2.

Lemma 8. *For any degree $d \ge 0$ and $\ell > 0$, the number of intervals in $(\bar{I}^-, \bar{I}^+)$ is at most $4\ell(d+1)/(2^{1/4} - 1)$.*

The lemma ensures that we get the desired partition, $\overline{[-1,1]}^{d,k}$, with k intervals by setting

$$\ell \stackrel{\text{def}}{=} k(2^{1/4} - 1)/(4(d+1)). \tag{4}$$

For any interval $I = [a, b]$, we obtain $\overline{I}^{d,k}$ by a linear translation of $\overline{[-1,1]}^{d,k}$. For example, $[c, d) \in \overline{[-1,1]}^{d,k}$ translates to $[a + (b-a)(c+1)/2, a + (b-a)(d+1)/2) \in \overline{I}^{d,k}$.

Recall that $\bar{p}_{\overline{I}^{d,k}}$ denotes the histogram of p on $\overline{I}^{d,k}$. The following lemma, proven in Appendix B.3 using Equations (2), (4), and Lemma 7, shows that the ℓ_1 distance of any degree-d polynomial to its histogram on $\overline{I}^{d,k}$ is a factor $\mathcal{O}((d+1)/k)$ times than the ℓ_1 norm of the polynomial.

Lemma 9. *Given an interval I, for some universal constant $c_1 > 1$, for all $p \in \mathcal{P}_d$, and integer $k \ge 4(d+1)/(2^{1/4}-1)$,*

$$\|p - \bar{p}_{\overline{I}^{d,k}}\|_{1,I} \le c_1 \cdot (d+1) \cdot \|p\|_{1,I}/k.$$

We obtain our *split estimator* f^{adj} for a given polynomial estimator $f^{poly} \in \mathcal{P}_d$ as

$$f^{adj} \stackrel{\text{def}}{=} f^{adj}_{I,f^{poly},d,k} \stackrel{\text{def}}{=} f^{poly} + \bar{f}^{emp}_{\overline{I}^{d,k}} - \bar{f}^{poly}_{\overline{I}^{d,k}}$$

that over each subinterval $J \in \overline{I}^{d,k}$ adds to f^{poly} a constant so that its mass over J equals that of f^{emp}. Since $\overline{I}^{d,k}$ has k intervals, it follows that $f^{adj} \in \mathcal{P}_{k,d}$.

The next lemma (essentially) upper bounds the ℓ_1 distance of f^{adj} from any $p \in \mathcal{P}_d$ that is close to f^{poly} in $\mathcal{A}_k$ distance by the ℓ_1 distance of p from any function f that is close to f^{emp} in $\mathcal{A}_k$ distance.

Lemma 10. *For any interval I, functions f and polynomials $p, f^{poly} \in \mathcal{P}_d$, $d \ge 0$, and $k \ge 4(d+1)/(2^{1/4} - 1)$, $f^{adj} = f^{adj}_{I,f^{poly},d,k}$ satisfies*

$$\begin{aligned}&\|f^{adj} - p\|_{1,I}\\&\le \frac{c_1(d+1)}{k}\|f^{poly} - p\|_{1,I} + \|f - p\|_{1,I} + \|f^{emp} - f\|_{\mathcal{A}_k,I},\end{aligned}$$

where c_1 is the universal constant in Lemma 9.

Proof Consider an interval $J \in \overline{I}^{d,k}$ and let $\overline{f^{adj}}, \overline{f^{emp}}, \overline{f^{poly}}, \bar{p}$ respectively denote the histograms of $f^{adj}, f^{emp}, f^{poly}, p$ respectively over J.

$$\begin{aligned}&\|f^{adj} - p\|_{1,J}\\&\stackrel{(a)}{\le} \|f^{adj} - p - \left(\overline{f^{adj}} - \bar{p}\right)\|_{1,J} + \|\overline{f^{adj}} - \bar{p}\|_{1,J}\\&\stackrel{(b)}{=} \|f^{adj} - p - \left(\overline{f^{adj}} - \bar{p}\right)\|_{1,J} + \|\overline{f^{emp}} - \bar{p}\|_{1,J}\\&\stackrel{(c)}{=} \|f^{poly} - p - \left(\overline{f^{poly}} - \bar{p}\right)\|_{1,J} + \|\overline{f^{emp}} - \bar{p}\|_{1,J}\\&\stackrel{(d)}{\le} \|f^{poly} - p - \left(\overline{f^{poly}} - \bar{p}\right)\|_{1,J} + \|\bar{f} - \bar{p}\|_{1,J} + \|\overline{f^{emp}} - \bar{f}\|_{1,J},\end{aligned}$$

where (a) and (d) follow from the triangle inequality, (b) follows since $\overline{f^{adj}}$ has the same mass as f^{emp} by construction, it implies $\|\overline{f^{adj}} - \bar{p}\|_{1,J} = \|\overline{f^{emp}} - \bar{p}\|_{1,J}$, and (c) follows because $f^{adj} - \overline{f^{adj}} = f^{poly} - \overline{f^{poly}}$ since f^{adj} and f^{poly} differ by a constant in each J.

The proof is complete by summing over $J \in \overline{I}^{d,k}$ using the fact that $\overline{I}^{d,k}$ has at most k intervals, and since $f^{poly} - p \in \mathcal{P}_d$ over I, from Lemma 9, the sum

$$\sum_{J \in \overline{I}^{d,k}} \|f^{poly} - p - \left(\overline{f^{poly}_J} - \bar{p}_J\right)\|_{1,J} \le \frac{c_1(d+1)}{k}\|f^{poly} - p\|_{1,I}.$$

4.3. Applying the Estimator

In this more technical section, we show how to use existing estimators in place of f^{poly} to achieve Theorem 12. The next lemma follows from a straightforward application of triangle inequality to Lemma 10 as shown in Appendix B.4. It shows that given an estimate f^{poly} whose distance to f is a constant multiple of $\|f - \mathcal{P}_d\|_1$ plus $\|f^{emp} - f\|_{\mathcal{A}_k,I}$, depending on the value of k, f^{adj} has nearly the optimal approximation factor of 2 at the expense of the larger $\|f^{emp} - f\|_{\mathcal{A}_k,I}$.

Lemma 11. *Given an interval I, $f^{poly} \in \mathcal{P}_d$, such that for some constants $c', c'', \eta > 0$,*

$$\|f^{poly} - f\|_{1,I} \le c'\|f - \mathcal{P}_d\|_{1,I} + c''\|f^{emp} - f\|_{\mathcal{A}_k,I} + \eta,$$

and the parameter $k \ge 4(d+1)/(2^{1/4}-1)$, the estimator $f^{adj} = f^{adj}_{I,f^{poly},d,k}$ satisfies

$$\|f^{adj} - f\|_{1,I} \le \left(2 + \frac{c_2(c'+1)}{k}\right)\|f - \mathcal{P}_d\|_{1,I} + \frac{c_2\eta}{k} + \left(1 + \frac{c_2c''}{k}\right)\|f^{emp} - f\|_{\mathcal{A}_k,I},$$

where $c_2 = c_1(d+1)$ and c_1 is the constant from Lemma 9.

Prior works (Acharya et al., 2017; Hao et al., 2020b) derive a polynomial estimator that achieves a constant factor approximation for $\mathcal{P}_{t,d}$. We may thus use them as f^{poly} in the above lemma. In particular, the estimator f^{adls} in (Acharya et al., 2017) achieves $c' = 3$ and $c'' = 2$ and f^{surf} in (Hao et al., 2020b) achieves a $c' = c_d \ge 2$ and $c'' = c_d$, where c_d increases with the degree d (e.g., $c' < 3 \,\forall\, d \le 8$). Define

$$\eta_d \stackrel{\text{def}}{=} \sqrt{(d+1)/n} \tag{5}$$

and for any $0 < \gamma < 1$, let

$$k(\gamma) \stackrel{\text{def}}{=} \left\lceil 8c_1(d+1)/\gamma \right\rceil, \tag{6}$$

where c_1 is the constant from Lemma 10. We obtain the following theorem for $0 < \gamma < 1$ by using $f^{poly} = f^{adls}$ with $\eta = \eta_d(\gamma)$ in Lemma 11, and then applying Lemma 4, all over $I = [X_{(0)}, X_{(n)}]$, i.e. the interval between the least and the largest sample.

Theorem 12. *Given $X^n \sim f$, for any $0 < \gamma < 1$, the estimator $f^{adj} = f^{adj}_{I,f^{adls}(\eta_d),d,k(\gamma)}$ for $I = [X_{(0)}, X_{(n)}]$, achieves*

$$\mathbb{E}\,\|f^{adj} - f\|_1 \le (2+\gamma)\|f - \mathcal{P}_d\|_1 + \mathcal{O}\left(\sqrt{\frac{d+1}{\gamma \cdot n}}\right).$$

We prove the above theorem in Appendix B.5, showing that f^{adj} is a 2-factor approximation for $\mathcal{P}_d$. Notice that when $\|f - \mathcal{P}_d\|_1 \gg \mathcal{O}\left(\sqrt{(d+1)/n}\right)$, as is the case when $n \nearrow \infty$, Theorem 12 gives f^{adj} a lower ℓ_1-distance bound to f than f^{adls}. We use the above procedure in the main TURF routine that we describe in the next section.

5. A 2-Factor Estimator for $\mathcal{P}_{t,d}$

In the previous section, we described an estimator that approximates a distribution to a distance only slightly larger than twice $\|f - \mathcal{P}_{1,d}\|_1$. We now extend this result to $\mathcal{P}_{t,d}$. Consider a $p^* \in \mathcal{P}_{t,d}$ that achieves $\|f - p^*\|_1 = \|f - \mathcal{P}_{t,d}\|_1$. If the t intervals corresponding to the different polynomial pieces of p^* are known, we may apply the routine in Section 4 to each interval and combine the estimate to obtain the 2-factor approximation for $\mathcal{P}_{t,d}$.

However, as these intervals are unknown, we instead use the partition returned by the ADLS routine in (Acharya et al., 2017). ADLS returns a partition with βt intervals where the parameter $\beta > 1$ by choice. Among these βt intervals, $p^* \in \mathcal{P}_d$ is not a degree-d polynomial in at most t intervals. Let I be an interval in this partition where p^* has more than one piece. The ADLS routine has the property that there are at-least $(\beta - 1)t$ other intervals in the partition in which p^* is a single-piece polynomial with a worse $\mathcal{A}_{d+1}$ distance to f. That is, for any interval J in the $(\beta - 1)t$ interval collection,

$$\|f - p^*\|_{\mathcal{A}_{d+1},I} \le \|f - p^*\|_{\mathcal{A}_{d+1},J} + \|f^{emp} - f\|_{\mathcal{A}_{d+1},J\cup I} + \eta.$$

This is used to bound the ℓ_1 distance in these intervals.

Our main routine TURF consists of simply applying the transformation discussed in Section 4 to the partition returned by the ADLS routine in (Acharya et al., 2017). Given samples X^n, the number of pieces-$t \ge 1$, degree-$d \ge 0$, for any $0 < \alpha < 1$, we first run the ADLS routine with input X^n, f^{emp}, and parameters t, d,

$$\beta = \beta(\alpha) \stackrel{\text{def}}{=} 1 + \frac{4k(\alpha)}{\alpha(d+1)}, \tag{7}$$

where $k(\alpha)$ is as defined in Equation (6), and $\eta_d = \sqrt{(d+1)/n}$. ADLS returns a partition $\bar{I}_{\text{ADLS}}$ of $\mathbb{R}$ with $2\beta t$ intervals and a degree-d, $2\beta t$-piecewise polynomial defined over the partition. For any interval $I \in \bar{I}_{\text{ADLS}}$, let f^{adls}_I denote the degree-d estimate output by ADLS over this interval. We obtain our output estimate $f^{out}_{t,d,\alpha}$ by applying the routine in Section 4 to f^{adls}_I for each $I \in \bar{I}_{\text{ADLS}}$ with $k = k(\alpha)$ (ref. Equation (6)). This is summarized below in Algorithm 1.

Algorithm 1 TURF

Input: X^n, t, d, α
$k \leftarrow \lceil 8c_1(d+1)/\alpha \rceil$ {c_1 is the constant in Lemma 9}
$\beta \leftarrow 1 + 4k/(\alpha(d+1))$
$\eta_d \leftarrow \sqrt{(d+1)/n}$
$\bar{I}_{\text{ADLS}}, \left(f^{adls}_I, I \in \bar{I}_{\text{ADLS}}\right) \leftarrow \text{ADLS}(X^n, t, d, \beta, \eta_d)$
Output: $f^{out}_{t,d,\alpha} \leftarrow \left(f^{adj}_{I,f^{adls}_I,d,k}, I \in \bar{I}_{\text{ADLS}}\right)$

Theorem 13 shows that $f^{out}_{t,d,\alpha}$ is a min-max 2-factor approximation for $\mathcal{P}_{t,d}$. We have a $\mathcal{O}(1/\alpha^{3/2})$ term in the 'variance' term in Theorem 13 that reflects the $\mathcal{O}(t/\alpha^{3/2})$ pieces in the output estimate. A small α corresponds to a low-bias, high-variance estimator with many pieces, and vice-versa. Note that the $3/2$ exponent here is larger than the corresponding $1/2$ in the result for $\mathcal{P}_d$ in Section 4 (Theorem 12). The increased exponent over $\mathcal{P}_d$ is due to the unknown locations of the polynomial pieces of $p^* \in \mathcal{P}_{t,d}$. Obtaining the exact exponent for 2-factor approximation for various classes may be an interesting question but beyond the scope of this paper. Let $\omega < 3$ be the matrix multiplication constant. As our transformation of f^{adls} takes $\mathcal{O}(n)$ time, the overall time complexity is the same as ADLS's near-linear $\tilde{\mathcal{O}}(nd^{3+\omega})$.

Theorem 13. *Given $X^n \sim f$, an integer number of pieces $t \geq 1$ and degree $d \geq 0$, the parameter $\alpha \geq 0$, $f^{out}_{t,d,\alpha}$ is returned by* TURF *in $\tilde{\mathcal{O}}(nd^{3+\omega})$ time such that*

$$\mathbb{E}\,\|f^{out}_{t,d,\alpha} - f\|_1 \leq (2+\alpha)\|f - \mathcal{P}_{t,d}\|_1 + \mathcal{O}\left(\sqrt{\frac{t(d+1)}{\alpha^3 n}}\right).$$

Theorem 13 is proven in Appendix C.1 and follows from the following lemma via a simple application of the VC inequality in Lemma 4 and using Property 5, 6. We prove the lemma in Appendix C.2.

Lemma 14. *Given samples $X^n \sim f$ for some $n \geq 1$, parameters $t \geq 1, d \geq 0$ and for $0 < \alpha < 1$, $f^{out}_{t,d,\alpha}$ returned by* TURF *satisfies*

$$\begin{aligned}\|f^{out}_{t,d,\alpha} - f\|_1 \leq& \left(3 + 2c_1 + \frac{2}{\beta - 1}\right)\|f^{emp} - f\|_{\mathcal{A}_{2\beta t\cdot k}} \\ &+ \left(2 + \frac{4c_1(d+1)}{k} + \frac{1 + k/(d+1)}{\beta - 1}\right)\|f - \mathcal{P}_{t,d}\|_1 \\ &+ \left(\frac{c_1(d+1)}{k} + \frac{k}{(\beta-1)(d+1)}\right)\eta_d,\end{aligned}$$

where c_1, $k = k(\alpha)$, $\beta = \beta(\alpha)$, are the constants in Lemma 9 and Equations (6), and (7) respectively, and $\eta_d \stackrel{\text{def}}{=} \sqrt{(d+1)/n}$.

6. Optimal Parameter Selection

Like many other statistical learning problems, learning distributions exhibits a fundamental trade-off between bias and variance. In Equation (1) increasing the parameters t and d enlarges the polynomial class $\mathcal{P}_{t,d}$, hence decreases the bias term $\|f - \mathcal{P}_{t,d}\|_1$ while increasing the variance term $\mathcal{O}(\sqrt{t(d+1)/n})$. As the number of samples n increases, asymptotically, it is always better to opt for larger t and d. Yet for any given n, some parameters t and d yield the smallest error. We consider the parameters minimizing the upper bound in Theorem 13.

6.1. Context and Results

For several popular structured distributions such as unimodal, log-concave, Gaussian, and their mixtures, low-degree polynomials, e.g. $d \leq 8$, are essentially optimal (Birgé, 1987; Chan et al., 2014; Hao et al., 2020b). Yet for the same classes, the range of the optimal t is large, between $\Theta(1)$ and $\Theta(n^{1/3})$. Therefore, for a given d, we seek the t minimizing the error upper bound in Equation (1).

In the next subsection, we describe a parameter selection algorithm that improves this result for the estimators we considered in the previous section. Following nearly identical steps as in the derivation of Theorem 13 from Lemma 14, and using the probabilistic version of the VC Lemma 4 (see (Devroye & Lugosi, 2012)), it may be shown that with high probability $f^{out}_{t,d,\alpha}$ is a c-factor approximation for $\mathcal{P}_{t,d}$. Namely, for any $\delta \geq 0$,

$$\|f^{out}_{t,d,\alpha} - f\|_1 \leq c \cdot \|f - \mathcal{P}_{t,d}\|_1 + \mathcal{O}\left(\sqrt{(t(d+1) + \log 1/\delta)/n}\right), \tag{8}$$

where $c = c(\alpha)$ is a function of the chosen α. We use the estimates $f^{out}_{t,d,\alpha}$ to find an estimate t^{est} such that $f^{out}_{t^{est},d,\alpha}$ has an error comparable to the c-factor approximation for $\mathcal{P}_{t,d}$ with the best t.

Theorem 15. *Given $n \in 2^{\mathcal{N}}$, $d \geq 0$, $0 < \alpha < 1$, c-factor estimates for $\mathcal{P}_{t,d}$ in high probability (see Equation (8)), $\{f^{out}_{t,d,\alpha} : 1 \leq t \leq n\}$, for any $0 < \beta < 1$, we find the estimate t^{est} such that w.p. $\geq 1 - \delta \cdot \log n$,*

$$\|f^{out}_{t^{est},d,\alpha} - f\|_1 \leq \min_{t \geq 1}\left((1+\beta)\cdot c \cdot \|f - \mathcal{P}_{t,d}\|_1 + \mathcal{O}\left(\sqrt{(t(d+1) + \log 1/\delta)/(\beta^2 n)}\right)\right).$$

The proof, provided in Appendix D.1, exploits the fact that the bias term of $f^{out}_{t,d,\alpha}$ is at most $c \cdot \|f - \mathcal{P}_{t,d}\|_1$, which decreases with t, and the variance term upper bounded by $\mathcal{O}(\sqrt{(t(d+1) + \log 1/\delta)/n})$, increasing with t.

6.2. Construction

We use the following algorithm derived in (Jain et al., 2022). Consider the set $\mathcal{V} = \{v_1, v_2, , \ldots, v_k\} \subseteq \mathcal{M}$, an *unknown* target $v \in \mathcal{M}$, an *unknown* non-increasing sequence b_i, and a *known* non-decreasing sequence c_i such that $d(v_i, v) \leq b_i + c_i\ \forall i$.

First consider selecting for a given $1 \leq i < j \leq k$, the point among v_i, v_j that is closer to v. Suppose for some constant $\gamma > 0$, $d(v_i, v_j) \leq \gamma c_j$. Then from the triangle inequality, $d(v_i, v) \leq d(v_j, v) + d(v_i, v_j) \leq b_j + c_j + \gamma c_j \leq b_j + (1+\gamma)c_j$. On the other hand if $d(v_i, v_j) > \gamma c_j$, since $b_j \leq b_i$ (as $j > i$), $d(v_j, v) \leq b_j + c_j \leq b_i + c_j \leq b_i + d(v_i, v_j)/\gamma$.

Therefore if we set γ to be sufficiently large and select $v'_\gamma = v_i$ if $d(v_i, v_j) \leq \gamma c_j$, and otherwise set $v'_\gamma = v_j$, we roughly obtain $d(v'_\gamma, v) \lesssim b_i + (1+\gamma)c_j$. We now generalize this approach to selecting between all points in $\mathcal{V}$. Let i_γ be the smallest index in $\{1, \ldots, k\}$ such that $\forall i_\gamma < i \leq k$,

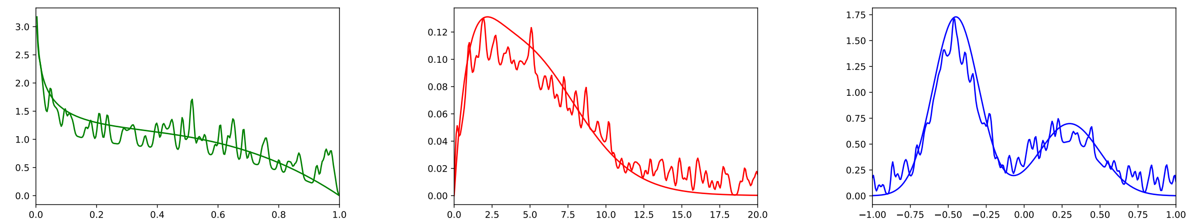

Figure 1. The Beta, Gamma and Gaussian mixtures, respectively. The smooth and coarse plots in each sub-figure correspond to the noise-free and noisy cases, respectively.

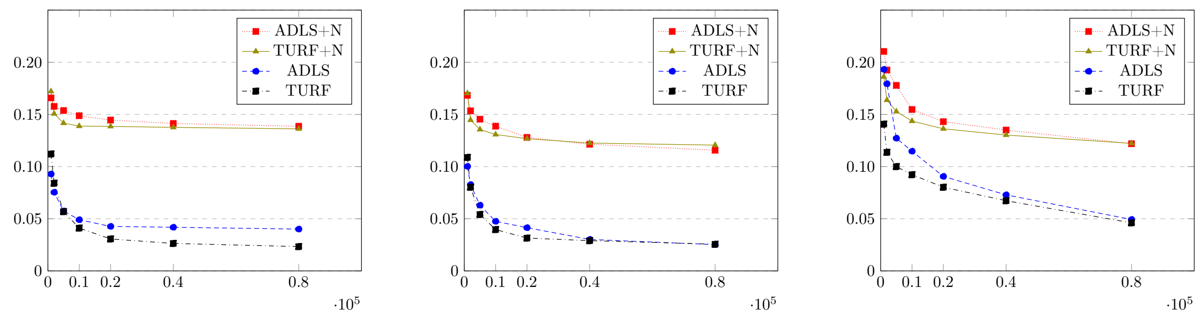

Figure 2. ℓ_1 error versus number of samples on the Beta, Gamma, and Gaussian mixtures respectively in Figure 1 for $d = 1$.

$d(v_{i_\gamma}, v_j) \le \gamma c_i$. Lemma 16 shows the favorable properties of v_{i_γ}, for example, that for a sufficiently large γ, $d(v_{i_\gamma}, v)$ is comparable to $\min_{i\in\{1,\ldots,k\}}(b_i+\lambda c_i)$ when $b_i \gg c_i$. The proof may be found in Appendix D.2.

Lemma 16. *Given a set $\mathcal{V} = \{v_1, v_2, \ldots, v_k\}$ in a metric space $(\mathcal{M}, d)$, a sequence $0 \le c_1 \le c_2 \ldots \le c_k$, and $\gamma > 2$, let $1 \le i_\gamma \le k$ be the smallest index such that for all $i_\gamma < i \le k$, $d(v_i, v_{i_\gamma}) \le \gamma c_i$. Then for all sequences $b_1 \ge b_2 \ge \ldots \ge b_k \ge 0$ such that for all i, $d(v_i, v) \le b_i + c_i$,*

$$d(v_{i_\gamma}, v) \le \min_{j\in\{1,\ldots,k\}} \left(\left(1 + \frac{2}{\gamma - 2}\right) \cdot b_j + (\gamma + 1)c_j\right).$$

The set of real integrable functions with TV distance forms a metric space. For simplicity for the given $d \ge 0$, $0 < \alpha < 1$, and n, denote $f_t^{out} \stackrel{\text{def}}{=} f_{t,d,\alpha}^{out}$. Assume n is a power of 2 and let $\mathcal{V} = \{f_1^{out}, f_2^{out}, f_4^{out}, \ldots, f_n^{out}\}$. Suppose for constants $c', c'' > 0$, and a chosen $0 < \delta < 1$, $\sqrt{c't + c'' \log 1/\delta}$ is f_t^{out}'s variance term. For any chosen $0 < \beta < 1$, we obtain t_β^{est} by applying the above method with $\mathcal{V}$ and the c_is corresponding to f_t^{out} as $\sqrt{c't + c'' \log 1/\delta}$. That is, t_β^{est} is the smallest $t \in \mathcal{I} = \{1, 2, 4, \ldots, n\}$ such that $\forall j \in \mathcal{I} : j \ge t$, $d(f_t^{out}, f_j^{out}) \le \gamma\sqrt{c'j + c'' \log 1/\delta}$ (where we select $\gamma = \gamma(\beta) = 2 + 2/\beta$). In Section 7, we experimentally evaluate the TURF estimator and the cross-validation technique.

7. Experiments

Direct comparison of TURF and ADLS for a given t, d is not straightforward as TURF outputs polynomials consisting of more pieces. To compare the algorithms more equitably, we apply the cross-validation technique in Section 6 to select the best t for each. The cross validation parameter δ is chosen to reflect the actual number of pieces output by ADLS and TURF. Note that while SURF (Hao et al., 2020a) is another piecewise polynomial based estimation method, it has an implicit method to cross-validate t, unlike ADLS and TURF. As comparisons against SURF may only reflect the relative strengths of the cross validation methods and not that of the underlying estimation procedure, we defer them to Appendix E. All experiments compare the ℓ_1 error, run for n between 1,000 and 80,000, and averaged over 50 runs. For ADLS we use the code provided in (Acharya et al., 2017), and for TURF we use the algorithm in Section 5.

The experiments consider the structured distributions addressed in (Acharya et al., 2017), namely mixtures of Beta: $.4\text{B}(.8, 4) + .6\text{B}(2, 2)$, Gamma: $.7\Gamma(2, 2) + .3\Gamma(7.5, 1)$, and Gaussians: $.65\mathcal{N}(\text{-}.45, .15^2)$+$.35\mathcal{N}(.3, .2^2)$ as shown in Figure 1. Figure 2 considers approximation relative to $\mathcal{P}_{t,1}$. The blue-dashed and the black-dot-dash plots show that TURF modestly outperforms ADLS. It is especially significant for the Beta distribution as $\text{B}(.8, 4)$ has a large second derivative near 0, and approximating it may require many degree-1 pieces localized to that region. For this lower width region, the $\mathcal{A}_1$ distance may be too small to warrant many pieces in ADLS, unlike in TURF that forms intervals guided by shape constraints e.g., based on Lemma 7.

We perturb these distribution mixtures to increase their bias. For a given $k > 0$, select $\bar{\mu}_k \stackrel{\text{def}}{=} (\mu_1, \ldots, \mu_k)$ by independently choosing μ_i, $i \in \{1, \ldots, k\}$ uniformly from the effective support of f (we remove 5% tail mass on either

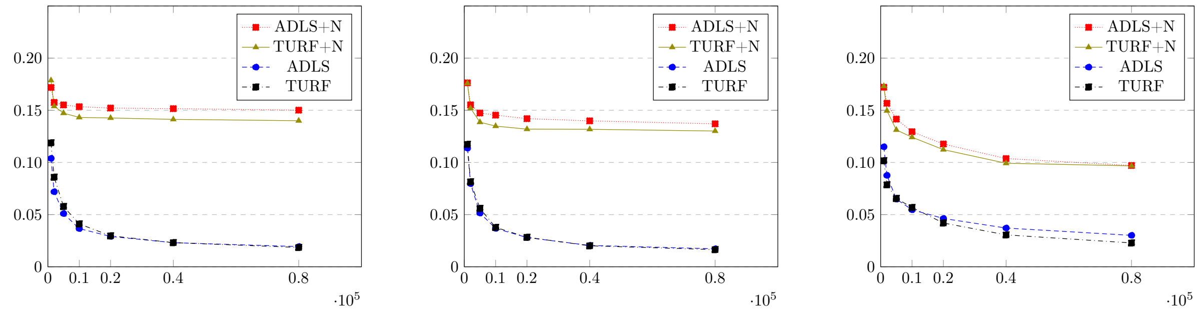

Figure 3. ℓ_1 error versus number of samples on the Beta, Gamma, and Gaussian mixtures respectively in Figure 1 for $d = 2$.

side). At each of these locations, apply a Gaussian noise of magnitude $0.25/k$ with standard deviation $\sigma = c_2/k$, for some constant $c_2 > 0$ that is chosen to scale with the effective support width of f. That is,

$$f_{\bar{\mu}} \stackrel{\text{def}}{=} \frac{3}{4} \cdot f + \frac{1}{4} \cdot \sum_{i=1}^{k} \frac{1}{k} \cdot \mathcal{N}\left(\mu_i, \frac{c_2^2}{k^2}\right).$$

We choose $k = 100$ and $c_2 = 0.05, 1, 0.1$ for the Beta, Gamma and Gaussian mixtures respectively, to yield the distributions shown in Figure 1. The red-dotted and olive-solid plots in Figure 2 compares ADLS and TURF on these distributions. While the overall errors are larger due to the added noise, TURF outperforms ADLS on nearly all distributions. A consistent trend across our experiments is that for large n, the performance gap between ADLS and TURF decreases. This may be explained by the fact that as n increases, the value of t output by the cross-validation method also increases, reducing the bias under both ADLS and TURF. However, the reduction in ADLS's bias is more significant due to its larger approximation factor compared to TURF, resulting in the smaller gap.

Figure 3 repeats the same experiments for $d = 2$. Increasing the degree leads to lower errors on both ADLS and TURF in the non-noisy case. However, the larger bias in the noisy case reveals the improved performance of TURF.

References

Acharya, J., Jafarpour, A., Orlitsky, A., and Suresh, A. T. Near-optimal-sample estimators for spherical gaussian mixtures. *arXiv preprint arXiv:1402.4746*, 2014.

Acharya, J., Diakonikolas, I., Li, J., and Schmidt, L. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1278–1289. SIAM, 2017.

Achieser, N. *Theory of Approximation*. Dover books on advanced mathematics. Dover Publications, 1992. ISBN 9780486671291.

Ashtiani, H., Ben-David, S., Harvey, N. J., Liaw, C., Mehrabian, A., and Plan, Y. Nearly tight sample complexity bounds for learning mixtures of gaussians via sample compression schemes. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 3416–3425, 2018.

Birgé, L. Estimating a density under order restrictions: Nonasymptotic minimax risk. *The Annals of Statistics*, pp. 995–1012, 1987.

Bithell, J. F. An application of density estimation to geographical epidemiology. *Statistics in medicine*, 9(6): 691–701, 1990.

Bousquet, O., Kane, D., and Moran, S. The optimal approximation factor in density estimation. *arXiv preprint arXiv:1902.05876*, 2019.

Bousquet, O., Braverman, M., Efremenko, K., Kol, G., and Moran, S. Statistically near-optimal hypothesis selection. *arXiv preprint arXiv:2108.07880*, 2021.

Chan, S.-O., Diakonikolas, I., Servedio, R. A., and Sun, X. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pp. 604–613. ACM, 2014.

Cohen, D., Kontorovich, A., and Wolfer, G. Learning discrete distributions with infinite support. *Advances in Neural Information Processing Systems*, 33:3942–3951, 2020.

Devroye, L. and Gyorfi, L. No empirical probability measure can converge in the total variation sense for all distributions. *The Annals of Statistics*, pp. 1496–1499, 1990.

Devroye, L. and Lugosi, G. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2012.

Diakonikolas, I. Learning structured distributions. *Handbook of Big Data*, pp. 267, 2016.

Gerber, M. S. Predicting crime using twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

Givens, G. H. and Hoeting, J. A. *Computational statistics*, volume 703. John Wiley & Sons, 2012.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

Han, Y., Jiao, J., and Weissman, T. Minimax estimation of discrete distributions under ℓ_1 loss. *IEEE Transactions on Information Theory*, 61(11):6343–6354, 2015.

Hao, Y. and Orlitsky, A. Unified sample-optimal property estimation in near-linear time. *Advances in Neural Information Processing Systems*, 32, 2019.

Hao, Y., Jain, A., Orlitsky, A., and Ravindrakumar, V. Surf: A simple, universal, robust, fast distribution learning algorithm. *Advances in Neural Information Processing Systems*, 33:10881–10890, 2020a.

Hao, Y., Jain, A., Orlitsky, A., and Ravindrakumar, V. Surf: A simple, universal, robust, fast distribution learning algorithm. In *NeurIPS*, 2020b.

Huber, P. J. Robust estimation of a location parameter. In *Breakthroughs in statistics*, pp. 492–518. Springer, 1992.

Jain, A., Orlitsky, A., and Ravindrakumar, V. Robust estimation algorithms don't need to know the corruption level. *arXiv preprint arXiv:2202.05453*, 2022.

Kamath, S., Orlitsky, A., Pichapati, D., and Suresh, A. T. On learning distributions from their samples. In *Conference on Learning Theory*, pp. 1066–1100. PMLR, 2015.

Pearson, K. X. contributions to the mathematical theory of evolution.—ii. skew variation in homogeneous material. *Philosophical Transactions of the Royal Society of London.(A.)*, (186):343–414, 1895.

Pimentel, M. A., Clifton, D. A., Clifton, L., and Tarassenko, L. A review of novelty detection. *Signal Processing*, 99: 215–249, 2014.

Rahman, Q. I., Schmeisser, G., et al. *Analytic theory of polynomials*. Number 26 in London Mathematical Society monographs. Clarendon Press, 2002.

Scott, D. W. Multivariate density estimation and visualization. In *Handbook of computational statistics*, pp. 549–569. Springer, 2012.

Silverman, B. W. *Density Estimation for Statistics and Data Analysis*, volume 26. CRC Press, 1986.

Tukey, J. W. Exploratory data analysis. *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, 1977.

Vapnik, V. N. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Wolfowitz, J. The minimum distance method. *The Annals of Mathematical Statistics*, pp. 75–88, 1957.

Yatracos, Y. G. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, pp. 768–774, 1985.

Zambom, A. Z. and Ronaldo, D. A review of kernel density estimation with applications to econometrics. *International Econometric Review*, 5(1):20–42, 2013.

A. Proofs for Section 1

A.1. Proof of Lemma 3

Proof

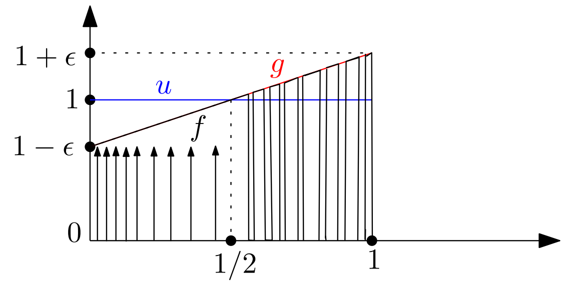

Figure 4. f that is indistinguishable from u in the proof of Lemma 3

Note that for any $t_1 \le t$ and $d_1 \le d$, $\mathcal{P}_{t_1,d_1} \subseteq \mathcal{P}_{t,d}$. Therefore $c_{t,d}$ increases with t, d. (Chan et al., 2014) showed that $c_{2,0} \ge 2$. We show below that that $c_{1,1} \ge 2$. Together they imply $c_{t,d} \ge 2$ when $t \ge 2$ or $d \ge 1$.

Let u be the uniform distribution on $[0, 1]$. Fix an $\epsilon > 0$ and consider the distribution $g(x) = 1 - \epsilon + 2\epsilon x$. Note that $g \in \mathcal{P}_{1,1}$. For a fixed $k \ge 1$, we construct two random distributions f_k and f'_k that are essentially indistinguishable using n samples as k is made large, and such that all functions have an ℓ_1 distance to either f_k or f'_k that is at least twice as far as their respective ℓ_1 approximations in $\mathcal{P}_{1,1}$.

To construct f_k we perturb g separately on the left, $[0, 1/2]$, and right, $(1/2, 1]$, halves of $[0, 1]$.

For the left half, we use the discrete sub-distribution h_k that assigns a mass $\epsilon/4 \cdot 1/k$ to k values drawn according to the distribution $h(x) = 4 - 8x$ over $[0, 1/2]$. Then let

$$f_k(x) = g(x) + h_k(x), \quad \text{for} \quad x \in [0, 1/2].$$

Thus f_k consists of discrete atoms added to g on $[0, 1/2]$.

For the right half, assuming Wolog that k is even, first partition $(1/2, 1]$ into $k/2$ intervals of width $1/k$ by letting $I_i \stackrel{\text{def}}{=} (1/2 + (i - 1)/k, 1/2 + i/k]$ for $i \in \{1, \ldots, k/2\}$. Let $|I|$ denote the width of interval I. For each $i \in \{1, \ldots, k/2\}$, select a random circular sub-interval $J_i \subseteq I_i$ of width

$$w_i = \frac{1}{1 + 2\epsilon i/k} \cdot |I_i|$$

as follows: Suppose $I_i = [a_i, b_i]$ for simplicity. Choose a point x_i uniformly at random in I_i and define

$$J_i \stackrel{\text{def}}{=} (a_i, a_i + \max\{0, x_i + w_i - b_i\}] \cup [x_i, \min\{x + w_i, b_i\}].$$

Let f_k be g over J_i and 0 over $I_i \setminus J_i$, hence as illustrated in Figure 4, for $x \in (1/2, 1]$,

$$f_k(x) \stackrel{\text{def}}{=} \begin{cases} g(x) & x \in J_i, \quad i \in \{1, \ldots, k/2\}, \\ 0 & x \in I_i \setminus J_i, \quad i \in \{1, \ldots, k/2\}. \end{cases}$$

It is easy to show that on any sub-interval of I_i, the area of f_k is within that of u up to an additive $\mathcal{O}(\epsilon/k)$.

Construct f'_k via the same method for f_k but mirrored along $1/2$, with $g' = 1 + \epsilon - 2\epsilon x$, adding atoms to $(1/2, 1]$ and alternating between g' and 0 on $[0, 1/2]$ as described in the construction of f_k.

By a birthday paradox type argument, for any $\delta \ge 0$, it is easy to see that the distributions f_k, f'_k are indistinguishable from u with probability $\ge 1 - \delta$ using any finitely many n samples (by choosing an appropriately large $k = k(\delta, n)$). Thus w.p.

$\geq 1-2\delta$, the estimate $f^{est} = f^{est}_{X^n}$ is identical under both f_k and f'_k. Therefore any estimator f^{est} suffers a factor

$$c \geq (1-2\delta) \cdot \min_{f^{est}} \max \left\{ \frac{\|f^{est} - f_k\|_1}{\|f_k - \mathcal{P}_{1,1}\|_1}, \frac{\|f^{est} - f'_k\|_1}{\|f'_k - \mathcal{P}_{1,1}\|_1} \right\},$$

By the mirror image symmetry between f_k and f'_k about 1/2, $f^{est} = u$ is the optimal estimate to within an additive $\mathcal{O}(\epsilon/k)$. This lower bounds c as

$$\begin{aligned}
\frac{c}{1-2\delta} \geq \frac{\|f_k - u\|_1 - \mathcal{O}(\epsilon/k)}{\|f_k - \mathcal{P}_{1,1}\|_1} &\overset{(a)}{\geq} \frac{\|f_k - u\|_1 - \mathcal{O}(\epsilon/k)}{\|f_k - g\|_1} \\
&= \frac{\|f_k - u\|_{1,[0,1/2]} + \|f_k - u\|_{1,(1/2,1]} - \mathcal{O}(\epsilon/k)}{\|f_k - g\|_{1,[0,1/2]} + \|f_k - g\|_{1,(1/2,1]}} \\
&\overset{(b)}{=} \frac{\|h_k\|_{1,[0,1/2]} + \|g - u\|_{1,[0,1/2]} + \|f_k - u\|_{1,(1/2,1]} - \mathcal{O}(\epsilon/k)}{\|h_k\|_{1,[0,1/2]} + \|f_k - g\|_{1,(1/2,1]}} \\
&\overset{(c)}{=} \frac{\epsilon/4 + \epsilon/4 + \|f_k - u\|_{1,[1/2,1]} - \mathcal{O}(\epsilon/k)}{\epsilon/4 + \|f_k - g\|_{1,(1/2,1]}} \\
&\overset{(d)}{=} \frac{\epsilon/4 + \epsilon/4 + \|f_k - u\|_{1,[1/2,1]} - \mathcal{O}(\epsilon/k)}{\epsilon/4 + \int_{1/2}^1 g(x)dx - \int_{1/2}^1 f_k(x)dx} \\
&= \frac{\epsilon/4 + \epsilon/4 + \|f_k - u\|_{1,[1/2,1]} - \mathcal{O}(\epsilon/k)}{\epsilon/4 + 1/2 + \epsilon/4 - \int_{1/2}^1 f_k(x)dx} \\
&\overset{(e)}{=} \frac{\epsilon/4 + \epsilon/4 + \|f_k - u\|_{1,[1/2,1]} - \mathcal{O}(\epsilon/k)}{\epsilon/2 + \mathcal{O}(\epsilon/k)} \\
&\overset{(f)}{\geq} \frac{\epsilon/4 + \epsilon/4 + \epsilon/2 - \mathcal{O}(\epsilon/k)}{\epsilon/2 + \mathcal{O}(\epsilon/k)} = 2 - \mathcal{O}\left(\frac{1}{k}\right),
\end{aligned}$$

where (a) follows since $g \in \mathcal{P}_{1,1}$, (b) follows since h_k is a discrete distribution, (c) follows since h_k has a total mass $\epsilon/4$ and since $\|g-u\|_{1,[0,1/2]} = \epsilon/4$ by a straightforward calculation, (d) follows since $g \geq f_k$ in $(1/2, 1]$, (e) follows since the area of f_k and u on $I = (1/2, 1]$ are equal to within an additive $\mathcal{O}(\epsilon/k)$, and (f) follows since $\|f_k - u\|_{1,(1/2,1]} \geq 2\|f_k - g\|_{1,(1/2,1]} - \mathcal{O}(\epsilon/k)$. Choosing $\delta \searrow 0$ and $k \nearrow \infty$ completes the proof.

A.2. Description and proof of Lemma 17

The following lemma shows that if f^{est} is a c-factor approximation for $\mathcal{P}_{t,d}$ for some t and d and achieves the min-max rate of a distribution class $\mathcal{C}$, then f^{est} is also a c-factor approximation for $\mathcal{C}$.

Lemma 17. *If f^{est} is a c-factor approximation for $\mathcal{P}_{t,d}$ and for all f in a class $\mathcal{C}$, $c \cdot \|f - \mathcal{P}_{t,d}\|_1 + \mathcal{O}(\mathcal{R}_n(\mathcal{P}_{t,d})) \leq \mathcal{O}(\mathcal{R}_n(\mathcal{C}))$, then for any f, not necessarily in $\mathcal{C}$,*

$$\|f^{est} - f\|_1 \leq c \cdot \|f - \mathcal{C}\|_1 + \mathcal{O}(\mathcal{R}_n(\mathcal{C})).$$

Proof For a distribution g and class $\mathcal{D}$, let $g_{\mathcal{D}} \in \mathcal{D}$ be the closest approximation to g from $\mathcal{D}$, namely achieving $\|g - g_{\mathcal{D}}\|_1 = \|g - \mathcal{D}\|_1$. Then for any distribution f,

$$\begin{aligned}
\|f^{est} - f\|_1 &\leq c \cdot \|f - \mathcal{P}_{t,d}\|_1 + \mathcal{O}(\mathcal{R}_n(\mathcal{P}_{t,d})) \\
&\overset{(a)}{\leq} c \cdot \|f - f_{\mathcal{C}_{\mathcal{P}_{t,d}}}\|_1 + \mathcal{O}(\mathcal{R}_n(\mathcal{P}_{t,d})) \\
&\leq c \cdot \|f - f_{\mathcal{C}}\|_1 + c \cdot \|f_{\mathcal{C}_{\mathcal{P}_{t,d}}} - f_{\mathcal{C}}\|_1 + \mathcal{O}(\mathcal{R}_n(\mathcal{P}_{t,d})) \\
&\overset{(b)}{=} c \cdot \|f - \mathcal{C}\|_1 + c \cdot \|f_{\mathcal{C}} - \mathcal{P}_{t,d}\|_1 + \mathcal{O}(\mathcal{R}_n(\mathcal{P}_{t,d})) \\
&\overset{(c)}{=} c \cdot \|f - \mathcal{C}\|_1 + \mathcal{O}(\mathcal{R}_n(\mathcal{C})),
\end{aligned}$$

where in (a), just as $f_{\mathcal{C}}$ is the $\mathcal{C}$ distribution closest to f, $f_{\mathcal{C}_{\mathcal{P}_{t,d}}}$ is the $\mathcal{P}_{t,d}$ distribution closest to $f_{\mathcal{C}}$ and the inequality follows since $f_{\mathcal{C}_{\mathcal{P}_{t,d}}} \in \mathcal{P}_{t,d}$ and by definition, $\|f - \mathcal{P}_{t,d}\|_1$ is the least distance from f to any $q \in \mathcal{P}_{t,d}$, (b) follows since by definition, $f_{\mathcal{C}_{\mathcal{P}_{t,d}}}$ is the best approximation to $f_{\mathcal{C}}$ from $\mathcal{P}_{t,d}$, and (c) follows from the property of $\mathcal{C}$ considered in the lemma as $f_{\mathcal{C}} \in \mathcal{C}$.

B. Proofs for Section 4

B.1. Proof of Lemma 7

Proof Observe that for any $q \in \mathcal{P}_d$ and interval J,

$$\Delta_J(q) \overset{\text{def}}{=} \max_{x\in J} q(x) - \min_{x \in J} q(x) \leq \int_J |q'(t)|dt,$$

where q' denotes the first derivative of q. The case of $d = 0$ is trivial. We give a proof for $d \geq 1$. Consider

$$f_d(x) \overset{\text{def}}{=} \frac{1}{x}\sin(d \arcsin(x)).$$

The following two claims (indicated with the overset (*)) may be verified via Wolfram Mathematica e.g. version 12.3.
Claim 1 $||f_d||_2 \leq \sqrt{\pi\, d}$, because

$$||f_d||_2^2 \overset{(*)}{=} 2\int_0^1 \frac{\sin^2(d\arcsin(x))}{x^2}\, dx \leq \int_0^{\frac{\pi}{2}} \frac{1-\cos(2d\theta)}{\sin^2\theta}\, d\theta = \pi\, d.$$

Claim 2 $f_d(0) \overset{(*)}{=} d$ and $f_d'(0) \overset{(*)}{=} 0$.
Let $f(t) \overset{\text{def}}{=} q(t)\sqrt{1-t^2}$ and note that for any $x \in [0,1]$,

$$\sqrt{1-x^2}\int_{-x}^{x}|q'(t)|\,dt \leq \int_{-1}^{1}|q'(t)|\sqrt{1-t^2}\,dt \leq \int_{-1}^{1}|f'(t)|\,dt + \int_{-1}^{1}\frac{|t\,q(t)|}{\sqrt{1-t^2}}, dt,$$

where the final inequality follows by integration by parts.

For simplicity let $g_d(x) \overset{\text{def}}{=} \sqrt{\frac{2}{\pi d}} \cdot f_{\frac{d}{2}}(x)$, then $g_d(0) = \sqrt{\frac{d}{2\pi}}$ and $g_d'(0) = 0$. Hence,

$$\frac{1}{2\pi}|f'(0)| = \frac{1}{d}\,|f'(0)|\,g_d^2(0) \overset{(a)}{\leq} 2\max_t |f(t)|\,g_d^2(t) \overset{(b)}{\leq} 4d\int_{-1}^{1}|q(t)|\,g_d^2(t)\,dt,$$

where (b) follows from Bernstein's inequality since $|q(t)|\sqrt{1-t^2} \leq d\int_{-1}^{1}|q(z)|dz$ for a degree-d polynomial q and any $t \in [0,1]$, and applying the inequality to $q(t)g_d^2(t)$ of degree $2d-2$. For (a) we apply Bernstein's inequality to $q(z) = \frac{d}{dz}\big(f(z)g_d^2(z)\big)$, set $t = 0$ and notice that $g_d'(0) = 0$.
Equivalently by a change of variables, for $c \overset{\text{def}}{=} 8\pi$ and $f_{\text{T}}(\theta) \overset{\text{def}}{=} f(\sin\theta)$,

$$|f_{\text{T}}'(0)| = |f'(0)| \leq c\,d\int_{-1}^{1}|q(t)|\,g_d^2(t)\,dt \leq c\,d\int_{-\pi}^{\pi}|f_{\text{T}}(\theta)|\,g_d^2(\sin\theta)d\theta.$$

Using the same reasoning for $f_{\text{T}}(\theta+\alpha)$ instead of $f_{\text{T}}(\theta)$ we obtain

$$|f_{\text{T}}'(\alpha)| \leq c\,d\int_{-\pi}^{\pi}|f_{\text{T}}(\theta+\alpha)|\,g_d^2(\sin\theta)\,d\theta = c\,d\int_{-\pi}^{\pi}|f_{\text{T}}(\theta)|\,g_d^2(\sin(\theta-\alpha))\,d\theta,$$

where the equality follows since $|f_{\text{T}}(\theta+\alpha)|\,g_d^2(\sin(\theta))$ is a periodic function with period 2π. Integrating both sides for α from $-\pi$ to π yields

$$\int_{-1}^{1}|f'(t)|\,dt = \int_{-\pi}^{\pi}|f_{\text{T}}'(\alpha)|\,d\alpha \overset{(a)}{\leq} c\,d\int_{-\pi}^{\pi}|f_{\text{T}}(\theta)|\,d\theta = c\,d\int_{-1}^{1}|q(t)|\,dt,$$

where (a) follows since $||g_d||_2 \leq 1$ from Claim 1 and the definition of g_d. Finally, for a proper $\tau \in (0,1)$ of order $1/d^2$,

$$\begin{aligned}
\int_{-1}^{1} \frac{|t\, q(t)|}{\sqrt{1-t^2}}\, dt &\leq \int_{-1+\tau}^{1-\tau} \frac{|q(t)|}{\sqrt{1-t^2}}\, dt + 2\max_t |q(t)| \int_{1-\tau}^{1} \frac{t}{\sqrt{1-t^2}}\, dt \\
&\overset{(a)}{\leq} \frac{\int_{-1}^{1} |q(t)|\, dt}{\sqrt{2\tau-\tau^2}} + 2\max_t |q(t)| \int_{1-\tau}^{1} \frac{t}{\sqrt{1-t^2}}\, dt \\
&= \frac{\int_{-1}^{1} |q(t)|\, dt}{\sqrt{2\tau-\tau^2}} + 2\max_t |q(t)| \cdot \sqrt{2\tau-\tau^2} \\
&\overset{(b)}{\leq} \frac{\int_{-1}^{1} |q(t)|\, dt}{\sqrt{2\tau-\tau^2}} + 2\sqrt{2\tau-\tau^2} \cdot (d+1)^2 \int_{-1}^{1} |q(t)|\, dt \\
&\overset{(c)}{=} 2\sqrt{2}(d+1) \int_{-1}^{1} |q(t)|\, dt,
\end{aligned}$$

where (a) follows since $1/\sqrt{1-t^2} \leq 1/\sqrt{2\tau-\tau^2}$ for $t \in [-1+\tau, 1-\tau]$, (b) follows from the Markov Brothers' inequality, and (c) holds for some $\tau = \mathcal{O}(1/d^2)$. Proof is complete from the fact that $8\pi + 2\sqrt{2} < 28$.

B.2. Proof of Lemma 8

Since for any $i \in \{1, \ldots, m\}$, $\bar{I}_i^+ \in \bar{I}^+$ and $\bar{I}_i^- \in \bar{I}^-$ both have $\lceil \ell(d+1)/2^{i/4} \rceil$ intervals for $i \in \{1, \ldots, m\}$, it follows that the total number of intervals in $(\bar{I}^-, \bar{I}^+)$ is upper bounded as

$$\begin{aligned}
2\left(\sum_{i=1}^{m} |\bar{I}_i| + 1\right) &= 2\left(\sum_{i=1}^{m} \left\lceil \frac{\ell(d+1)}{2^{i/4}} \right\rceil + 1\right) \\
&\leq 2\left(\sum_{i=1}^{m} \left(\frac{\ell(d+1)}{2^{i/4}} + 1\right) + 1\right) \\
&\leq 2\left(\sum_{i=1}^{\infty} \frac{\ell(d+1)}{2^{i/4}} + m + 1\right) \\
&= 2\left(\frac{\ell(d+1)}{2^{1/4}} \cdot \frac{1}{1-2^{-1/4}} + \log_2(\ell(d+1)^2) + 1\right) \\
&\overset{(a)}{\leq} 2\left(\frac{\ell(d+1)}{2^{1/4}} \cdot \frac{1}{1-2^{-1/4}} + 2\log_2(\ell(d+1)) + 1\right) \\
&\overset{(b)}{\leq} 2\left(\frac{\ell(d+1)}{2^{1/4}} \cdot \frac{1}{1-2^{-1/4}} + \frac{2\ell(d+1)}{\log 2}\right) \\
&= 2\ell(d+1)\left(\frac{1}{2^{1/4}-1} + \frac{2}{\log 2}\right) \\
&\leq \frac{4\ell(d+1)}{2^{1/4}-1}.
\end{aligned}$$

where (a) follows since $\ell \geq 1$ and (b) follows from the identity that for any $x \geq 1, \log(x) \leq x - 1$.

B.3. Proof of Lemma 9

Proof We provide a proof for $I = [-1,1]$ by considering $\bar{I} = \overline{[-1,1]}^{d,k}$. An identical proof follows for any other interval I as its partition $\bar{I}^{d,k}$ is obtained by a linear translation of $\overline{[-1,1]}^{d,k}$. For $i \in \{1, \ldots, m\}$, let

$$I_i \overset{\text{def}}{=} I_i^+ \cup I_i^-.$$

Similarly let $E_m \stackrel{\text{def}}{=} E_m^+ \cup E_m^-$. Applying Lemma 7 with $a = 1 - 1/2^i$, we obtain

$$\begin{aligned}
\int_{I_i} |p'(x)|dx \stackrel{(a)}{\leq} \int_{-(1-1/2^i)}^{(1-1/2^i)} |p'(x)|dx &\leq \frac{28(d+1)\|p\|_{1,I}}{(1-(1-1/2^i)^2)^{1/2}} \\
&\leq \frac{28(d+1)\|p\|_{1,I}}{(1-(1-1/2^i))^{1/2}} \\
&\leq 28 \cdot 2^{i/2} \cdot (d+1)\|p\|_{1,I},
\end{aligned}$$

where (a) follows since $I_i \subseteq [-(1-1/2^i), (1-1/2^i)]$. As $\bar{I}_i^+$ consists of $\lceil \ell(d+1)/2^{i/4} \rceil$ equal width intervals from Equation (4) and since $\bar{I}_i^+$ is of width $|\bar{I}_i^+| = 1/2^i$, it follows that each interval in $\bar{I}_i^+$ (and similarly for $\bar{I}^-$) is of width $\leq 1/(2^{3i/4} \cdot l(d+1))$. Thus from Equation (2), the ℓ_1 difference between p and $\bar{p}_{\bar{I}}$ over I_i is given by

$$\begin{aligned}
\|p - \bar{p}_{\bar{I}}\|_{1,I_i} &\leq \sum_{J \in \bar{I}_i} \Delta_J(p) \cdot |J| \\
&\leq \sum_{J \in \bar{I}_i} \Delta_{I_i}(p) \cdot \frac{1}{2^{3i/4} \cdot \ell(d+1)} \\
&\leq \left(\sum_{J \in \bar{I}_i} \int_J |p'(x)|dx \right) \cdot \frac{1}{2^{3i/4} \cdot \ell(d+1)} \\
&\leq \int_{I_i} |p'(x)|dx \cdot \frac{1}{2^{3i/4} \cdot \ell(d+1)} \\
&\leq 28 \cdot 2^{i/2} \cdot (d+1)\|p\|_{1,I} \cdot \frac{1}{2^{3i/4} \cdot \ell(d+1)} = \frac{28 \cdot 2^{-i/4}\|p\|_{1,I}}{\ell}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\|p - \bar{p}_{\bar{I}}\|_{1,I} &= \sum_{i=1}^{m} \|p - \bar{p}_{\bar{I}}\|_{1,I_i} + \|p - \bar{p}_{\bar{I}}\|_{1,E_m} \\
&\leq \sum_{i=1}^{m} \frac{28 \cdot 2^{-i/4}\|p\|_{1,I}}{\ell} + \max_{x \in E_m} p(x) \cdot \frac{2}{\ell(d+1)^2} \\
&\stackrel{(a)}{\leq} \sum_{i=1}^{m} \frac{28 \cdot 2^{-i/4}\|p\|_{1,I}}{\ell} + (d+1)^2\|p\|_{1,I} \frac{2}{\ell(d+1)^2} \\
&\stackrel{(b)}{\leq} \frac{\|p\|_{1,I}}{\ell} \left(\frac{28}{1-2^{-1/4}} + 2 \right) \\
&\stackrel{(c)}{\leq} \frac{4(d+1)\|p\|_{1,I}}{k(2^{1/4}-1)} \left(\frac{28}{1-2^{-1/4}} + 2 \right) \\
&\leq \frac{3764(d+1)\|p\|_{1,I}}{k},
\end{aligned}$$

where (a) follows since I is a symmetric interval, $\forall x \in I$, $p(x) \leq (d+1)^2\|p\|_{1,I}$ from the Markov Brothers' inequality, (b) follows from the infinite negative geometric sum and (c) follows since $\ell \stackrel{\text{def}}{=} k(2^{1/4}-1)/(4(d+1))$ as defined in Equation (4).

B.4. Proof of Lemma 11

Proof Select a $p^* \in \mathcal{P}_d$ that achieves $\|f - p^*\|_{1,I} = \|f - \mathcal{P}_d\|_{1,I}$. Then

$$\begin{aligned}
\|f^{adj} - f\|_{1,I} &\le \|f^{adj} - p^*\|_{1,I} + \|p^* - f\|_{1,I} \\
&\overset{(a)}{\le} 2 \cdot \|f - p^*\|_{1,I} + \frac{c_1(d+1)\|p^* - f^{poly}\|_{1,I}}{k} + \|f^{emp} - f\|_{\mathcal{A}_k,I} \\
&\le 2 \cdot \|f - p^*\|_{1,I} + \frac{c_1(d+1)\big(\|p^* - f\|_{1,I} + \|f - f^{poly}\|_{1,I}\big)}{k} + \|f^{emp} - f\|_{\mathcal{A}_k,I} \\
&\overset{(b)}{\le} \left(2 + \frac{c_1(d+1)(1+c')}{k}\right)\|f - p^*\|_{1,I} + \left(\frac{c_1(d+1)c''}{k} + 1\right)\|f^{emp} - f\|_{\mathcal{A}_k,I} + \frac{c_1(d+1)}{k} \cdot \eta,
\end{aligned}$$

where (a) follows from setting $p = p^*$ in Lemma 10, and (b) follows from using $\|f^{poly} - f\|_{1,I} \le c'\|f - p^*\|_{1,I} + c''\|f^{emp} - f\|_{\mathcal{A}_{d+1},I} + \eta$ along with the fact that $k \ge d+1$ (so that $\|f^{emp} - f\|_{\mathcal{A}_{d+1},I} \le \|f^{emp} - f\|_{\mathcal{A}_k,I}$).

B.5. Proof of Theorem 12

Proof Since $c = 3$ and $c' = 2$ for the f^{adls} estimate, it follows for $I = [X_{(0)}, X_{(n)}]$ from Lemma 11 that

$$\begin{aligned}
\|f^{adj} - f\|_{1,I} &\le \left(2 + \frac{3c_1(d+1)}{k}\right)\|f - p^*\|_1 + \left(\frac{2c_1(d+1)}{k} + 1\right)\|f^{emp} - f\|_{\mathcal{A}_{d+1}} + \frac{c_1(d+1)}{k} \cdot \eta_d \\
&\le (2+\gamma)\|f - \mathcal{P}_d\|_1 + \left(\frac{\gamma}{4} + 1\right)\|f^{emp} - f\|_{\mathcal{A}_{d+1}} + \frac{\gamma}{8} \cdot \eta_d
\end{aligned}$$

where the last inequality follows since we choose $k = k(\gamma) \ge 8c_1(d+1)/\gamma$. Let $J = \mathbb{R} \setminus I$. Using Lemma 4,

$$\begin{aligned}
\mathbb{E}\,\|f^{adj} - f\|_1 &\le (2+\gamma)\|f - \mathcal{P}_d\|_{1,I} + \left(\frac{\gamma}{4} + 1\right)\mathbb{E}\,\|f^{emp} - f\|_{\mathcal{A}_{d+1}} + \frac{\gamma}{8} \cdot \eta_d + (2+\gamma)\|f\|_{1,J} \\
&= (2+\gamma)\|f - \mathcal{P}_d\|_{1,I} + \left(\frac{\gamma}{4} + 1\right)\mathbb{E}\,\|f^{emp} - f\|_{\mathcal{A}_{d+1}} + \frac{\gamma}{8} \cdot \eta_d + (2+\gamma)\,\mathbb{E}\,\|f - f^{emp}\|_{\mathcal{A}_2,J} \\
&\overset{(a)}{\le} (2+\gamma)\|f - \mathcal{P}_d\|_{1,I} + \left(\frac{\gamma}{4} + 1\right)\mathcal{O}\left(\sqrt{\frac{k}{n}}\right) + \frac{\gamma}{8} \cdot \eta_d + 3 \cdot \mathcal{O}\left(\sqrt{\frac{2}{n}}\right) \\
&\overset{(b)}{\le} (2+\gamma)\|f - \mathcal{P}_d\|_1 + \mathcal{O}\left(\sqrt{\frac{d+1}{\gamma \cdot n}}\right),
\end{aligned}$$

where (a) follows since $\gamma < 1$ and from Lemma 4, and (b) follows as $\eta_d = \sqrt{(d+1)/n}$, $k = \mathcal{O}(d+1)$ and $0 < \gamma < 1$.

C. Proofs for Section 5

C.1. Proof of Theorem 13

From Lemma 14,

$$\begin{aligned}
\|f^{out}_{t,d,\alpha} - f\|_1 &\le \left(2 + \frac{4c_1(d+1)}{k} + \frac{1 + k/(d+1)}{\beta - 1}\right)\|f - \mathcal{P}_{t,d}\|_1 + \left(3 + 2c_1 + \frac{2}{\beta - 1}\right)\|f^{emp} - f\|_{\mathcal{A}_{2\beta t \cdot k}} \\
&\quad + \left(\frac{c_1(d+1)}{k} + \frac{k}{(\beta-1)(d+1)}\right)\eta_d \\
&\overset{(a)}{\le} \left(2 + \frac{4c_1(d+1)}{k} + \frac{\alpha(d+1)}{4k} \cdot \left(1 + \frac{k}{d+1}\right)\right)\|f - \mathcal{P}_{t,d}\|_1 + \left(3 + 2c_1 + \frac{\alpha(d+1)}{2k}\right)\|f^{emp} - f\|_{\mathcal{A}_{2\beta t \cdot k}} \\
&\quad + \left(\frac{c_1(d+1)}{k} + \frac{\alpha(d+1)}{4k} \cdot \frac{k}{d+1}\right)\eta_d \\
&\overset{(b)}{\le} \left(2 + \frac{\alpha}{2} + \frac{\alpha}{2}\right)\|f - \mathcal{P}_{t,d}\|_1 + \left(3 + 2c_1 + \frac{\alpha}{2}\right)\|f^{emp} - f\|_{\mathcal{A}_{2\beta t \cdot k}} + \left(\frac{\alpha}{8} + \frac{\alpha}{4}\right)\eta_d
\end{aligned}$$

where (a) follows since by definition $\beta - 1 = 4k/(\alpha(d+1))$, (b) follows since $k \stackrel{\text{def}}{=} \lceil 8c_1(d+1)/\alpha \rceil$ and since $0 < \alpha < 1$, $c_1 > 1$ imply $k \geq d+1$. From Lemma 4,

$$\begin{aligned}
\mathbb{E}\,\|f^{out}_{t,d,\alpha} - f\|_1 &\leq (2+\alpha)\|f - \mathcal{P}_{t,d}\|_1 + \left(3 + 2c_1 + \frac{\alpha}{2}\right)\mathbb{E}\,\|f^{emp} - f\|_{\mathcal{A}_{2\beta t\cdot k}} + \frac{3\alpha\eta_d}{8} \\
&\leq (2+\alpha)\|f - \mathcal{P}_{t,d}\|_1 + \left(3 + 2c_1 + \frac{\alpha}{2}\right)\mathcal{O}\left(\sqrt{\frac{2\beta t k}{n}}\right) + \frac{3\alpha\eta_d}{8} \\
&\stackrel{(a)}{\leq} (2+\alpha)\|f - \mathcal{P}_{t,d}\|_1 + \left(3 + 2c_1 + \frac{\alpha}{2}\right)\mathcal{O}\left(\sqrt{\frac{k^2 t}{\alpha(d+1)}}\right) + \frac{3\alpha\eta_d}{8} \\
&\stackrel{(b)}{\leq} (2+\alpha)\|f - \mathcal{P}_{t,d}\|_1 + \left(3 + 2c_1 + \frac{\alpha}{2}\right)\mathcal{O}\left(\sqrt{\frac{c_1^2(d+1)^2 t}{\alpha^3(d+1)n}}\right) + \frac{3\alpha\eta_d}{8} \\
&\stackrel{(c)}{\leq} (2+\alpha)\|f - \mathcal{P}_{t,d}\|_1 + \mathcal{O}\left(\sqrt{\frac{t(d+1)}{\alpha^3 n}}\right),
\end{aligned}$$

where (a), (b) both follow from the definitions of β, k in Equations (7), (6) and (c) follows since $n_d \stackrel{\text{def}}{=} \sqrt{(d+1)/n}$ and $0 < \alpha < 1$.

C.2. Proof of Lemma 14

For simplicity denote $f^{out} \stackrel{\text{def}}{=} f^{out}_{t,d,\alpha}$ and consider a particular $p^* \in \mathcal{P}_{t,d}$ that achieves $\|f - \mathcal{P}_{t,d}\|_1$. Let $\bar{F}$ denote the set of intervals in $\bar{I}_{\text{ADLS}}$ that has p^* as a single piece polynomial in I. Let $\bar{J} \stackrel{\text{def}}{=} \bar{I}_{\text{ADLS}} \setminus \bar{F}$ be the remaining intervals where p^* has more than one polynomial piece. Since $p^* \in \mathcal{P}_{t,d}$ has t polynomial pieces, the number of intervals in $\bar{J}$ is $\leq t$.

Recall that for any subset $S \subseteq \mathbb{R}$, and integrable functions g_1, g_2, and integer $m \geq 1$, $\|g_1 - g_2\|_{1,S}$, $\|g_1 - g_2\|_{\mathcal{A}_m,S}$ denote the ℓ_1 and $\mathcal{A}_m$ distances over S respectively. Equation (14) in (Acharya et al., 2017) shows that over $\bar{F}$,

$$\sum_{I\in\bar{F}} \|f^{adls}_I - f\|_1 \leq 3\|f - p^*\|_{1,\bar{F}} + 2||f^{emp} - f||_{\mathcal{A}_{|\bar{F}|\cdot(d+1)},\bar{F}} + \eta_d. \tag{9}$$

We bound the error in $\bar{F}$ by setting $p = p^*,\ f^{poly} = f^{adls}_I$ in Lemma 10, using Equation (9), and noting that $k \geq d+1$ as $c_1 \geq 1$ from Equation (6):

$$\begin{aligned}
\sum_{I\in\bar{F}} \|f^{out} - f\|_{1,I} &\leq 2\sum_{I\in\bar{F}} \|f - p^*\|_{1,I} + ||f^{emp} - f||_{\mathcal{A}_{k\cdot|\bar{F}|}} + \frac{c_1(d+1)}{k}(\|p^* - f\|_1 + \|f - f^{adls}_I\|_1) \\
&\leq\ 2\|f - p^*\|_{1,\bar{F}} + (1+2c_1)||f^{emp} - f||_{\mathcal{A}_{k\cdot|\bar{F}|}} + \frac{c_1(d+1)}{k}(4\|f - p^*\|_1 + \eta_d).
\end{aligned} \tag{10}$$

From Lemma 49 (Acharya et al., 2017), for all intervals $I \in \bar{J}$, the following Equation (11) holds that they use to derive Equation (12).

$$\|f^{adls}_I - f^{emp}\|_{\mathcal{A}_{d+1},I} \leq \frac{\|f - p^*\|_1 + \|f^{emp} - f\|_{\mathcal{A}_{2\beta t\cdot(d+1)}} + \eta_d}{(\beta-1)t}. \tag{11}$$

$$\sum_{I\in\bar{J}} \|f^{adls}_I - f\|_{1,I} \leq \frac{\|f - p^*\|_1 + \|f^{emp} - f\|_{\mathcal{A}_{2\beta t\cdot(d+1)}}}{\beta - 1} + 2\|f^{emp} - f\|_{\mathcal{A}_{2\beta t\cdot(d+1)}} + 2\|f - p^*\|_{1,\bar{J}} + \frac{\eta_d}{2(\beta-1)}. \tag{12}$$

Recall that we obtain f^{out} by adding a constant to f^{adls}_I along each interval $I \in \overline{I}^{d,k}$ to match its area to f^{emp} in that interval. Since $\overline{I}^{d,k}$ has $\leq k$ intervals (Lemma 8), $\forall I \in \bar{J}$,

$$\begin{aligned}
\|f^{out} - f^{adls}_I\|_{1,I} &\leq \|f^{adls}_I - f^{emp}\|_{\mathcal{A}_k,I} \\
&\leq \frac{k}{d+1}\|f^{adls}_I - f^{emp}\|_{\mathcal{A}_{d+1},I}
\end{aligned} \tag{13}$$

where the last inequality follows from Property 6.

Adding Equations (12) and (13) over intervals in $\bar{J}$ by noting $\bar{J}$ has $\leq t$ intervals and $k \geq d+1$ implies

$$\sum_{I\in\bar{J}} \|f^{out} - f\|_{1,I} \leq \frac{1+k/(d+1)}{\beta-1}\cdot\|f-p^*\|_1 + 2\|f-p^*\|_{1,\bar{J}} + 2\left(1+\frac{1}{\beta-1}\right)\|f^{emp}-f\|_{\mathcal{A}_{2\beta t\cdot k}} + \frac{k\cdot\eta_d}{(d+1)(\beta-1)}. \tag{14}$$

Adding Equations (10) and (14) proves Lemma 14 (since p^* satisfies $\|f-p^*\|_1 = \|f-\mathcal{P}_{t,d}\|_1$).

D. Proofs for Section 6

D.1. Proof of Theorem 15

Proof Applying the probabilistic version of the VC inequality, i.e. Lemma 4, (see (Devroye & Lugosi, 2012)) to Lemma 14 we have with probability $\geq 1-\delta$,

$$\|f^{out}_{t,d,\alpha} - f\|_1 \leq c\cdot\|f-\mathcal{P}_{t,d}\|_1 + \mathcal{O}\left(\sqrt{\frac{t(d+1)+\log 1/\delta}{n}}\right).$$

From the union bound, the above condition holds true for the $\log n$ sized estimate collection $\{f^{out}_{t,d,\alpha} : t\in\{1,2,4,\ldots,n\}\}$ with probability $\geq 1-\delta\cdot\log n$. Apply the method discussed in Section 6.2 with $\gamma = 2+2/\beta$, to obtain $t^{est} = t^{est}_\beta$. Using Lemma 16 that w.p. $\geq 1-\delta\cdot\log n$,

$$\begin{aligned}\|f^{out}_{t^{est},d,\alpha} - f\|_1 &\leq \min_{t\in\{1,2,4,\ldots,n\},d}\left(\left(1+\frac{2}{\gamma-2}\right)\cdot c\cdot\|f-\mathcal{P}_{t,d}\|_1 + (\gamma+1)\chi\sqrt{\frac{t(d+1)}{n}}\right)\\ &\overset{(a)}{\leq} \min_{0\leq t\leq n,d}\left(\left(1+\frac{2}{\gamma-2}\right)\cdot c\cdot\|f-\mathcal{P}_{t,d}\|_1 + \sqrt{2}(\gamma+1)\chi\sqrt{\frac{t(d+1)+\log 1/\delta}{n}}\right)\\ &\overset{(b)}{\leq} \min_{0\leq t\leq n,d}\left((1+\beta)\cdot c\cdot\|f-\mathcal{P}_{t,d}\|_1 + \mathcal{O}\left(\sqrt{\frac{t(d+1)+\log 1/\delta}{\beta^2 n}}\right)\right),\end{aligned}$$

where (a) follows from the fact that for any $1\leq t\leq n$, $\exists t'\in\{1,2,4,\ldots,n\} : t'\in[t,2t]$ (so that $\|f-\mathcal{P}_{t',d}\|_1 \leq \|f-\mathcal{P}_{t,d}\|_1$ and (b) follows since $\gamma = 2+2/\beta$.

D.2. Proof of Lemma 16

Proof For $i\geq i_\gamma$, from the triangle inequality, and as by definition, $d(v_{i_\gamma}, v_i) \leq \gamma c_i$ for all $i\geq i_\gamma$,

$$d(v_{i_\gamma}, v) \leq d(v_i, v) + d(v_{i_\gamma}, v_i) \leq b_i + c_i + \gamma c_i = b_i + (1+\gamma)c_i.$$

For $i < i_\gamma$, if

$$b_{i_\gamma - 1} \geq \frac{\gamma-2}{2} c_{i_\gamma},$$

the proof follows since for any $1\leq j'\leq i_\gamma - 1$,

$$d(v_{i_\gamma}, v) \leq b_{i_\gamma} + c_{i_\gamma} \leq b_{i_\gamma} + \frac{2}{\gamma-2} b_j \overset{(a)}{\leq} \left(1+\frac{2}{\gamma-2}\right) b_{j'},$$

where (a) follows since $j'\leq j < i_\gamma$. On the other hand if

$$b_{i_\gamma - 1} < \frac{\gamma-2}{2} c_{i_\gamma},$$

then $\forall j'' \geq j+1$,

$$d(v_j, v_{j''}) \leq b_j + b_{j''} + c_j + c_{j''} \overset{(a)}{\leq} 2b_j + 2c_{j''} \leq 2\cdot\frac{\gamma-2}{2}c_{i_\gamma} + 2c_{j''} \overset{(b)}{\leq} \gamma c_{j''},$$

where (a) follows since $j''\geq j$, and (b) follows since $j''\geq j = i_\gamma$, contradicting the definition of i_γ.

E. Comparisons against SURF (Hao et al., 2020a)

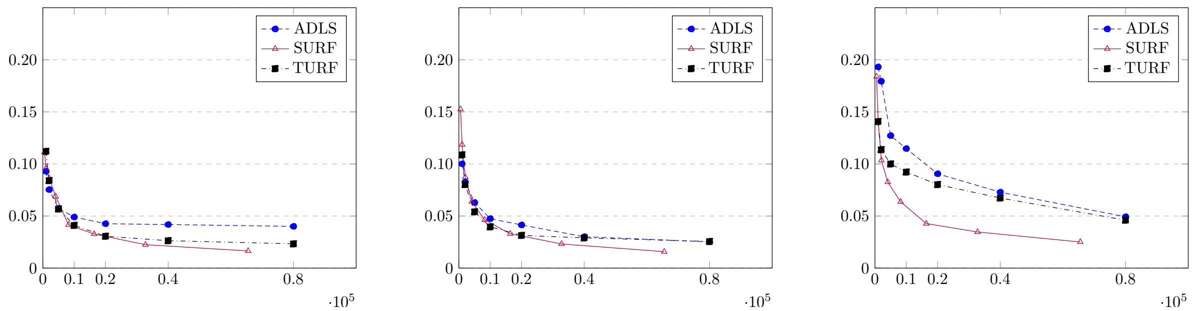

Figure 5. ℓ_1 error versus number of samples on the Beta, Gamma, and Gaussian mixtures respectively in Figure 1 for $d = 1$.

Figure 5 compares SURF against TURF and ADLS for the non-noisy distributions considered in Section 7, namely mixtures of Beta: $.4\mathrm{B}(.8, 4) + .6\mathrm{B}(2, 2)$, Gamma: $.7\Gamma(2, 2) + .3\Gamma(7.5, 1)$, and Gaussians: $.65\mathcal{N}(\text{-}.45,.15^2)+.35\mathcal{N}(.3,.2^2)$. While SURF achieves a lower error, this may be due to its implicit cross-validation method, unlike in ADLS and TURF that relies on our independent cross-validation procedure in Section 6. While the primary focus of our work was in determining the optimal approximation constant, evaluating the experimental performance of the various piecewise polynomial estimators may be an interesting topic for future research.