SURF: A Simple, Universal, Robust, Fast Distribution Learning Algorithm

Yi Hao, Ayush Jain, Alon Orlitsky, Vaishakh Ravindrakumar
Dept. of Electrical and Computer Engineering
University of California, San Diego
{yih179, ayjain, aorlitsky, varavind}@eng.ucsd.edu

Abstract

Sample- and computationally-efficient distribution estimation is a fundamental tenet in statistics and machine learning. We present SURF, an algorithm for approximating distributions by piecewise polynomials. SURF is: simple, replacing prior complex optimization techniques by straight-forward empirical probability approximation of each potential polynomial piece through simple empirical-probability interpolation, and using plain divide-and-conquer to merge the pieces; universal, as well-known polynomial-approximation results imply that it accurately approximates a large class of common distributions; robust to distribution mis-specification as for any degree $d \leq 8$, it estimates any distribution to an ℓ_1 distance < 3 times that of the nearest degree-d piecewise polynomial, improving known factor upper bounds of 3 for single polynomials and 15 for polynomials with arbitrarily many pieces; fast, using optimal sample complexity, running in near sample-linear time, and if given sorted samples it may be parallelized to run in sub-linear time. In experiments, SURF outperforms state-of-the art algorithms.

1 Introduction

1.1 Background

Estimating an unknown distribution from its samples is a fundamental statistical problem arising in many applications such as modeling language, stocks, weather, traffic patterns, and many more. It has therefore been studied for over a century, e.g. [15].

Consider an unknown univariate distribution f over $\mathbb{R}$, generating n samples $X^n \stackrel{\text{def}}{=} X_1, \ldots, X_n$. An *estimator* for f is a mapping $\hat{f} : \mathbb{R}^n \to \mathbb{R}$. As in many of the prior works, we evaluate $\hat{f}$ using the ℓ_1 distance, $\|\hat{f} - f\|_1$. The ℓ_1 distance professes several desirable properties, including scale and location invariance, and provides provable guarantees on the values of Lipschitz functionals of f [6].

Ideally, we would prefer an estimator that learns any distribution. However, arbitrary distributions cannot be learned with any number of samples. Let u be the continuous uniform distribution over $[0, 1]$. For any number n of samples, uniformly select n^3 points from $[0, 1]$ and let p be the discrete uniform distribution over these n^3 points. Since with high probability collisions do not occur within samples under either distribution, u and p cannot be distinguished from the uniformly occurring samples. As $\|u - p\|_1 = 2$, it follows that for any estimator $\hat{f}$, $\max_{f \in \{u,p\}} \mathbb{E}\|\hat{f} - f\|_1 \gtrsim 1$.

A common modification, motivated by PAC agnostic learning, assumes that f is close to a natural distribution class $\mathcal{C}$, and tries to find the distribution in $\mathcal{C}$ closest to f. The following notion of $\mathrm{OPT}_{\mathcal{C}}(f)$ considers this lowest distance, and the usual *minimax learning rate* of $\mathcal{C}$, $\mathcal{R}_n(\mathcal{C})$, is the lowest worst-case expected distance achieved by any estimator,

$$\mathrm{OPT}_{\mathcal{C}}(f) \stackrel{\text{def}}{=} \inf_{g \in \mathcal{C}} \|f - g\|_1, \;\; \mathcal{R}_n(\mathcal{C}) \stackrel{\text{def}}{=} \min_{\hat{f}} \max_{f \in \mathcal{C}} \mathop{\mathbb{E}}_{X^n \sim f} \|\hat{f} - f\|_1.$$

As has been considered in [2], $\hat{f}$ is said to be a factor-c approximation for $\mathcal{C}$ if

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

$$\mathbb{E}\|\hat{f} - f\|_1 \le c \cdot \text{OPT}_{\mathcal{C}}(f) + \epsilon_n$$

where as $n \nearrow \infty$, the *statistical rate*, $\epsilon_n \searrow 0$ at a rate independent of f, namely, the estimator's error is essentially at most c times the optimal. Since for $f \in \mathcal{C}$ has $\text{OPT}_{\mathcal{C}}(f) = 0$, we see that $\epsilon_n \ge \mathcal{R}_n(\mathcal{C})$ for any estimator.

The key challenge is to obtain such an estimate for dense approximation classes $\mathcal{C}$. One such class is the set of degree-d polynomials, $\mathcal{P}_d$ and its t-piecewise extension, $\mathcal{P}_{t,d}$. It is known that by tuning the parameters t, d, the bias and variance under $\mathcal{P}_{t,d}$ can be suitably tailored to achieve several in-class minimax rates. For example, if f is a log-concave distribution, choosing $t = n^{1/5}$ and $d = 1$, $\text{OPT}_{\mathcal{P}_{t,d}}(f) + \mathcal{R}_n(\mathcal{P}_{t,d}) = \mathcal{O}(1/n^{2/5})$ [3], matching the minimax rate of learning log-concave distributions. Similarly, minimax rates may be attained for many other structured classes including uni-modal, Gaussian, and mixtures of all three.

The VC dimension, $\text{VC}(\mathcal{C})$, measures the complexity of a class $\mathcal{C}$. For many dense classes, including $\mathcal{P}_{t,d}$, $\mathcal{R}_n(\mathcal{C}) = \Theta(\sqrt{\text{VC}(\mathcal{C})/n})$. For such classes, a cross-validation based estimator $\hat{f}$, such as the minimum distance based selection [6], across a sufficiently fine cover of $\mathcal{C}$, achieves a factor-3 approximation to $\mathcal{C}$,

$$\mathbb{E}\|\hat{f} - f\|_1 \le 3\text{OPT}_{\mathcal{C}}(f) + \mathcal{O}(\sqrt{\text{VC}(\mathcal{C})/n}).$$

However, in general, such methods might have time complexity exponential in n. This is especially significant in modern applications that process a large number of samples. [1] provided a near-linear $\mathcal{O}(n \log^3 n)$ time algorithm, ADLS, that still achieves the same factor-3 approximation for $\mathcal{P}_{t,d}$ and the statistical rate $\epsilon_n = \mathcal{O}(\sqrt{t(d+1)/n})$. However it leaves some important questions unanswered.

- **Q1:** ADLS shares the same factor-3 approximation as the generic minimum distance selection. However, for the constant-polynomial class $\mathcal{P}_0$, it is easy to see that the empirical histogram $\hat{f}$ achieves a factor-2 approximation, matching a known lower bound [6]. This raises the question if the factor-3 upper bound can be reduced for higher-degree polynomials as well, and if it can be achieved with statistical rate near the optimal $\sqrt{d(t+1)/n}$.
- **Q2:** ADLS requires prior knowledge of the number t of polynomial pieces, which may be impractical in real applications. Even for structured distribution families, the t achieving their minimax rate can vary significantly. For example, for log-concave distributions, $t = \Theta(n^{1/5})$, and for unimodal distributions, $t = \Theta(n^{1/3})$. This raises the question of whether there are estimators that are optimal for $\mathcal{P}_{t,d}$ simultaneously over all $\forall t \ge 0$.

A partial answer for Q1 was provided in [2] who recently showed that any *finite* class $\mathcal{C}$ can be approximated with the optimal approximation factor of 2, and with statistical rate $\epsilon_n = \tilde{\mathcal{O}}\left(|\mathcal{C}|^{1/5}/n^{2/5}\right)$. While this result can be adapted to infinite classes like $\mathcal{P}_{t,d}$ by constructing finite covers, as Lemma 15 in Appendix E shows, even for the basic single piece quadratic polynomial class $\mathcal{P}_2$, this yields $\epsilon_n = \tilde{\mathcal{O}}(n^{-1/4}) \gg \Theta(n^{-1/2}) = \mathcal{R}_n(\mathcal{P}_2)$. And as with the minimum distance selection discussed above, the result is only information-theoretic without a matching algorithm.

Q2 can be partially addressed by using cross-validation techniques, for example based on the minimum distance selection that compare results for different t's and finds the best. However, as shown in [6], this would add an extra approximation factor of at least 3, and perhaps even 5 as ADLS's estimates are un-normalized, resulting in $c = 5 \cdot 3 = 15$. Furthermore this step raises the statistical rate by an additive $\mathcal{O}(\log n/\sqrt{n})$.

SURF answers both questions in the affirmative. Theorem 1 achieves factors ≤ 3 for all degrees ≤ 8 with optimal $\epsilon_n = \mathcal{O}(\mathcal{R}_n(\mathcal{P}_d))$. Corollary 3 achieves the same factors and a near-optimal $\epsilon_n = \tilde{\mathcal{O}}(\mathcal{R}_n(\mathcal{P}_{t,d}))$ for any $t \ge 0$, even unknown, and runs in time $\mathcal{O}(n \log^2 n)$.

The rest of the paper is organized as follows. In Section 2 we describe the construction of intervals and partitions based on statistically equivalent blocks. In Section 3 we present INT, a polynomial approximation method for any queried interval based on a novel empirical mass interpolation. In Section 4 we explain the MERGE and COMP routines, that respectively combine and compare between piecewise polynomial approximations. We conclude in Section 5 with a detailed comparison of SURF and ADLS, and show experimental results that confirm the theory and show that SURF performs well for a variety of distributions. Proofs of all theorems and lemmas may be found in the supplementary material.

1.2 Relation to Prior Work

In terms of objectives, SURF is most closely related to ADLS. Briefly, SURF is simpler, because of which it has a $\mathcal{O}(n \log^2 n)$ time complexity compared to $\mathcal{O}(n \log^3 n)$, it is parallizable to run in sub-linear time given sorted samples unlike ADLS that uses VC dimension based approaches. As mentioned above, it is also more adaptive. On the other hand, when t is known in advance, ADLS achieves a factor-3 approximation with optimal ϵ_n. For a more detailed comparison, see Section 5.

Among the many other methods that have been employed in distribution estimation, see [16, 5], SURF is inspired by the concept of statistically equivalent blocks introduced in [19, 20]. Distribution estimation methods using this concept partition the domain into regions identified by a fixed number of samples, and perform local estimation on these regions. These methods have the advantage that they are simple to describe, almost always of polynomial time complexity in n, and easy to interpret.

The first estimator that used this technique is found in [13]. Expanding on several subsequent works, the notable work [12] shows consistency of a family of equivalent block based estimators for multivariate distributions. See [5] for a more extensive treatment of this subject. Ours is the first work that provides agnostic error guarantees for an equivalent block based estimator.

Other popular estimation methods are the Kernel, nearest neighbor, MLE, and wavelets, see [17]. Another related method uses splines, for example [21, 9]. While MLE and splines may be used for polynomial estimation, MLE is intractable in general, and neither provide agnostic error guarantees.

1.3 Main Results

SURF first uses an interpolation routine INT that outputs an estimate, $\hat{f}_{I,\mathrm{INT}} \in \mathcal{P}_d$ for any queried interval I. Notice that a degree-d polynomial is determined by the measure it assigns to any $d+1$ distinct sub-intervals of I. While ADLS considers fitting the polynomial that minimizes difference in measure to the empirical mass on the worst set of $d+1$ sub-intervals, we show that for low-degree polynomials, it suffices to consider certain special sub-intervals. Provided in Lemma 8, they are functions of d and are sample independent. For $d \le 8$, the resulting estimate is a factor < 3 approximation to $\mathcal{P}_d$, with $\epsilon_n = \mathcal{O}(\mathcal{R}_n(\mathcal{P}_d))$, the optimal statistical rate for $\mathcal{P}_d$.

Theorem 1. *Given samples $X^{n-1} \sim f$ for some $n \ge 128$, degree d, and an interval I with n_I samples within I,* INT *takes $\mathcal{O}(d^\tau + n_I)$ time, and outputs $\hat{f}_{I,\mathrm{INT}} \in \mathcal{P}_d$ such that*

$$\mathbb{E}\|\hat{f}_{I,\mathrm{INT}} - f\|_I \le (r_d + 1) \cdot \inf_{h \in \mathcal{P}_d} \|h - f\|_I + r_d \cdot \sqrt{\frac{2(d+1)q_I}{\pi n}},$$

where $q_I \stackrel{\text{def}}{=} (n_I + 1)/n$, $\|.\|_I$ is the ℓ_1 norm evaluated on I, $\tau < 2.4$ is the matrix inversion exponent, r_d is a fundamental constant whose values are $r_0 = 1, r_1 = 1.25, r_2 \approx 1.42, r_3 \approx 1.55$, $r_4 \le 1.675$, $r_5 \le 1.774$, $r_6 \le 1.857$, $r_7 \le 1.930$, $r_8 \le 1.999$ for $4 \le d \le 8$.

A few remarks are in order. The additive $\mathcal{O}(\sqrt{q_I/n})$ here is related to the standard deviation in the measure associated with an interval that has q_I fraction of samples. For $d > 8$, $r_d > 3$ and they may be evaluated using Lemma 8.

The main routine of SURF, MERGE, then calls INT to obtain a piecewise estimate for any partition of the domain. MERGE uses COMP to compare between the different piecewise estimates. By imposing a special binary structure on the space of partitions, we allow for COMP to efficiently make this comparison via a divide-and-conquer approach. This allows MERGE, and in turn SURF, to output $\hat{f}_{\mathrm{SURF}}$ in $\mathcal{O}((d^\tau + \log n) n \log n)$ time, where τ is the matrix inversion exponent. $\hat{f}_{\mathrm{SURF}}$ is a factor-$(r_d + 1)$ approximation for $\mathcal{P}_{t,d}$ $\forall t \ge 0$. The simplicity of SURF, both the polynomial interpolation and divide-and-conquer, allow us to derive all constants explicitly unlike in the previous works. This result is summarized below in Theorem 2 and Corollary 3.

Theorem 2. *Given $X^{n-1} \sim f$ for some $n \ge 128$ such that n is a power of 2, and parameters $d \le 8$, $\alpha > 2$,* SURF *takes $\mathcal{O}((d^\tau + \log n) n \log n)$ time, and outputs $\hat{f}_{SURF}$ such that w.p. $\ge 1 - \delta$,*

$$\|\hat{f}_{SURF} - f\|_1 \le \min_{\bar{I} \in \Delta_{\mathbb{R}}(X^{n-1})} \sum_{I \in \bar{I}} \left(\frac{(r_d+1)\alpha}{\alpha - 2} \inf_{h \in \mathcal{P}_d} \|h - f\|_I + \frac{r_d(\alpha\sqrt{2} + \sqrt{2} - 1)}{(\sqrt{2}-1)^2} \sqrt{\frac{5(d+1)q_I \log \frac{n}{\delta}}{n}} \right),$$

where q_I is the fraction of samples in interval I, $\Delta_{\mathbb{R}}(X^n)$ is the collection of all partitions of $\mathbb{R}$ whose intervals start and end at a sample point, $\|\cdot\|_I$ is the ℓ_1 distance evaluated in interval I, $\tau < 2.4$ is the matrix inversion exponent, and $r_d > 0$ is the constant in Theorem 1.

Corollary 3. *Running* SURF *with $d \le 8$, $\alpha > 4$,*

$$\mathbb{E}\|\hat{f}_{\text{SURF}} - f\|_1 \le \min_{t\ge 0}\left((r_d+1)\left(1+\frac{4}{\alpha}\right)\cdot \text{OPT}_{\mathcal{P}_{t,d}}(f) + \tilde{\mathcal{O}}\left(\alpha\sqrt{\frac{t\cdot(d+1)}{n}}\right)\right).$$

2 Intervals and Partitions

For $n \ge 1$, let $X^{(n-1)} \stackrel{\text{def}}{=} X_{(1)},\ldots,X_{(n-1)}$ be the increasingly-sorted values of X^{n-1}. For integers $0 \le a < b \le n$, these samples define intervals on the real line $\mathbb{R}$,

$$I_{a,b} = (-\infty, X_{(b)}) \text{ if } a=0,\ I_{a,b} = [X_{(a)}, X_{(b)}) \text{ if } 0<a<b<n,\ I_{a,b} = [X_{(a)},\infty) \text{ if } b=n.$$

The *interval-* and *empirical-probabilities* are $P_{a,b} \stackrel{\text{def}}{=} \int_{I_{a,b}} dF$, and $q_{a,b} \stackrel{\text{def}}{=} \frac{b-a}{n}$. For any $0 \le a < b \le n$, $I_{a,b}$ forms a *statistically equivalent* block [19], wherein $P_{a,b} \sim \text{Beta}(b-a, n-(b-a))$ regardless of f, and $P_{a,b}$ concentrates to $q_{a,b}$.

Lemma 4. *For any $0 \le a < b \le n$, $\epsilon \ge 0$,*

$$\Pr[|P_{a,b} - q_{a,b}| \ge \epsilon\sqrt{q_{a,b}}] \le e^{-(n-1)\epsilon^2/2} + e^{-(n-1)\epsilon^2 q_{a,b}/(2q_{a,b}+2\epsilon\sqrt{q_{a,b}})}.$$

We extend this concentration from one interval to many. For a fixed $\epsilon > 0$, let $\mathcal{Q}_\epsilon$ be the event that

$$\forall\, 0 \le a < b \le n,\ |P_{a,b} - q_{a,b}| \le \epsilon\sqrt{q_{a,b}}.$$

Lemma 5. *For any $n \ge 128$ and $\epsilon \ge 0$,*

$$P[\mathcal{Q}_\epsilon] \ge 1 - n(n+1)/2\cdot\left(e^{-(n-1)\epsilon^2/2} + e^{-(n-1)\epsilon^2/(2+2\epsilon\sqrt{n})}\right).$$

Notice that $\mathcal{Q}_\epsilon$ refers to a stronger concentration event that involves $\sqrt{q_{a,b}}$ $\forall 0 \le a < b \le n$ and standard VC dimension based bounds cannot be readily applied to obtain Lemma 5.

A collection of countably many disjoint intervals whose union is $\mathbb{R}$ is said to be a partition of $\mathbb{R}$. A distribution $\bar{q}$, consisting of interval empirical probabilities is called an *empirical distribution*, or that each probability in $\bar{q}$ is a multiple of $1/n$. The set of all empirical distributions is denoted by $\Delta_{\text{emp,n}}$. Since each $q \in \bar{q} \in \Delta_{\text{emp,n}} \ge 1/n$, $\bar{q}$ may be split into its finitely many probabilities as $\bar{q} = (q_1,\ldots,q_k)$. These probabilities define a partition if we consider the first increasingly sorted $q_1 n$ samples, the next $q_2 n$ samples and so on. For $1 \le i \le k$, let $r_i \stackrel{\text{def}}{=} \sum_{j=1}^{i-1} q_j$ (note that $r_1 = 0$). The empirical distribution defines the following *interval partition*:

$$\bar{I}_{\bar{q}} \stackrel{\text{def}}{=} (I_{r_1 n,(r_1+q_1)n}, I_{r_2 n,(r_2+q_2)n},\ldots,I_{r_k n,(r_k+q_k)n}).$$

3 The Interpolation Routine

This section describes INT, which outputs an estimate $\hat{f}_{I,\text{INT}} \in \mathcal{P}_d$ for any queried interval I. WLOG let $I = [0,1]$. A collection, $\bar{n}_d = (n_0,\ldots,n_{d+1})$ such that $0 = n_0 \le n_1 \le \cdots \le n_d \le n_{d+1} = 1$ is said to be a node partition of $[0,1]$. Let $\mathcal{N}_d$ be the set of node partitions and for the set of non-zero polynomials, $\mathcal{P}_d \setminus \{0\}$, define $r : \mathcal{N}_d, \mathcal{P}_d \to [1,\infty)$ and its suprema

$$r(\bar{n}_d, h) \stackrel{\text{def}}{=} \frac{\int_0^1 |h|}{\sum_{i=1}^{d+1} |\int_{n_{i-1}}^{n_i} h|},\ r_d(\bar{n}_d) = \sup_{h\in\mathcal{P}_d\setminus\{0\}} r(\bar{n}_d, h). \tag{1}$$

Notice that $r(\bar{n}_d, h) \ge 1$ since the absolute integral $\ge$ the sum of absolute areas. For any node partition $\bar{n}_d \in \mathcal{N}_d$, let $J_{\bar{n}_d,i} \stackrel{\text{def}}{=} [n_{i-1}, n_i]$, $i \in \{1,\cdots,d+1\}$ so that $\bar{J}_{\bar{n}_d} = (J_{\bar{n}_d,1},\ldots,J_{\bar{n}_d,d+1})$ partitions $[0,1]$. Let $\hat{f}_{\bar{n}_d} \in \mathcal{P}_d$ be the unique polynomial whose measure on all $d+1$ intervals in $\bar{J}_{\bar{n}_d}$ matches its empirical mass. It is defined as:

$$\hat{f}_{\bar{n}_d} \stackrel{\text{def}}{=} h \in \mathcal{P}_d : \forall i \in \{1,\cdots,d+1\}, \int_{n_{i-1}}^{n_i} h(z)dz = q_{J_{\bar{n}_d,i}}, \tag{2}$$

where for n_J samples that lie within an interval J, $q_J \stackrel{\text{def}}{=} (n_J+1)/n$. Computation of $\hat{f}_{\bar{n}_d}$ involves a calculation of $d+1$ empirical masses that takes $\mathcal{O}(n_J)$ time, and solving a system of $d+1$ linear equations that takes $\mathcal{O}(d^\tau)$ time, where $\tau < 2.4$ is the matrix inversion exponent, for a $\mathcal{O}(n_J + d^\tau)$ run time. The estimate $\hat{f}_{\bar{n}_d}$ corresponding to any choice of $\bar{n}_d \in \mathcal{N}_d$ satisfies the following:

Lemma 6. *For interval $I = [0,1]$ with empirical probability q_I, any $\bar{n}_d \in \mathcal{N}_d$, and $\epsilon > 0$, the estimate $\hat{f}_{\bar{n}_d}$ (2) is such that under event $\mathcal{Q}_\epsilon$,*

$$\|\hat{f}_{\bar{n}_d} - f\|_1 \le (1 + r_d(\bar{n}_d)) \inf_{h\in\mathcal{P}_d} \|h - f\|_1 + r_d(\bar{n}_d)\epsilon\sqrt{(d+1)q_I}.$$

In Lemma 7, we show that for any $\bar{n}_d \in \mathcal{N}_d$, there exists an $r_d(\bar{n}_d)$ achieving $h \in \mathcal{P}_d$, and that it belongs to a special set, $\mathcal{P}_{\bar{n}_d} \subseteq \mathcal{P}_d$,

$$\mathcal{P}_{\bar{n}_d} \stackrel{\text{def}}{=} \Big\{ h \in \mathcal{P}_d : \exists i_1 \in \{1,\ldots,d+1\} : \forall i \in \{1,\ldots,d+1\} \setminus \{i_1\}, \int_{n_{i-1}}^{n_i} h = 0 \Big\}.$$

In words, $\mathcal{P}_{\bar{n}_d}$ is the set of polynomials that has a non-zero area in at most one $I \in \bar{I}_{\bar{n}_d}$.

Lemma 7. *For any degree-d and $\bar{n}_d \in \mathcal{N}_d$,*

$$r_d(\bar{n}_d) = \sup_{h\in\mathcal{P}_d} r(\bar{n}_d, h) = \max_{h\in\mathcal{P}_{\bar{n}_d}} r(\bar{n}_d, h).$$

Let the smallest $r_d(\bar{n}_d)$ be denoted by $r_d^\star \stackrel{\text{def}}{=} \inf_{\bar{n}_d\in\mathcal{N}_d} r_d(\bar{n}_d)$. Lemma 8 shows that there exists an $\bar{n}_d$ that attains the infimum. It is denoted by $\bar{n}_d^\star = \arg\min_{\bar{n}_d\in\mathcal{N}_d} r_d(\bar{n}_d)$. For $d \le 3$, we calculate $r_d^\star$ and $\bar{n}_d^\star$. For $4 \le d \le 8$ we find a $\bar{n}_d \in \mathcal{N}_d$ such that the corresponding $r_d(\bar{n}_d) < 2$.

Lemma 8. *For $d \le 3$, there exists a node collection $\bar{n}_d^\star$ that achieves $r_d^\star$. These, and their respective $r_d^\star$ are given by*

d	$\bar{n}_d^\star$	$r_d^\star$
0	$(0,1)$	1
1	$(0,0.5,1)$	1.25
2	$\approx (0, 0.2599, 0.7401, 1)$	≈ 1.42
3	$\approx (0, 0.1548, 0.5, 0.8452, 1)$	≈ 1.56

Denoting $\bar{n}_2^\star = (0, \alpha_0, 1-\alpha_0, 1)$, and $\bar{n}_3^\star = (0, \beta_0, 0.5, 1-\beta_0, 1)$, the exact values of α_0, β_0, are obtained as roots to a degree-14 and degree-69 polynomial that we explicitly provide. For degrees $4 \le d \le 8$, the following $\bar{n}_d \in \mathcal{N}_d$ and $r_d(\bar{n}_d)$ provide upper bounds on $r_d^\star$.

d	$\bar{n}_d$	$r_d(\bar{n}_d)$
4	$(0, 0.1015, 0.348, 0.652, 0.8985, 1)$	< 1.675
5	$(0, 0.071, 0.254, 0.5, 0.746, 0.929, 1)$	< 1.774
6	$(0, 0.053, 0.192, 0.390, 0.610, 0.808, 0.947, 1)$	< 1.857
7	$(0, 0.0405, 0.149, 0.310, 0.5, 0.690, 0.851, 0.9595, 1)$	< 1.930
8	$(0, 0.032, 0.119, 0.252, 0.414, 0.586, 0.749, 0.881, 0.968, 1)$	< 1.999

For a given interval I and $d \le 8$, INT first scales and shifts I to obtain $[0,1]$. It then constructs $\hat{f}_{\bar{n}_d}$ using the $\bar{n}_d$ in Lemma 8. The output $\hat{f}_{I,\mathrm{INT}}$ is the re-scaled-shifted $\hat{f}_{\bar{n}_d}$.

4 The Compare and Merge Routines

This section presents MERGE and COMP, the main routines of SURF. For any contiguous collection of intervals $\bar{I}$, let $\hat{f}_{\bar{I},\mathrm{INT}}$ be the piecewise polynomial estimate consisting of $\hat{f}_{I,\mathrm{INT}} \in \mathcal{P}_d$ given by INT in each $I \in \bar{I}$. The key idea in SURF is to separate interval partitions into a binary hierarchy, effectively allowing a comparison of all the superpolynomially many (in n) estimates corresponding to the different interval partitions, but by using only $\tilde{\mathcal{O}}(n)$ comparisons.

Recall that n here a power of 2 and define the integer $D \stackrel{\text{def}}{=} \log_2 n$. An empirical distribution, $\bar{q} \in \Delta_{\mathrm{emp,n}}$, is called a *binary* distribution if each of its probability values take the form $1/2^d$, for some integer $0 \le d \le D$. The corresponding interval partition, $\bar{I}_{\bar{q}}$, is said to be a *binary partition*.

$$\Delta_{\mathrm{bin,n}} \stackrel{\text{def}}{=} \{\bar{q} \in \Delta_{\mathrm{emp,n}} : \forall q \in \bar{q}, q = 1/2^{\nu(q)}, 0 \le \nu(q) \le D,\ \nu(q) \in \mathbb{Z}\}.$$

For example $\bar{q} = (1)$, $\bar{q} = (1/2, 1/4, 1/4)$, $\bar{q} = (1/4, 1/8, 1/8, 1/2)$ are binary distributions. Similarly, $(1/n,\ldots,1/n) = (1/2^{\log n},\ldots,1/2^{\log n})$ is also a binary distribution since n here is a power of 2 (assume $n \ge 8$ so that they are all in $\Delta_{\mathrm{emp,n}}$). Lemma 9 shows that $\Delta_{\mathrm{bin,n}}$ retains most of the approximating power of $\Delta_{\mathrm{emp,n}}$. In particular, that for any $\bar{q} \in \Delta_{\mathrm{emp,n}}$, there exists a binary distribution $\bar{q}' \in \Delta_{\mathrm{bin,n}}$ such that $\bar{I}_{\bar{q}'}$ has a smaller bias than $\bar{I}_{\bar{q}}$, while its deviation under the concentration event, $\mathcal{Q}_\epsilon$, is larger by less than a factor of $1/(\sqrt{2}-1)$.

Lemma 9. *For any empirical distribution $\bar{q} \in \Delta_{\mathrm{emp,n}}$, there exists $\bar{q}' \in \Delta_{\mathrm{bin,n}}$ such that*

$$\|f^\star_{\bar{I}_{\bar{q}'}} - f\|_1 \le \|f^\star_{\bar{I}_{\bar{q}}} - f\|_1, \ \sum_{q \in \bar{q}'} \epsilon\sqrt{q} \le \sum_{q\in\bar{q}} \frac{1}{\sqrt{2}-1}\epsilon\sqrt{q},$$

where for any $d > 0$, $f^\star_{\bar{I}}$ is the piecewise degree-d polynomial closest to f on the partition $\bar{I}$.

For a fixed $\bar{p} \in \Delta_{\mathrm{bin,n}}$, let $\Delta_{\mathrm{bin,n},\le\bar{p}}$ be the set of binary distributions such that for any $\bar{q} \in \Delta_{\mathrm{bin,n},\le\bar{p}}$, each $I_1 \in \bar{I}_{\bar{q}}$ is contained in some $I_2 \in \bar{I}_{\bar{p}}$.

$$\Delta_{\mathrm{bin,n},\le\bar{p}} \stackrel{\mathrm{def}}{=} \{\bar{q} \in \Delta_{\mathrm{bin,n}} : \forall I_1 \in \bar{I}_{\bar{q}},\ \exists I_2 \in \bar{I}_{\bar{p}},\ I_1 \subseteq I_2\}. \tag{3}$$

For example if $\bar{p} = (1/2, 1/4, 1/4)$ is the binary distribution, $(1/4, 1/4, 1/8, 1/8, 1/4)$, $(1/2, 1/4, 1/8, 1/8) \in \Delta_{\mathrm{bin,n},\le\bar{p}}$, whereas $(1/2, 1/2) \notin \Delta_{\mathrm{bin,n},\le\bar{p}}$.

4.1 The MERGE Routine

The MERGE routine operates in $i \in \{1, \ldots, D\}$ steps (recall $D = \log_2 n$) where at the end of each step i, MERGE holds onto a binary distribution q_i. At at the last step $i = D$, SURF outputs the piecewise estimate on the partition given by $\bar{q}_D$, i.e. $\hat{f}_{\mathrm{SURF}} = \hat{f}_{\bar{I}_{\bar{q}_D},\mathrm{INT}}$. Let

$$D(i) \stackrel{\mathrm{def}}{=} D - i \text{ and let } \bar{u}_i \stackrel{\mathrm{def}}{=} \left(1/2^{D(i)}, \ldots, 1/2^{D(i)}\right).$$

Initialize $\bar{q}_0 \leftarrow (1/n, \ldots, 1/n)$. Start with $i = 1$ and assign $\bar{s} \leftarrow \bar{q}_{i-1}$. Throughout its run MERGE maintains $\bar{s} = \bar{q}_{i-1} \in \Delta_{\mathrm{bin,n},\le\bar{u}_i}$. For instance this holds for $i = 1$ since $\bar{u}_1 = (2/n, \ldots, 2/n)$. MERGE considers merging the probability values in $\bar{s}$ to match it with $\bar{u}_i$. For example if at step $i = D - 1$, $\bar{s} = (1/8, 1/8, 1/4, 1/4, 1/4)$, it considers merging $(1/8, 1/8, 1/4)$ and $(1/4, 1/4)$ to obtain $\bar{u}_{D-1} = (1/2, 1/2)$.

This decision is made by invoking the COMP routine on intervals corresponding to the merged probability value. In this case COMP is called on intervals $\bar{I} \in \bar{I}_{\bar{s}}$ corresponding to $(1/8, 1/8, 1/4)$ and $(1/4, 1/4)$ respectively, along with the tuning parameter γ,

$$\gamma \stackrel{\mathrm{def}}{=} \alpha \cdot r_d \cdot \epsilon\sqrt{d+1}.$$

While COMP decides to merge depending on the increment in bias on the merged interval versus the decrease in variance, γ tunes this trade-off. A large γ results in a decision to merge while a small γ has the opposite effect. If $\mathrm{COMP}(\bar{I}, \gamma) \le 0$ the probabilities in $\bar{s}$ corresponding to $\bar{I}$ are merged and copied into $\bar{q}_i$. Otherwise they are copied as is into $\bar{q}_i$. See Appendix F.1 for a detailed description.

At each step $i \in \{1, \ldots, D\}$, MERGE calls COMP on $2^{D(i)}$ intervals, each consisting of 2^i samples. Thus each step of MERGE takes $\mathcal{O}(2^{D(i)} \cdot (d^\tau + \log(2^i)) \cdot 2^i) = \mathcal{O}((d^\tau + \log n)2^D)$ time. The total time complexity is therefore $\mathcal{O}((d^\tau + \log n)2^D D) = \mathcal{O}((d^\tau + \log n)n \log n)$.

4.2 The COMP Routine

COMP receives an interval partition $\bar{I}$ consisting of m samples and the parameter γ as input, and returns a real value that indicates its decision to merge the probabilities under $\bar{I}$. Let $\bar{s}$ be the set of empirical probabilities corresponding to $\bar{I}$. Let the merged interval be I and let $\hat{f} = \hat{f}_{\mathrm{INT},I}$ be the polynomial estimate on I.

For simplicity suppose $\bar{I} = (I_1, I_2)$ with empirical mass s_{I_1}, s_{I_2} respectively, and let $\mathrm{OPT}_{I,\mathcal{P}_{t,d}}(f) = \min_{h\in\mathcal{P}_{t,d}} \|h - f\|_I$. If $\bar{I}$ is merged, observe that the bias $\mathrm{OPT}_{I,\mathcal{P}_{t,d}}(f) \ge \mathrm{OPT}_{I_1,\mathcal{P}_{t,d}}(f) + \mathrm{OPT}_{I_2,\mathcal{P}_{t,d}}(f)$ increases but since $s_I = s_{I_1} + s_{I_2}$, $\sqrt{s_I} \le \sqrt{s_{I_1}} + \sqrt{s_{I_2}}$, resulting in a smaller ϵ-deviation under event $\mathcal{Q}_\epsilon$ in Lemma 5. Consider their difference parameterized by the constant γ,

$$\mu'_\gamma(f) \stackrel{\mathrm{def}}{=} (\mathrm{OPT}_{I_1,\mathcal{P}_{t,d}}(f) + \mathrm{OPT}_{I_2,\mathcal{P}_{t,d}}(f) - \mathrm{OPT}_{I,\mathcal{P}_{t,d}}(f)) - \gamma(\sqrt{s_{I_1}} + \sqrt{s_{I_2}} - \sqrt{s_I}).$$

If $\mu'_\gamma(f) \le 0$, it indicates that the overall ℓ_1 error is smaller under the merged I. While $\mu'_\gamma(f)$ cannot be evaluated without access to the underlying f, we use a proxy, $\mu_{\bar{I},\gamma}(f)$ that is defined next.

Normalize $\bar{s}$ so that it is a distribution, and consider $\bar{p} \in \Delta_{\mathrm{bin,m}}$ such that $\bar{s} \in \Delta_{\mathrm{bin,m},\le\bar{p}}$ and the piecewise estimate on $\bar{I}_{\bar{p}}$, i.e. $\hat{f}_{\bar{I}_{\bar{p}},\mathrm{INT}}$. Define $\Lambda_{\bar{I}_{\bar{p}}}(\hat{f}) \stackrel{\mathrm{def}}{=} \|\hat{f}_{\bar{I}_{\bar{p}},\mathrm{INT}} - \hat{f}\|_{\bar{I}_{\bar{p}}}$, $\lambda_{\bar{p},\gamma} \stackrel{\mathrm{def}}{=} \sum_{p\in\bar{p}} \gamma\sqrt{p}$,

$$\mu_{\bar{I}_{\bar{s}},\gamma}(\hat{f}) \stackrel{\mathrm{def}}{=} \max_{\bar{p}:\bar{s}\in\Delta_{\mathrm{bin,m},\le\bar{p}}} \Lambda_{\bar{I}_{\bar{p}}}(\hat{f}) - \lambda_{\bar{p},\gamma}.$$

COMP returns $\mu_{\bar{I}_{\bar{s}},\gamma}(\hat{f})$ via a divide-and-conquer based implementation, and results in $\mathcal{O}((d^\tau + \log m)m)$ time. A detailed description is provided in Appendix F.3. Lemma 10 shows that under event $\mathcal{Q}_\epsilon$, $\hat{f}_{\mathrm{SURF}}$ is within a constant factor of the best piecewise polynomial approximation over any binary partition, plus its deviation in probability under $\mathcal{Q}_\epsilon$ times $\mathcal{O}(\sqrt{d+1})$.

Lemma 10. *Given samples $X^{n-1} \sim f$, for some n that is a power of 2, degree $d \le 8$ and the threshold $\alpha > 2$,* SURF *outputs $\hat{f}_{SURF}$ in time $\mathcal{O}((d^\tau + \log n)n \log n)$ such that under event $\mathcal{Q}_\epsilon$,*

$$\|\hat{f}_{SURF} - f\|_1 \le \min_{\bar{p} \in \Delta_{\mathrm{bin,n}}(X^{n-1})} \sum_{I \in \bar{I}_{\bar{p}}} \left(\frac{(r_d+1)\alpha}{\alpha-2} \inf_{h \in \mathcal{P}_d} \|h - f\|_I + \frac{r_d(\alpha\sqrt{2} + \sqrt{2} - 1)}{\sqrt{2}-1} \epsilon \sqrt{(d+1)q_I} \right),$$

where q_I is the empirical mass under interval I, r_d is the constant in Theorem 1.

5 Comparison and Experiments

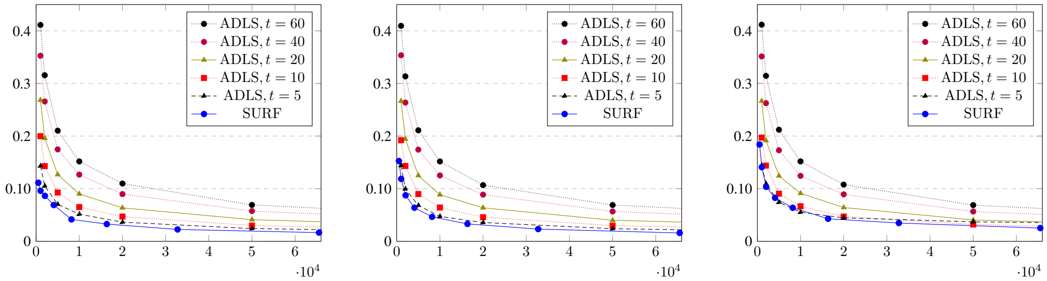
We compare the factor improvement of SURF with ADLS, expand on larger degrees-d polynomial approximation, and in particular, address learning Gaussians optimally. We also describe how SURF benefits from its local nature, enabling a distributed computation. Our experiments show that SURF is more adaptive than ADLS, and perform additional experiments on both synthetic and real datasets.

The following table compares SURF with ADLS in terms of the expected error. For $d \le 8$, $r_d \in [2, 3)$ is the factor in Theorem 1, and $\tau, \omega \in [2, 2.4]$ are constants. We achieve a lesser factor approximation at nearly the optimal statistical rate, with an improved time complexity in both n and d.

	SURF	ADLS
$\mathcal{P}_d$	$r_d\,\mathrm{OPT}_{\mathcal{P}_{t,d}}(f) + \sqrt{\frac{2d}{\pi n}}$	$3\mathrm{OPT}_{\mathcal{P}_{t,d}}(f) + \mathcal{O}\left(\sqrt{\frac{d}{n}}\right)$
$\mathcal{P}_{t,d}$ and known t	$r_d\,\mathrm{OPT}_{\mathcal{P}_{t,d}}(f) + \mathcal{O}\left(\sqrt{\frac{t(d+1)\log n}{n}}\right)$	$3\mathrm{OPT}_{\mathcal{P}_{t,d}}(f) + \mathcal{O}\left(\sqrt{\frac{t(d+1)}{n}}\right)$
$\mathcal{P}_{t,d}$ and unknown t	$r_d \min_{t\ge 0} \Big(\mathrm{OPT}_{\mathcal{P}_{t,d}}(f) + \mathcal{O}\left(\sqrt{\frac{t(d+1)\log n}{n}}\right)\Big)$	$15 \min_{t\ge 0} \Big(\mathrm{OPT}_{\mathcal{P}_{t,d}}(f) + \mathcal{O}\left(\sqrt{\frac{t(d+1)}{n}}\right)\Big) + \mathcal{O}\left(\frac{\log n}{\sqrt{n}}\right)$
Time complexity	$\mathcal{O}(n \log^2 n d^\tau)$	$\mathcal{O}(n \log^3 n d^{3+\omega})$

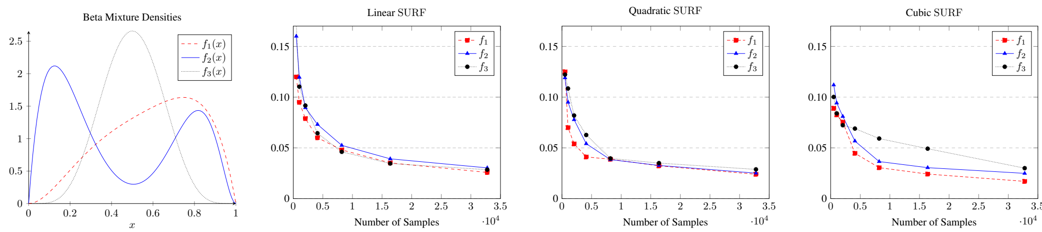
While for $d > 8$, SURF does not improve the approximation factor below < 3, we note that polynomial approximations of larger degrees exhibit oscillatory behavior, for example around the edges when approximating a pulse. Called the Runge phenomenon [18], this may result in an unbounded ℓ_p distance for $p > 1$. In this scenario it may be preferred to use a lower degree polynomial, but with an appropriately large t. Consider the important case when f is a Gaussian distribution. As shown in Lemma 16, $\mathrm{OPT}_{\mathcal{P}_{t,d}}(f) = \mathcal{O}(1/t^{d-1})$. Using the fact that $\epsilon_n = \tilde{\mathcal{O}}(\sqrt{t(d+1)/n})$ and minimizing $\mathrm{OPT}_{\mathcal{P}_{t,d}}(f) + \epsilon_n$ over t for a fixed d, we obtain $\|\hat{f}_{\mathrm{SURF}} - f\|_1 = \tilde{\mathcal{O}}((d+1)/n)^{\frac{1}{2} - \frac{1}{4d-2}}$. Even for an astronomical $n = 2^{100}$ samples, choosing $d = 8$ ensures that $n^{\frac{1}{4d-2}} \le 11$. Thus in almost all scenarios of practical interest we nearly match (upto a $\sqrt{\log n}$ factor) the minimax rate $\mathcal{O}(1/n)^{\frac{1}{2}}$ of learning Gaussians. While ADLS avoids this factor of $n^{\frac{1}{4d-2}}$, they do so by using $d = \mathcal{O}(\log n)$ which may present the above drawbacks. For degrees that are even larger, the $\Omega(d^5 n \log^3 n)$ time taken by ADLS may make it impractical.

In terms of time complexity, SURF benefits from its local nature, enabling a distributed computation. As detailed in Appendix F, if provided with pre-sorted samples, a known t and memory $m \ge t$, it can be adapted to run in time $\mathcal{O}((d^\tau + \log n)n \max\{1/t, \log n/m\}) \ll \mathcal{O}(n)$, if $t \approx n$. We now follow up with an experimental comparison. SURF is run with $\alpha = 0.25$ and the errors are averaged over 10 runs. In running ADLS we use the provided code as is. Figure 1 compares the ℓ_1 error in piecewise-linear estimation using SURF vs ADLS on the distributions considered in [1], namely, a beta, Gamma, and Gaussian mixture. The plots correspond to the errors incurred on running SURF, and ADLS with pieces $t = 5, 10, 20, 40, 60$. While some hyperparameter optimizations may aid



(a) .4Beta(.8, 4)+.6Beta(2, 2)　(b) .7Gam(2, 2)+.3Gam(7.5, 1)　(c) .65$\mathcal{N}$(-.45,.15^2)+.35$\mathcal{N}$(.3,.2^2)

Figure 1: ℓ_1 error versus number of samples of piece-wise linear SURF and ADLS.



(a) Densities f_1, f_2, f_3.　(b) ℓ_1 error with $d = 1$.　(c) ℓ_1 error with $d = 2$.　(d) ℓ_1 error with $d = 3$.

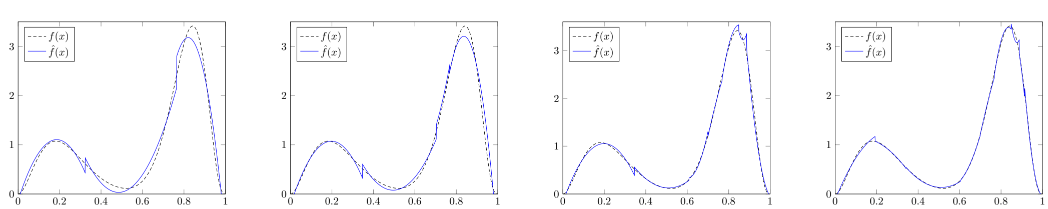
Figure 2: Evaluation of the estimate output by SURF with degrees $d = 1, 2, 3$, $\alpha = 0.25$, on $f_1 = 0.4\text{Beta}(3, 4) + 0.6\text{Beta}(5, 2)$, $f_2 = 0.4\text{Beta}(10, 3) + 0.6\text{Beta}(2, 8)$, and $f_3 = \text{Beta}(6, 6)$.

either algorithms, observe that the errors can be much larger with the wrong t. Significantly, the $t = 5$ for which the results are comparable, is also roughly the number of pieces that SURF outputs.

Experiments show that SURF learns a wide range of parametric families such as the beta, Gaussian and Gamma distributions. In Figure 2 we show results on beta mixture distributions over $[0, 1]$, as they accommodate a wide range of shapes. Other results may be found in Appendix G. Let Beta(α, β) be the beta density with parameters α, β. We run SURF to estimate three distributions, as shown in Figure 2(a). SURF estimates them using piecewise polynomials of degree $d = 1, 2, 3$. Figures 2(b)–2(d) show the resulting ℓ_1 errors. Observe that the errors are decaying, and are similar between distributions. This is not surprising since low degree polynomial approximations largely rely on local smoothness, which all of the considered densities possess. By the same reasoning, on increasing d from 1 to 3, the variation in error between distributions increases. The smoother f_1 starts incurring a smaller ℓ_1 error than f_2 and f_3.

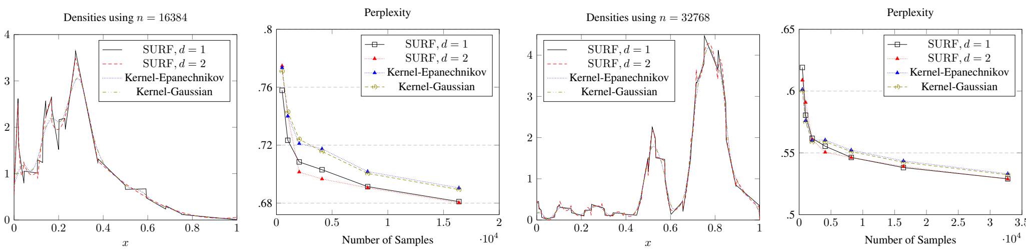
Next, we run SURF with $d = 2$ to estimate $f = 0.3f_{\text{Beta},3,10} + 0.7f_{\text{Beta},17,4}$ with $n = 1024, 4096, 16384, 65536$. Figure 3 plots the resulting estimates against f. Notice that the estimate not only successively better estimates f in ℓ_1 distance, but also pointwise converges to f.

Finally, we ran SURF on real data sets consisting of salaries from the 1994 US census and electric signals from the sensorless drive diagnosis dataset [8], that have been used to evaluate classification



(a) $n = 1024$.　(b) $n = 4096$.　(c) $n = 16384$.　(d) $n = 65536$.

Figure 3: SURF with degree $d = 2$, $\alpha = 0.25$ estimating $f = 0.3f_{\text{Beta},3,10} + 0.7f_{\text{Beta},17,4}$ with $n = 1024, 4096, 16384, 65536$ samples.



(a) Salary Estimate (b) Salary Perplexity (c) Signal Estimate (d) Signal Perplexity

Figure 4: Real data estimates and perplexity of SURF vs MLE based Kernel estimators

algorithms [10, 4, 14]. We trim 0.5% of samples on either side and re-scale to obtain 57923 samples that lie in $[0, 1]$. Figures 4(a) and 4(c) show the estimate output by SURF and the similarly non-parametric, popularly used Kernel estimator with Epanechnikov and Gaussian kernels via the `fitdist()` function in `MATLAB`®. As it can be observed, SURF, without any hidden parameter, recovers characteristic features of the distribution such as the clusters, mode values, and tails. This is in contrast with ADLS, that, strictly speaking, cannot be used in this context as it requires additional cross-validation to tune t based on the number of clusters, etc. The perplexity, or the exponent of the average negative log-likelihood on unseen samples, is a commonly used measure in practice to evaluate an estimate. Figures 4(d), 4(b) compares the perplexity on a test set with one-fourth the number of samples. As it can be seen, even as `fitdist()` outputs the perplexity minimizer on the training set, SURF performs better.

Broader Impact

SURF is a simple, universal, robust, and fast algorithm for the important problem of estimating distributions by piecewise polynomials. Real-life applications are likely to be approximated by relatively low-degree polynomials and require fast algorithms. SURF is particularly well-suited for these regimes.

Acknowledgments

We are grateful to the National Science Foundation (NSF) for supporting this work through grants CIF-1564355 and CIF-1619448.

References

[1] Jayadev Acharya, Ilias Diakonikolas, Jerry Li, and Ludwig Schmidt. Sample-optimal density estimation in nearly-linear time. In *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1278–1289. SIAM, 2017.

[2] Olivier Bousquet, Daniel Kane, and Shay Moran. The optimal approximation factor in density estimation. *arXiv preprint arXiv:1902.05876*, 2019.

[3] Siu-On Chan, Ilias Diakonikolas, Rocco A Servedio, and Xiaorui Sun. Efficient density estimation via piecewise polynomial approximation. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 2014.

[4] Jie Cheng and Russell Greiner. Comparing bayesian network classifiers. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 101–108. Morgan Kaufmann Publishers Inc., 1999.

[5] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

[6] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Science & Business Media, 2001.

[7] *NIST Digital Library of Mathematical Functions, 8.17*. http://dlmf.nist.gov/8.17, Release 1.0.25 of 2019-12-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

[8] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[9] Chong Gu and Chunfu Qiu. Smoothing spline density estimation: Theory. *The Annals of Statistics*, pages 217–234, 1993.

[10] Ron Kohavi. Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pages 202–207, 1996.

[11] Ilia Krasikov. New bounds on the hermite polynomials. *East journal on approximations*, 10(3):355–362, 2004.

[12] Gábor Lugosi, Andrew Nobel, et al. Consistency of data-driven histogram methods for density estimation and classification. *The Annals of Statistics*, 24(2):687–706, 1996.

[13] KR Parthasarathy and PK Bhattacharya. Some limit theorems in regression theory. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 91–102, 1961.

[14] Fabian Paschke, Christian Bayer, Martyna Bator, Uwe Mönks, Alexander Dicks, Olaf Enge-Rosenblatt, and Volker Lohweg. Sensorlose zustandsüberwachung an synchronmotoren. In *23. Workshop Computational Intelligence 2013. Proceedings*, volume 46, 2013.

[15] Karl Pearson. 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.

[16] David W Scott. Multivariate density estimation and visualization. In *Handbook of computational statistics*, pages 549–569. Springer, 2012.

[17] Bernard W Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

[18] Lloyd N Trefethen. *Approximation theory and approximation practice*, volume 128. Siam, 2013.

[19] John W Tukey. Non-parametric estimation ii. statistically equivalent blocks and tolerance regions–the continuous case. *The Annals of Mathematical Statistics*, pages 529–539, 1947.

[20] John W Tukey. Nonparametric estimation, iii. statistically equivalent blocks and multivariate tolerance regions–the discontinuous case. *The Annals of Mathematical Statistics*, pages 30–39, 1948.

[21] Edward J Wegman and Ian W Wright. Splines in statistics. *Journal of the American Statistical Association*, 78(382):351–365, 1983.

A Introduction

A.1 Proof of Theorem 1

Proof For a given d, INT outputs $\hat{f}_{I,\mathrm{INT}}$, the re-scaled-shifted $\hat{f}_{\bar{n}_d}$ given by the corresponding $\bar{n}_d \in \mathcal{N}_d$ in Lemma 8. Choosing $\epsilon(\delta) = \sqrt{5\log(1/\delta)/n}$, for $n \geq 128$, $\mathcal{Q}_{\epsilon(\delta)}$ occurs with probability $\geq 1-\delta$ from Lemma 5. Using Lemma 6 with $\epsilon(\delta)$ completes the proof.

A.2 Proof of Theorem 2

Proof Choosing $\epsilon(\delta) = \sqrt{5\log(n/\delta)/n}$, for $n \geq 128$, $\mathcal{Q}_{\epsilon(\delta)}$ occurs with probability $\geq 1-\delta$ from Lemma 5. Using Lemma 9 on top of Lemma 10 proves the theorem.

A.3 Proof of Corollary 3

Proof From Theorem 2, w.p. $\geq 1-\delta$,

$$\begin{aligned}
\|\hat{f}_{\mathcal{A}(X^{n-1})} - f\|_1 \leq \min_{\bar{I}\in\Delta_{\mathbb{R}}(X^{n-1})} \sum_{I\in\bar{I}} \Bigg(&\frac{(r_d+1)\cdot\alpha}{\alpha-2} \cdot \inf_{h\in\mathcal{P}_d} \|h-f\|_I \\
&+ \frac{r_d\cdot(\alpha\sqrt{2}+\sqrt{2}-1)}{(\sqrt{2}-1)^2} \sqrt{\frac{5(d+1)q_I\log\frac{n}{\delta}}{n}} \Bigg) \\
&\overset{(a)}{\leq} \min_{t\geq 0} \left(\frac{(r_d+1)\cdot\alpha}{\alpha-2}\mathrm{OPT}_{\mathcal{P}_{t,d}} + \frac{r_d\cdot(\alpha\sqrt{2}+\sqrt{2}-1)}{(\sqrt{2}-1)^2} \sqrt{\frac{5t\cdot(d+1)\log\frac{n}{\delta}}{n}} \right) \\
&\overset{(b)}{\leq} \min_{t\geq 0} \Bigg((r_d+1)\left(1+\frac{4}{\alpha}\right)\cdot \mathrm{OPT}_{\mathcal{P}_{t,d}}(f) \\
&\quad + \frac{r_d\cdot(\alpha\sqrt{2}+\sqrt{2}-1)}{(\sqrt{2}-1)^2} \sqrt{\frac{5t\cdot(d+1)\log\frac{n}{\delta}}{n}} \Bigg),
\end{aligned}$$

where (a) follows since for any partition with t pieces, $\sum_{I\in\bar{I}}\sqrt{q_I} \leq \sqrt{t}$, and (b) follows since for any $x > 4$ $x/(x-2) < 1 + 4/x$. Letting $\alpha \to \infty$ and choosing $\delta \approx 1/n$ completes the proof.

B Intervals and Partitions

B.1 Proof of Lemma 4

Proof For simplicity, let

$$X \overset{\text{def}}{=} P_{a,b}, \;\; p \overset{\text{def}}{=} q_{a,b}$$

so that $X = P_{a,b} \sim \mathrm{Beta}(nq_{a,b}, n(1-q_{a,b})) = \mathrm{Beta}(np, n(1-p))$. For any $x, y \in \mathbb{R}^+$, let $\mathrm{B}(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ denote the beta function and let $a, b > 0$ and $x \in [0,1]$,

$$I_x(a,b) \overset{\text{def}}{=} \int_0^x \frac{z^{a-1}(1-z)^{b-1}}{\mathrm{B}(a,b)} dz$$

be the incomplete beta function. Then,

$$\begin{aligned}
\Pr[X \leq p - \epsilon\sqrt{p}] &\overset{(a)}{=} I_{p-\epsilon\sqrt{p}}(np, n(1-p)) \\
&\overset{(b)}{=} \sum_{i=np}^{n-1} \binom{n-1}{i} (p-\epsilon\sqrt{p})^i (1-p+\epsilon\sqrt{p})^{n-1-i} \\
&\overset{(c)}{\leq} e^{-(n-1)D(p||p-\epsilon\sqrt{p})} \\
&\overset{(d)}{\leq} e^{-(n-1)\frac{\epsilon^2}{2}},
\end{aligned}$$

where (a) follows by definition, (b) follows by the property of incomplete beta function [7], (c) follows from the Chernoff bound applied to the right tail of a Binom$(n, p - \epsilon\sqrt{p})$ random variable, and (d) follows since $D(x||y) \le (x-y)^2/\max\{x,y\}$. Similarly,

$$
\begin{aligned}
\Pr[X \ge p + \epsilon\sqrt{p}] &\stackrel{(a)}{=} 1 - I_{(p+\epsilon\sqrt{p})}(np, n(1-p)) \\
&\stackrel{(b)}{=} 1 - \sum_{i=np}^{n-1} \binom{n-1}{i} (p+\epsilon\sqrt{p})^i (1-p-\epsilon\sqrt{p})^{n-1-i} \\
&\le \sum_{i=0}^{np} \binom{n-1}{i} (p+\epsilon\sqrt{p})^i (1-p-\epsilon\sqrt{p})^{n-1-i} \\
&\stackrel{(c)}{\le} e^{-(n-1)D(p||p+\epsilon\sqrt{p})} \\
&\stackrel{(d)}{\le} e^{-(n-1)\frac{\epsilon^2 p}{2(p+\epsilon\sqrt{p})}},
\end{aligned}
$$

where (a) follows by definition, (b) follows by the property of incomplete beta function [7], and (c) follows from Chernoff bound applied to the left tail of a Binom$(n, p + \epsilon\sqrt{p})$ random variable, and (d) follows since $D(x||y) \le (x-y)^2/\max\{x,y\}$.

B.2 Proof of Lemma 5

Proof From using the union bound, we have

$$
\begin{aligned}
1 - \Pr[\mathcal{Q}_\epsilon] &= \Pr\left[\exists 0 \le a < b \le n : |P_{a,b} - q_{a,b}| \ge \epsilon\sqrt{q_{a,b}}\right] \\
&\le \sum_{0 \le a < b \le n} \Pr\left[|P_{a,b} - q_{a,b}| \ge \epsilon\sqrt{q_{a,b}}\right] \\
&\stackrel{(a)}{\le} \sum_{0 \le a < b \le n} \left(e^{\frac{-(n-1)\epsilon^2}{2}} + e^{\frac{-(n-1)\epsilon^2 q_{a,b}}{2(q_{a,b}+\epsilon\sqrt{q_{a,b}})}} \right) \\
&\stackrel{(b)}{=} \frac{n(n+1)}{2} \left(e^{\frac{-(n-1)\epsilon^2}{2}} + e^{\frac{-(n-1)\epsilon^2 q_{a,b}}{2(q_{a,b}+\epsilon\sqrt{q_{a,b}})}} \right) \\
&\stackrel{(c)}{\le} \frac{n(n+1)}{2} \left(e^{\frac{-(n-1)\epsilon^2}{2}} + e^{-\frac{(n-1)\epsilon^2}{2(1+\epsilon\sqrt{n})}} \right),
\end{aligned}
$$

where (a) follows from Lemma 4, (b) follows since $|0 \le a < b \le n| = \binom{n+1}{2}$, and (c) follows since $q_{a,b} \ge 1/n$.

C The Interpolation Routine

C.1 Proof of Lemma 6

Proof For a partition $\bar{I}$ of $I = [0,1]$, and integrable functions g_1, g_2, define the distance

$$
d_{\bar{I}}(g_1, g_2) \stackrel{\text{def}}{=} \sum_{J \in \bar{I}} \Big| \int_J g_1 - \int_J g_2 \Big|. \tag{4}
$$

In words, $d_{\bar{I}}(g_1, g_2)$ is the sum of absolute differences between measures under g_1 and g_2 across all intervals in $\bar{I}$. For any $h \in \mathcal{P}_d$,

$$
\begin{aligned}
\|\hat{f}_{\bar{n}_d} - f\|_I &\le \|h - f\|_I + \|\hat{f}_{\bar{n}_d} - h\|_I \\
&\stackrel{(a)}{\le} \|h - f\|_I + r_d(\bar{n}_d) d_{\bar{I}_{\bar{n}_d}}(h, \hat{f}_{\bar{n}_d}) \\
&\stackrel{(b)}{\le} \|h - f\|_I + r_d(\bar{n}_d)\left(d_{\bar{I}_{\bar{n}_d}}(h, f) + d_{\bar{I}_{\bar{n}_d}}(f, \hat{f}_{\bar{n}_d})\right) \\
&\stackrel{(c)}{\le} (1 + r_d(\bar{n}_d))\|h - f\|_I + r_d(\bar{n}_d) d_{\bar{I}_{\bar{n}_d}}(f, \hat{f}_{\bar{n}_d}) \\
&\stackrel{(d)}{\le} (1 + r_d(\bar{n}_d))\|h - f\|_I + r_d(\bar{n}_d) \sum_{J \in \bar{I}_{\bar{n}_d}} |P_J - q_J|.
\end{aligned}
$$

where (a) follows since $(h - \hat{f}_{I,\bar{n}_d}) \in \mathcal{P}_d$, and from definitions of the ratio $r_d(\bar{n}_d)$ in Equation (1), (b) follows since the $d_{\bar{I}}$-distance satisfies the triangle inequality, (c) follows since the ℓ_1 distance upper bounds $d_{\bar{I}}-$distance, (d) follows since $\hat{f}_{\bar{n}_d}$, by definition, is the polynomial such that $\int_J \hat{f}_{\bar{n}_d} = q_J$ $\forall J \in \bar{I}_{\bar{n}_d}$, and the interval probability $P_J \stackrel{\text{def}}{=} \int_J f$.

Since $P_J \sim \text{Beta}(q_j n, (1 - P_J)n)$, it follows that

$$
\begin{aligned}
\mathbb{E}\|\hat{f}_{\bar{n}_d} - f\|_I &\le (1 - r_d(\bar{n}_d))\|h - f\|_I + r_d(\bar{n}_d) \sum_{J \in \bar{I}_{\bar{n}_d}} \mathbb{E}|P_J - q_J| \\
&\stackrel{(a)}{\le} (1 - r_d(\bar{n}_d))\|h - f\|_I + r_d(\bar{n}_d) \sum_{J \in \bar{I}_{\bar{n}_d}} \sqrt{\frac{2q_J}{\pi n}} \\
&\stackrel{(b)}{\le} (1 + r_d(\bar{n}_d))\|h - f\|_I + r_d(\bar{n}_d)\sqrt{\frac{2(d+1)q_I}{\pi n}}.
\end{aligned}
$$

where (a) follows from the mean absolute deviation of the Beta distribution, and (b) follows since by the concavity of $\sqrt{x}$ for $x \ge 0$, the sum $\sum_{J \in \bar{I}_{\bar{n}_d}} \sqrt{q_J}$ is maximized if for each $J \in \bar{I}_{\bar{n}_d}$, $q_J = (\sum_{J \in \bar{I}_{\bar{n}_d}} q_J)/|\bar{I}_{\bar{n}_d}| = q_I/(d+1)$.

The following version will be useful in the proof of Lemma 10. Under event $\mathcal{Q}_\epsilon$,

$$
\begin{aligned}
\mathbb{E}\|\hat{f}_{\bar{n}_d} - f\|_I &\le (1 - r_d(\bar{n}_d))\|h - f\|_I + r_d(\bar{n}_d) \sum_{J \in \bar{I}_{\bar{n}_d}} \mathbb{E}|P_J - q_J| \\
&\le (1 + r_d(\bar{n}_d))\|h - f\|_I + r_d(\bar{n}_d) \sum_{J \in \bar{I}_{\bar{n}_d}} \epsilon\sqrt{q_J} \\
&\stackrel{(a)}{\le} (1 + r_d(\bar{n}_d))\|h - f\|_I + r_d(\bar{n}_d)\epsilon\sqrt{(d+1)q_I},
\end{aligned}
$$

where (a) follows due to the same reasoning as above.

C.2 Proof of Lemma 7

Proof Fix $h \in \mathcal{P}_d$. Let $(\beta_1, , \ldots, , \beta_{d_0})$ be the roots of h in $[0,1]$ for some $\beta_1 \le, \ldots, \le \beta_{d_0}, 0 \le d_0 \le d$. Let $\beta_0 \stackrel{\text{def}}{=} 0, \beta_{d_0+1} \stackrel{\text{def}}{=} 1$. Notice that

$$
\begin{aligned}
\int_0^1 |h| = \sum_{i=1}^{d_0+1} \Big| \int_{\beta_{i-1}}^{\beta_i} h \Big| &\stackrel{(a)}{\le} \sup_{\bar{m}_d \in \mathcal{N}_d} \sum_{i=1}^{d+1} \Big| \int_{m_{i-1}}^{m_i} h \Big| \\
&= \sup_{\bar{m}_d \in \mathcal{N}_d} \max_{\bar{s} \in \{0,1\}^{d+1}} \sum_{i=1}^{d+1} (-1)^{s_i} \int_{m_{i-1}}^{m_i} h \\
&\stackrel{(b)}{\le} \int_0^1 |h|,
\end{aligned}
$$

where (a) follows since on padding $d-d_0$ zeros, $(0,\cdots,0,\beta_0,\ldots,\beta_{d+1})\in\mathcal{N}_d$. Thus (b) is, in fact, an equality, implying

$$\begin{aligned}
r_d(\bar{n}_d) = \sup_{h\in\mathcal{P}_d} r(\bar{n}_d,h) &= \sup_{h\in\mathcal{P}_d} \frac{\int_0^1 |h|}{\sum_{i=1}^{d+1} |\int_{n_{i-1}}^{n_i} h|} \\
&= \sup_{h\in\mathcal{P}_d} \frac{\sup_{\bar{m}_d\in\mathcal{N}_d} \max_{\bar{s}\in\{0,1\}^{d+1}} \sum_{i=1}^{d+1}(-1)^{s_i}\int_{m_{i-1}}^{m_i} h}{\sum_{i=1}^{d+1} |\int_{n_{i-1}}^{n_i} h|} \\
&= \sup_{\bar{m}_d\in\mathcal{N}_d} \max_{\bar{s}\in\{0,1\}^{d+1}} \sup_{h\in\mathcal{P}_d} \frac{\sum_{i=1}^{d+1}(-1)^{s_i}\int_{m_{i-1}}^{m_i} h}{\sum_{i=1}^{d+1} |\int_{n_{i-1}}^{n_i} h|}.
\end{aligned}$$

Denote $h=\sum_{i=1}^{d+1} c_i\cdot x^{i-1}$ and let $\bar{c}\overset{\text{def}}{=}(c_1,\ldots,c_{d+1})$. Notice that since $r(\bar{n}_d,h)\geq 1$ for any $h\in\mathcal{P}_d$, and since $r_d(\bar{n}_d,0)\overset{\text{def}}{=}1$, WLOG assume $h\neq 0$ or $\bar{c}\neq\bar{0}\overset{\text{def}}{=}(0,\ldots,0)$. By linearity of the integral of h in $\bar{c}$, recast $r_d(\bar{n}_d)$ into

$$r_d(\bar{n}_d) = \sup_{\bar{m}_d\in\mathcal{N}_d} \max_{\bar{s}\in\{0,1\}^{d+1}} \sup_{\bar{c}\in\mathbb{R}^{d+1}\setminus\{\bar{0}\}} \frac{\sum_{i=1}^{d+1} c_i\mu_i}{\sum_{i=1}^{d+1} |\sum_{j=1}^{d+1} c_j\lambda_{i,j}|},$$

where for any $i,j\in\{1,\ldots,d+1\}$, $\mu_i\in\mathbb{R}$ is a function of $\bar{m}_d,\bar{s}$ and $\lambda_{i,j}\in\mathbb{R}$ is a function of $\bar{n}_d$. Observe that $\bar{n}_d$ is given, and additionally fix $\bar{m}_d\in\mathcal{N}_d$, $\bar{s}\in\{0,1\}^{d+1}$. Since the objective function here is a ratio whose denominator is positive (since $h\neq 0$), WLOG set the numerator to 1 via the constraint $\sum_{i=1}^{d+1} c_i\mu_i=1$ and convert it to a linear program as:

$$\max \frac{1}{\sum_{i=1}^{d+1} v_i} : \bar{c},\bar{v}\in\mathbb{R}^{d+1},\ \ v_i\geq\sum_{j=1}^{d+1} c_j\lambda_{i,j},\ v_i\geq-\sum_{j=1}^{d+1} c_j\lambda_{i,j}, \sum_{i=1}^{d+1} c_i\mu_i=1,$$

where $\bar{v}\overset{\text{def}}{=}(v_1,\ldots,v_{d+1})$. Observe that these constraints give rise to a bounded region, and since this is a linear program, there exists a solution at some corner point involving at least $2\cdot(d+1)$ equalities, one for each variable. In any such solution, since the equality: $\sum_{i=1}^{d+1} c_i\mu_i=1$ is always active, at least $2\cdot(d+1)-1$ of the other inequalities attain equality. Notice that for any $i\in\{1,\ldots,d+1\}$, $v_i=0$ if both

$$v_i=\sum_{j=1}^{d+1} c_j\lambda_{i,j} \text{ and } v_i=-\sum_{j=1}^{d+1} c_j\lambda_{i,j} \text{ hold.}$$

Thus in this corner point solution, $v_i\neq 0$ for at most one $i\in\{1,\ldots,d+1\}$. Let

$$\mathcal{D}_{\bar{n}_d}=\left\{\bar{c}\in\mathbb{R}^{d+1}\setminus\{\bar{0}\} : \exists i_1\in\{1,\ldots,d+1\} : \forall i\neq i_1,\ |\sum_{j=1}^{d+1} c_j\lambda_{i,j}|=0\right\}$$

This implies

$$\begin{aligned}
r_d(\bar{n}_d) &= \sup_{\bar{m}_d\in\mathcal{N}_d} \max_{\bar{s}\in\{0,1\}^{d+1}} \sup_{\bar{c}\in\mathbb{R}^{d+1}\setminus\{\bar{0}\}} \frac{\sum_{i=1}^{d+1} c_i\mu_i}{\sum_{i=1}^{d+1} |\sum_{j=1}^{d+1} c_j\lambda_{i,j}|} \\
&= \sup_{\bar{m}_d\in\mathcal{N}_d} \max_{\bar{s}\in\{0,1\}^{d+1}} \max_{\bar{c}\in\mathcal{D}_{\bar{n}_d}} \frac{\sum_{i=1}^{d+1} c_i\mu_i}{\sum_{i=1}^{d+1} |\sum_{j=1}^{d+1} c_j\lambda_{i,j}|} \\
&= \sup_{\bar{m}_d\in\mathcal{N}_d} \max_{\bar{s}\in\{0,1\}^{d+1}} \max_{h\in\mathcal{P}_{\bar{n}_d}} \frac{\sum_{i=1}^{d+1}(-1)^{s_i}\int_{m_{i-1}}^{m_i} h}{\sum_{i=1}^{d+1} |\int_{n_{i-1}}^{n_i} h|} \\
&= \max_{h\in\mathcal{P}_{\bar{n}_d}} \sup_{\bar{m}_d\in\mathcal{N}_d} \max_{\bar{s}\in\{0,1\}^{d+1}} \frac{\sum_{i=1}^{d+1}(-1)^{s_i}\int_{m_{i-1}}^{m_i} h}{\sum_{i=1}^{d+1} |\int_{n_{i-1}}^{n_i} h|} \\
&= \max_{h\in\mathcal{P}_{\bar{n}_d}} \frac{\int_0^1 |h|}{\sum_{i=1}^{d+1} |\int_{n_{i-1}}^{n_i} h|} = \max_{h\in\mathcal{P}_{\bar{n}_d}} r_d(\bar{n}_d,h).
\end{aligned}$$

C.3 Proof of Lemma 8

Proof For any polynomial $h \in \mathcal{P}_d$, the ratio $r(\bar{n}_d, h)$ is invariant to multiplying both the numerator and denominator by a constant. Thus, WLOG consider polynomials whose leading coefficient is 1. Then for any $\bar{n}_d \in \mathcal{N}_d$, $\mathcal{P}_{\bar{n}_d} = (h_{\bar{n}_d,1}, \ldots, h_{\bar{n}_d,d+1})$, is a set consisting of $d+1$ unique polynomials, where each $h_{\bar{n}_d,i},\ i \in \{1, \ldots, d+1\}$ is that polynomial with 0 area in all intervals in $\bar{I}_{\bar{n}_d}$ except $I_{\bar{n}_d,i}$.

Case $\mathbf{d=0}$: Here $\mathcal{N}_0 = \{(0,1)\}$ and is a singleton set. Since any $h \in \mathcal{P}_0$ is a constant value, $\int_0^1 |h| = |\int_0^1 h|$. Therefore $r_0^\star = \max_{h \in \mathcal{P}_0} r(\bar{n}_d, h) = 1$.

Case $\mathbf{d=1}$: Let $\bar{n}_1 = (0, m, 1)$. In this case $h_{\bar{n}_d,1}(x) = x - m/2$ and $h_{\bar{n}_d,2}(x) = x - (1+m)/2$. Using Lemma 7,

$$\begin{aligned} r_1(\bar{n}_d) &= \max_{h \in \mathcal{P}_{\bar{n}_d}} r(\bar{n}_d, h) = \max\{r(\bar{n}_d, h_{\bar{n}_d,1}), r(\bar{n}_d, h_{\bar{n}_d,2})\} \\ &= \max\Big\{\frac{m^2/4 + (1-m/2)^2}{1-m}, \frac{(1-m)^2/4 + ((1+m)/2)^2}{m}\Big\}. \end{aligned}$$

$r_1(\bar{n}_d)$ is minimized for $m^\star = 1/2$, giving $r_1^\star = (1/16 + 9/16)/(1/2) = 1.25$.

Case $\mathbf{d=2}$: By symmetry, the minimizing node partition is symmetric about 0.5. Thus WLOG let $\bar{n}_2 = (0, m, 1-m, 1)$ for some $m \leq 0.5$. Among the $d+1 = 3$ polynomials in $\mathcal{P}_{\bar{n}_2}$, by symmetry of $\bar{n}_2$, $r_2(h_{\bar{n}_d,1}) = r_2(h_{\bar{n}_d,3})$. Thus we consider the larger ratio across only two polynomials, $h_{\bar{n}_d,2}, h_{\bar{n}_d,3}$.

Denote the polynomial as $h_{\bar{n}_d,2}(x) = (x-a_2)^2 - b_2^2$ and upon setting the respective integrals to 0,

$$\Big|\frac{m^3}{3} - a_2 m^2 + (a_2^2 - b_2^2)m\Big| = 0,\ \Big|\frac{1-(1-m)^3}{3} - a_2(1-(1-m)^2) + (a_2^2 - b_2^2)m\Big| = 0$$

$$\implies a_2 = \frac{1}{2}, b_2^2 = \frac{3(m^2-m)+1}{9}.$$

Representing $h_{\bar{n}_d,3}(x) = (x-a_3)^2 - b_3^2$ and repeating the same steps,

$$\Big|\frac{m^3}{3} - a_3 m^2 + (a_3^2 - b_3^2)m\Big| = 0,\ \Big|\frac{(1-m)^3 - m^3}{3} - a_3((1-m)^2 - m^2) + (a_3^2 - b_3^2)(1-2m)\Big| = 0$$

$$\implies a_3 = \frac{1}{3}, b_3^2 = \frac{4m^2 - 6m + 3}{3}.$$

The corresponding $r(\bar{n}_d, h_{\bar{n}_d,2})$ and $r(\bar{n}_d, h_{\bar{n}_d,3})$ are given by

$$\frac{8\left(\frac{1-3m(1-m)}{9}\right)^{3/2}}{m(1-m)} + 1,\ \frac{2\left(\frac{(2m-1)(2m-2)+1}{3}\right)^{3/2}}{(2m-1)(m-1)} - 1.$$

From simultaneously minimizing the above expressions by equating them, the optimal m is the root of

$$\begin{aligned} q_2(m) = &-\frac{26624}{729}m^{14} + \frac{193280}{729}m^{13} - \frac{211024}{243}m^{12} + \frac{3703648}{2187}m^{11} - \frac{4790776}{2187}m^{10} \\ &+ \frac{39108232}{19683}m^9 - \frac{8554775}{6561}m^8 + \frac{12357280}{19683}m^7 - \frac{13004032}{59049}m^6 + \frac{1061792}{19683}m^5 \\ &- \frac{4350752}{531441}m^4 + \frac{246976}{531441}m^3 + \frac{11840}{177147}m^2 - \frac{6656}{531441}m + \frac{256}{531441} \end{aligned}$$

near 0.26. Thus the optimal $m^\star \approx 0.2599$ and the corresponding $r_2^\star \approx 1.423$.

Case $\mathbf{d=3}$: By symmetry, as before, WLOG let $\bar{n}_d = (0, m, 0.5, 1-m, 1)$. This reduces the search space to just two polynomials, $h_{\bar{n}_d,1}, h_{\bar{n}_d,2}$. The optimal m occurs as the root of

$$q_3(m) \stackrel{\text{def}}{=} m^{69} + \frac{2233}{46}m^{68} + \frac{3394851}{2944}m^{67} - \frac{26295551}{1472}m^{66} + \frac{76466381715}{376832}m^{65}$$

$$
\begin{aligned}
&- \frac{1357944230009}{753664}m^{64} + \frac{627961733592749}{48234496}m^{63} - \frac{3795194179761079}{48234496}m^{62} \\
&+ \frac{1252499739594399621}{3087007744}m^{61} - \frac{5593584650474780121}{3087007744}m^{60} \\
&+ \frac{87541700408454835933}{12348030976}m^{59} - \frac{9689649149944354300097}{395136991232}m^{58} \\
&+ \frac{477481388280085878102175}{6322191859712}m^{57} - \frac{1316816736336377796401265}{6322191859712}m^{56} \\
&+ \frac{104518853645525535726426411}{202310139510784}m^{55} - \frac{935729957191660731330480575}{809240558043136}m^{54} \\
&+ \frac{1894032003216065918256250147}{809240558043136}m^{53} - \frac{111070063686665905121657252873}{25895697857380352}m^{52} \\
&+ \frac{2947966937880382636398337723253}{414331165718085632}m^{51} - \frac{96053130159826779148472826511}{9007199254740992}m^{50} \\
&+ \frac{5957768773291898355944143881565}{414331165718085632}m^{49} \\
&- \frac{28666092800309188568756667723285}{1657324662872342528}m^{48} \\
&+ \frac{1936825259310147713677259087614429}{106068778423829921792}m^{47} \\
&- \frac{21666345695949567710248390395 5187}{13258597302978740224}m^{46} \\
&+ \frac{4797689934446961630160031643189779}{424275113695319687168}m^{45} \\
&- \frac{6609917603386978813128872815736861}{1697100454781278748672}m^{44} \\
&- \frac{1831051294952734349349124229564767}{424275113695319687168}m^{43} \\
&+ \frac{77448257249275962672807384094624197}{6788401819125114994688}m^{42} \\
&- \frac{343821157459641486443539964855757 5571}{21722885821200367983 0016}m^{41} \\
&+ \frac{14691873341415043555417961121466375911}{868915432848014719320064}m^{40} \\
&- \frac{52197058412928213666930233026164438477}{3475661731392058877280256}m^{39} \\
&+ \frac{626588406181557032617836659688444683449}{55610587702272942036484096}m^{38} \\
&- \frac{1555525051509188771980730278198868813547}{222442350809091768145936384}m^{37} \\
&+ \frac{2911728348738530370986950039544396475929}{889769403236367072583745536}m^{36} \\
&- \frac{415629763783269606797247480824967018319}{618970019642690137449562112}m^{35} \\
&- \frac{21293950227855325096203381076755307029285}{28472620903563746322679857152}m^{34} \\
&+ \frac{568117182824342053622453013472967651433283}{455561934457019941162877714432}m^{33} \\
&- \frac{544176155073471826876603760623685652791461}{455561934457019941162877714432}m^{32} \\
&+ \frac{26510792231823207111468208052472537291091795}{29155963805249276234424173723648}m^{31} \\
&- \frac{433925296736104935336380333442235614348 4711}{7288990951312319058606043430912}m^{30}
\end{aligned}
$$

$$
\begin{aligned}
&+ \frac{80732188658524254663695919868749449919506951}{2332477104419942098753933897891 84}m^{29} \\
&- \frac{21178612540977853104278860543475186392271681}{116623855220997104937696694894592}m^{28} \\
&+ \frac{80945977995703186772569013525777965350114711}{932990841767976839501573559156736}m^{27} \\
&- \frac{17702912107607481296775724392303230443070095}{466495420883988419750786779578368}m^{26} \\
&+ \frac{113795689628398207709330045266274999410341749}{7463926734143814716012588473253888}m^{25} \\
&- \frac{21034631455622622763256522489198166970683613}{373196336707190735800629423662694 4}m^{24} \\
&+ \frac{229167532475013797669274973054071362594872031}{11942282774630103545620141557206 2208}m^{23} \\
&- \frac{17963282649430914596699742603270996080670419}{29855706936575258864050353893015552}m^{22} \\
&+ \frac{165857532854257459359651249063603517430127219}{95538262197040828364961132457649766 4}m^{21} \\
&- \frac{21988393007280087042133518345726499678514793}{477691310985204141824805662288248832}m^{20} \\
&+ \frac{2673062517068208952962202401457887091620815}{238845655492602070912402831144124416}m^{19} \\
&- \frac{4753573760248527366859497435267453417384687}{1910765243940816567299222649152995328}m^{18} \\
&+ \frac{7698530432612172216840174126420917394335697}{15286121951526532538393781193223962624}m^{17} \\
&- \frac{882543917938320973760720615032337965405 39}{9553826219704082836496113245764976 64}m^{16} \\
&+ \frac{46618175901212180036351783955592685629293 1}{30572243903053065076787562386447925248}m^{15} \\
&- \frac{34314489379863699139383468530966229926169}{152861219515265325383937811932239626 24}m^{14} \\
&+ \frac{71176987410160907949890583502121075660389}{244577951224424520614300499091583401984}m^{13} \\
&- \frac{3987333962073343668163889901656755306225}{122288975612212260307150249545791700992}m^{12} \\
&+ \frac{748668655114740354600745066271570265031}{244577951224424520614300499091583401984}m^{11} \\
&- \frac{110845256761306163606440669442292637027}{489155902448849041228600998183166803968}m^{10} \\
&+ \frac{1899176964083859703283044829743633209}{1701411834604692317316873037158841057 28}m^{9} \\
&- \frac{20739215847923545637371202116833 41}{425352958651173079329218259289710264 32}m^{8} \\
&- \frac{34402390201821002633790825125211 75}{611444878061061301535751247728958504 96}m^{7} \\
&+ \frac{4264138856148752641548430451717}{66461399789245793645190353014017228 8}m^{6} \\
&- \frac{311750355103578111892964488373 1}{764306097576326626919689059661198131 2}m^{5} \\
&+ \frac{2903707718203711911272212542 3}{1910765243940816567299222649152995328}m^{4} \\
&- \frac{460833093614235731783726 79}{238845655492602070912402831144124416}m^{3}
\end{aligned}
$$

$$
\begin{aligned}
&- \frac{5978119273184036056855541}{59711413873150517728100707786031104} m^2 \\
&+ \frac{3010486164998286948083 1}{59711413873150517728100707786031104} m \\
&- \frac{105905655782897976459}{1492785346828762943202517694650 7776}
\end{aligned}
$$

near 0.155. This gives $m^\star \approx 0.1548$ and the corresponding $r_3^\star \approx 1.559$.

For degrees $4 \le d \le 8$, we use numerical methods on top of Lemma 7 to derive $\bar{n}_d \in \mathcal{N}_d$ and the corresponding $r_d(\bar{n}_d)$. These values populate the second table in Lemma 8.

D The Compare and Stitch Routines

D.1 Proof of Lemma 9

Proof As observed in Equation (**??**), any $q \in \bar{q} \in \Delta_{\text{emp,n}}$ is an integral multiple of $1/n$. Observing that $\log_2 n$ is an integer, we may decompose q along its binary expansion as

$$q = \sum_{j=0}^{\log_2 n} 2^{-j} b_j,$$

for some $b_j \in \{0,1\}$, $j \in \{1, \ldots, \log_2 n\}$. Replace each $q \in \bar{q}$ with the vector $(2^{-0}b_0, 2^{-1}b_1, \cdots)$ to obtain $\bar{q}' \in \Delta_{\text{bin,n}}$. From the property of the geometric sum,

$$\sum_{j=0}^{\log_2 n} \sqrt{2^{-j} b_j} \le \frac{\sqrt{q}}{\sqrt{2}-1}.$$

Finally $\|f^\star_{\bar{I}_{\bar{q}'}} - f\|_1 \le \|f^\star_{\bar{I}_{\bar{q}}} - f\|_1$ since $\bar{I}_{\bar{q}'}$ being a finer partition than $\bar{I}_{\bar{q}}$, $f^\star_{\bar{I}_{\bar{q}'}}$ is a closer approximation to f than $f^\star_{\bar{I}_{\bar{q}}}$.

D.2 Proof of Lemma 10

Proof For any interval I, let

$$f^\star_I \stackrel{\text{def}}{=} \arg\min_{h \in \mathcal{P}_d} \|h - f\|_I,$$

and for any partition $\bar{I}$, let $f^\star_{\bar{I}}$ be the piecewise polynomial that equals $f^\star_I$ in each $I \in \bar{I}$. For simplicity let $I_{\bar{q}} \stackrel{\text{def}}{=} I_{\bar{q}_D}$ denote the final partition and $\bar{q} \stackrel{\text{def}}{=} \bar{q}_D$ the corresponding empirical distribution. Consider any $\bar{p} \in \Delta_{\text{bin,n}}$ and its associated interval partition, $\bar{I}_{\bar{p}}$. Two interval partitions $\bar{I}_1, \bar{I}_2$ corresponding to binary distributions have the following property: Any interval in $\bar{I}_1$ is either completely contained within some interval in $\bar{I}_2$, or is a union of contiguous intervals from $\bar{I}_2$. As a result $\bar{I}_{\bar{q}}$ may partitioned into three classes of intervals:

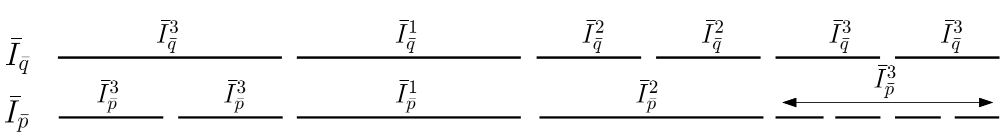


Figure 5: Illustration of $\bar{I}_{\bar{q}}$ being partitioned into $\bar{I}^1_{\bar{q}}$, $\bar{I}^2_{\bar{q}}$ and $\bar{I}^3_{\bar{q}}$ using $\bar{I}_{\bar{p}}$.

- $\bar{I}^1_{\bar{q}}$, composed of intervals that are equal to some interval in $\bar{I}_{\bar{p}}$,
- $\bar{I}^2_{\bar{q}}$, that consists of intervals that lie strictly within some interval in $\bar{I}_{\bar{p}}$,
- $\bar{I}^3_{\bar{q}}$, containing intervals that are unions of more than one interval from $\bar{I}_{\bar{p}}$.

This is shown in Figure 5. Lemmas 11, 12, 13 address each of these intervals separately. Combining the lemmas,

$$
\begin{aligned}
\|\hat{f}_{\bar{I}_{\bar{q}}} - f\|_1 &= \|\hat{f}_{\bar{I}_{\bar{q}}} - f\|_{\bar{I}^1_{\bar{q}}} + \|\hat{f}_{\bar{I}_{\bar{q}}} - f\|_{\bar{I}^2_{\bar{q}}} + \|\hat{f}_{\bar{I}_{\bar{q}}} - f\|_{\bar{I}^3_{\bar{q}}} \\
&\le (r_d+1)\cdot \|f^\star_{\bar{I}_{\bar{p}}} - f\|_{\bar{I}^1_{\bar{q}}} + \sum_{I\in \bar{I}^1_{\bar{p}}} r_d \cdot \epsilon\sqrt{(d+1)p_I} \\
&\quad + \frac{(r_d+1)\cdot\alpha}{\alpha-2}\cdot \|f^\star_{\bar{I}_{\bar{p}}} - f\|_{\bar{I}^2_{\bar{q}}} + \frac{1}{\alpha-1}\sum_{I\in \bar{I}^2_{\bar{p}}} r_d \cdot \epsilon\sqrt{(d+1)p_I} \\
&\quad + (r_d+1)\cdot \|f^\star_{\bar{I}_{\bar{p}}} - f\|_{\bar{I}^3_{\bar{q}}} + \frac{\alpha\sqrt{2}+\sqrt{2}-1}{\sqrt{2}-1}\sum_{I\in \bar{I}^3_{\bar{p}}} r_d \cdot \epsilon\sqrt{(d+1)p_I} \\
&\overset{(a)}{\le} \frac{(r_d+1)\cdot\alpha}{\alpha-2}\|f^\star_{\bar{I}_{\bar{p}}} - f\|_1 + \frac{\alpha\sqrt{2}+\sqrt{2}-1}{\sqrt{2}-1}\sum_{I\in \bar{I}_{\bar{p}}} r_d \cdot \epsilon\sqrt{(d+1)p_I},
\end{aligned}
$$

where (a) follows since $\alpha > 2 \Rightarrow 1/(\alpha-1) < 1 < (\alpha\sqrt{2}+\sqrt{2}-1)/(\sqrt{2}-1)$.

Lemma 11. *For the final partition $\bar{I}_{\bar{q}}$ in the run of* MERGE *and any $\bar{p} \in \Delta_{\text{bin,n}}$, let $\bar{I}^1_{\bar{q}} \subseteq \bar{I}_{\bar{q}}$ be the intervals that intersect with $\bar{I}_{\bar{p}}$. Let $\bar{I}^1_{\bar{p}} = \bar{I}^1_{\bar{q}} \subseteq \bar{I}_{\bar{p}}$ denote the corresponding collection in $\bar{I}_{\bar{p}}$. Then,*

$$
\|\hat{f}_{\bar{I}_{\bar{q}}} - f\|_{\bar{I}^1_{\bar{q}}} \le (r_d+1)\cdot \|f^\star_{\bar{I}_{\bar{p}}} - f\|_{\bar{I}^1_{\bar{q}}} + \sum_{I\in \bar{I}^1_{\bar{p}}} r_d \cdot \epsilon\sqrt{(d+1)p_I},
$$

Proof Follows from Theorem 1 and noticing that intervals in $\bar{I}^1_{\bar{q}}$ and $\bar{I}^1_{\bar{p}}$ coincide.

Lemma 12. *For the final partition $\bar{I}_{\bar{q}}$ in the run of* MERGE *and any $\bar{p} \in \Delta_{\text{bin,n}}$, let $\bar{I}^2_{\bar{q}} \subseteq \bar{I}_{\bar{q}}$ be the intervals that do not intersect with, and strictly lie in some interval in $\bar{I}_{\bar{p}}$. Let $\bar{I}^2_{\bar{p}} \subseteq \bar{I}_{\bar{p}}$ be the corresponding intervals that contain $\bar{I}^2_{\bar{q}}$. Then,*

$$
\|\hat{f}_{\bar{I}_{\bar{q}}} - f\|_{\bar{I}^2_{\bar{q}}} \le \frac{(r_d+1)\cdot\alpha}{\alpha-2}\cdot \|f^\star_{\bar{I}_{\bar{p}}} - f\|_{\bar{I}^2_{\bar{q}}} + \frac{1}{\alpha-1}\sum_{I\in \bar{I}^2_{\bar{p}}} r_d \cdot \epsilon\sqrt{(d+1)p_I}.
$$

Proof Notice that all intervals in $\bar{I}^2_{\bar{q}}$ are strictly contained within some interval in $\bar{I}^2_{\bar{p}}$. Using this, we further partition $\bar{I}^2_{\bar{q}}$ using intervals in $\bar{I}^2_{\bar{p}}$. Fix an $I \in \bar{I}^2_{\bar{p}}$ and let $\bar{I} \in \bar{I}^2_{\bar{q}}$ be intervals whose union gives I. Let $\bar{q}_{\bar{I}} \subseteq \bar{q}$ denote the empirical probabilities corresponding to $\bar{I}$ and let p_I denote the empirical probability under I.

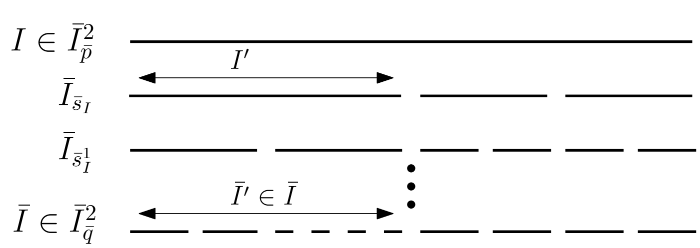


Figure 6: Illustration of $\bar{I} \in \bar{I}^2_{\bar{q}}$, $\bar{I}_{\bar{s}_I}$ and $\bar{I}_{\bar{s}^1_I}$ corresponding to a particular $I \in \bar{I}^2_{\bar{p}}$.

While $\bar{q}_{\bar{I}}$ is a sub-distribution in general, WLOG assume $\bar{q}_{\bar{I}}$ is a distribution. Now, at some point in the run of MERGE, COMP was called with $\hat{f}_{I,\text{INT}}$, $\bar{I}$, $\bar{q}_{\bar{I}}$, and it was in-turn declared that $\bar{I}$ was not to be merged into I. Therefore, for the $\mu_{\bar{I},\gamma}(\hat{f}_{I,\text{INT}})$ attaining binary distribution, $\bar{s}_I \in \Delta_{\text{bin,n},\ge \bar{q}_{\bar{I}}}$,

$\Lambda_{I_{\bar{s}_I}}(\hat{f}_{I,\text{INT}}) - \lambda_{\bar{s}_I,\gamma} \geq 0$. It follows that

$$\begin{aligned}
\sum_{s\in\bar{s}_I} \alpha\cdot r_d\cdot\epsilon\sqrt{(d+1)s} = \lambda_{\bar{s}_I,\gamma} &\leq \Lambda_{I_{\bar{s}_I}}(\hat{f}_{I,\text{INT}}) \\
&= \|\hat{f}_{I,\text{INT}} - \hat{f}_{\bar{I}_{\bar{s}_I}}\|_I \\
&\leq \|\hat{f}_{I,\text{INT}} - f\|_I + \|f - \hat{f}_{\bar{I}_{\bar{s}_I}}\|_I \\
&\overset{(a)}{\leq} (r_d+1)\cdot\|f_I^\star - f\|_I + r_d\cdot\epsilon\sqrt{(d+1)p_I} \\
&\quad + (r_d+1)\cdot\|f^\star_{\bar{I}_{\bar{s}_I}} - f\|_I + \sum_{s\in\bar{s}_I} r_d\cdot\epsilon\sqrt{(d+1)s} \\
&\overset{(b)}{\leq} 2(r_d+1)\cdot\|f_I^\star - f\|_I + r_d\cdot\epsilon\sqrt{(d+1)p_I} + \sum_{s\in\bar{s}_I} r_d\cdot\epsilon\sqrt{(d+1)s},
\end{aligned}$$

where (a) follows from in Theorem 1, (b) follows since I being the union of $\bar{I}_{\bar{q}_I}$ is also the union of $\bar{I}_{\bar{s}_I}$, and $f_I^\star$ is therefore a coarser approximation to f than $f^\star_{\bar{I}_{\bar{s}_I}}$, giving rise to a larger ℓ_1 distance. Rearrange this to obtain

$$\sum_{s\in\bar{s}_I} r_d\cdot\epsilon\sqrt{(d+1)s} \leq \frac{1}{\alpha-1}\cdot\left(2(r_d+1)\cdot\|f_I^\star - f\|_I + r_d\cdot\epsilon\sqrt{(d+1)p_I}\right). \tag{5}$$

Consider a fixed $I' \in \bar{I}_{\bar{s}_I}$ and let $\bar{I}' \in \bar{I}$ be the intervals under $\bar{I}$ whose union gives I'. We recursively use the same argument to bound the LHS of Equation (5). This is shown for the leftmost interval of $\bar{I}_{\bar{s}_I}$ in Figure 6. Let $\bar{q}_{\bar{I}'}$ be the corresponding probabilities under $\bar{I}'$ and let $s_{I'}$ denote the empirical probability under I'. Notice that in some previous step of MERGE, as was for I, COMP was invoked with $\hat{f}_{I',\text{INT}}$, $\bar{I}'$, $\bar{q}_{\bar{I}'}$, for which $\mu_{\bar{I}',\gamma}(\hat{f}_{I',\text{INT}}) \geq 0$. Repeat the same procedure as above to obtain

$$\begin{aligned}
\sum_{s\in\bar{s}_{I'}} r_d\cdot\epsilon\sqrt{(d+1)s} &\leq \frac{1}{\alpha-1}\cdot\left(2(r_d+1)\cdot\|f^\star_{I'} - f\|_{I'} + r_d\cdot\epsilon\sqrt{(d+1)s_{I'}}\right) \\
&\overset{(a)}{\leq} \frac{1}{\alpha-1}\cdot\left(2(r_d+1)\cdot\|f^\star_{I} - f\|_{I'} + r_d\cdot\epsilon\sqrt{(d+1)s_{I'}}\right),
\end{aligned} \tag{6}$$

where $\bar{s}_{I'}$ here is the binary distribution which attains $\mu_{\bar{I}',\gamma}(\hat{f}_{I',\text{INT}})$, and (a) follows because I' being an interval within I, $f_I^\star$ is a coarser approximation to f than $f^\star_{I'}$. Summing Equation (6) for each such I', accumulate the distribution $\bar{s}_I^1 \overset{\text{def}}{=} (\cup_{I'\in\bar{I}_{\bar{s}_I}}\bar{s}_{I'})$, and using Equation (5), the inequality,

$$\sum_{s\in\bar{s}_I^1} r_d\cdot\epsilon\sqrt{(d+1)s} \leq \left(\frac{1}{\alpha-1} + \frac{1}{(\alpha-1)^2}\right)\cdot 2(r_d+1)\cdot\|f_I^\star - f\|_I + \frac{1}{\alpha-1}\cdot r_d\cdot\epsilon\sqrt{(d+1)p_I}. \tag{7}$$

Notice that while both $\bar{s}_I, \bar{s}_I^1 \in \Delta_{\text{bin,n},\geq\bar{q}_I}$, $\bar{s}_I^1$ is at least one notch closer to $\bar{q}_I$ as $\bar{s}_I^1 \in \Delta_{\text{bin,n},<\bar{s}_I}$. Since the number of binary distributions is finite, on recursively using this argument, summation across $\bar{q}_I$ is eventually obtained on the LHS. Iterating on this procedure yields the upper bound

$$\begin{aligned}
\sum_{q\in\bar{q}_{\bar{I}}} r_d\cdot\epsilon\sqrt{(d+1)q} &\leq \left(\frac{1}{\alpha-1} + \frac{1}{(\alpha-1)^2} + \cdots\right)\cdot 2(r_d+1)\cdot\|f_I^\star - f\|_I \\
&\quad + \frac{1}{\alpha-1} r_d\cdot\epsilon\sqrt{(d+1)p_I} \\
&\overset{(a)}{\leq} \frac{2(r_d+1)}{\alpha-2}\cdot\|f_I^\star - f\|_I + \frac{1}{\alpha-1}\cdot r_d\cdot\epsilon\sqrt{(d+1)p_I},
\end{aligned}$$

where (a) follows since $\alpha > 2$. Repeating this argument across each $I \in \bar{I}^2_{\bar{p}}$,

$$\sum_{I\in\bar{I}^2_{\bar{q}}} r_d\cdot\epsilon\sqrt{(d+1)q_I} \leq \frac{2(r_d+1)}{\alpha-2}\cdot\|f^\star_{\bar{I}_{\bar{p}}} - f\|_{\bar{I}^2_{\bar{q}}} + \frac{1}{\alpha-1}\sum_{I\in\bar{I}^2_{\bar{p}}} r_d\cdot\epsilon\sqrt{(d+1)p_I}. \tag{8}$$

This finally gives us

$$
\begin{aligned}
\|\hat{f}_{\bar{I}_{\bar{q}}} - f\|_{\bar{I}^2_{\bar{q}}} &\overset{(a)}{\leq} (r_d+1)\cdot\|f^\star_{\bar{I}_{\bar{q}}} - f\|_{\bar{I}^2_{\bar{q}}} + \sum_{I\in\bar{I}^2_{\bar{q}}} r_d\cdot\epsilon\sqrt{(d+1)q_I} \\
&\overset{(b)}{\leq} (r_d+1)\cdot\|f^\star_{\bar{I}_{\bar{p}}} - f\|_{\bar{I}^2_{\bar{q}}} + \sum_{I\in\bar{I}^2_{\bar{q}}} r_d\cdot\epsilon\sqrt{(d+1)q_I} \\
&\leq (r_d+1)\left(1+\frac{2}{\alpha-2}\right)\|f^\star_{\bar{I}_{\bar{p}}} - f\|_{\bar{I}^2_{\bar{q}}} + \frac{1}{\alpha-1}\sum_{I\in\bar{I}^2_{\bar{p}}} r_d\cdot\epsilon\sqrt{(d+1)p_I},
\end{aligned}
$$

where (a) follows from Theorem 1, (b) follows since, by definition, intervals in $\bar{I}^2_{\bar{q}}$ lie within those in $\bar{I}^2_{\bar{p}}$, and thus $f^\star_{\bar{I}_{\bar{p}}}$ is a coarser approximation to f than $f^\star_{\bar{I}_{\bar{q}}}$ in $\bar{I}^2_{\bar{q}}$, and finally (c) follows by plugging in Equation (8).

Lemma 13. *For the final partition $\bar{I}_{\bar{q}}$ in the run of* MERGE *and any $\bar{p}\in\Delta_{\mathrm{bin,n}}$, let $\bar{I}^3_{\bar{q}}\subseteq\bar{I}_{\bar{q}}$ be intervals that are unions of more than one interval from $\bar{I}_{\bar{p}}$. Let $\bar{I}^3_{\bar{p}}\subseteq\bar{I}_{\bar{p}}$ be the corresponding intervals whose union gives $\bar{I}^3_{\bar{q}}$. Then,*

$$
\|\hat{f}_{\bar{I}_{\bar{q}}} - f\|_{\bar{I}^3_{\bar{q}}} \leq (r_d+1)\cdot\|f^\star_{\bar{I}_{\bar{p}}} - f\|_{\bar{I}^3_{\bar{q}}} + \frac{\alpha\sqrt{2}+\sqrt{2}-1}{\sqrt{2}-1}\sum_{I\in\bar{I}^3_{\bar{p}}} r_d\cdot\epsilon\sqrt{(d+1)p_I}.
$$

Proof Fix an $I\in\bar{I}^1_{\bar{q}}$ and let q_I be its empirical probability. Let $\bar{I}_{\bar{p},I}\in\bar{I}_{\bar{p}}$ indicate intervals under $\bar{I}_{\bar{p}}$ whose union gives I and let $\bar{p}_I\subseteq\bar{p}$ denote the corresponding empirical probabilities under $\bar{I}_{\bar{p},I}$. In run of MERGE, let the interval collection that was merged to create I be denoted by $\bar{I}_{\bar{q},I}$, and its collection of empirical probabilities by $\bar{q}_I$. While $\bar{p}_I$ is a sub-distribution in general, WLOG assume it is a distribution. This also implies $\bar{q}_I$ is a distribution.

Using $\bar{I}_{\bar{q},I}$, separate $\bar{I}_{\bar{p},I}$ into

- $\bar{I}^1_{\bar{p},I}$, consisting of intervals in $\bar{I}_{\bar{p},I}$ that are equal to, or unions of intervals from $\bar{I}_{\bar{q},I}$.
- $\bar{I}^2_{\bar{p},I}$, intervals in $\bar{I}_{\bar{p},I}$ that lie strictly inside some interval in $\bar{I}_{\bar{q},I}$.

Let $\bar{I}^2_{\bar{q},I}\subseteq$ be the corresponding intervals in $\bar{I}_{\bar{q},I}$ that contain $\bar{I}^2_{\bar{p},I}$. Let $\bar{p}^1_I,\bar{p}^2_I$ be empirical probabilities corresponding to $\bar{I}^1_{\bar{p},I}$, $\bar{I}^2_{\bar{p},I}$ respectively. Similarly let $\bar{q}^2_I$ correspond to $\bar{I}^2_{\bar{q},I}$. This is shown in Figure 7, where the arrow indicates the collection of intervals merged by MERGE.

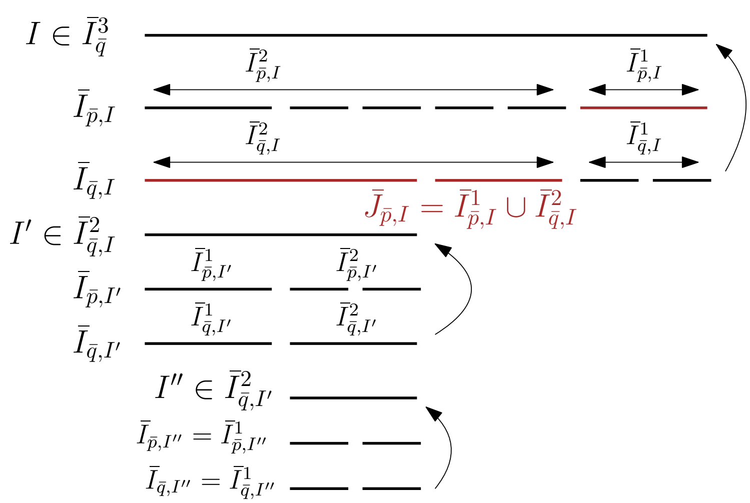


Figure 7: Illustration of proof construction for a particular $I\in\bar{I}^3_{\bar{q}}$.

Modify $\bar{I}_{\bar{p},I}$ to obtain a new partition $\bar{J}_{\bar{p},I} \stackrel{\text{def}}{=} \bar{I}^1_{\bar{p},I} \cup \bar{I}^2_{\bar{q},I}$. Now each interval in $\bar{J}_{\bar{p},I}$ is equal to, or is a union of intervals from $\bar{I}_{\bar{q},I}$. Equivalently, if $\bar{s}$ is the empirical distribution corresponding to $\bar{J}_{\bar{p},I}$, $\bar{s} \in \Delta_{\text{bin,n},\geq \bar{q}_I}$. Since I was merged when the merging routine was called with $\hat{f}_{I,\text{INT}}$, $\bar{I}_{\bar{q},I}$, $\bar{q}_I$, it implies $\lambda_{\bar{s},\gamma} \geq \Lambda_{I_{\bar{s}}}(\hat{f}_{I,\text{INT}})$. Therefore

$$
\begin{aligned}
\|\hat{f}_{\bar{I}_{\bar{q}}} - f\|_I &\leq \|\hat{f}_{\bar{J}_{\bar{p},I}} - f\|_I + \|\hat{f}_{\bar{I}_{\bar{q}}} - \hat{f}_{\bar{J}_{\bar{p},I}}\|_I \\
&\stackrel{(a)}{=} \|\hat{f}_{\bar{J}_{\bar{p},I}} - f\|_I + \Lambda_{I_{\bar{s}}}(\hat{f}_{I,\text{INT}}) \\
&\leq \|\hat{f}_{\bar{J}_{\bar{p},I}} - f\|_I + \lambda_{\bar{s},\gamma} \\
&\stackrel{(b)}{=} \|\hat{f}_{\bar{J}_{\bar{p},I}} - f\|_{\bar{J}_{\bar{p},I}} + \lambda_{\bar{s},\gamma} \\
&= \|\hat{f}_{\bar{J}_{\bar{p},I}} - f\|_{\bar{J}_{\bar{p},I}} + \alpha \sum_{s\in\bar{s}} r_d \cdot \epsilon\sqrt{(d+1)s} \\
&\stackrel{(c)}{=} \|\hat{f}_{\bar{I}^1_{\bar{p},I}} - f\|_{\bar{I}^1_{\bar{p},I}} + \alpha \sum_{p\in\bar{p}^1_I} r_d \cdot \epsilon\sqrt{(d+1)p} \\
&\quad + \|\hat{f}_{\bar{I}^2_{\bar{q},I}} - f\|_{\bar{I}^2_{\bar{q},I}} + \alpha \sum_{q\in\bar{q}^2_I} r_d \cdot \epsilon\sqrt{(d+1)q} \\
&\stackrel{(d)}{=} \|\hat{f}_{\bar{I}_{\bar{p}}} - f\|_{\bar{I}^1_{\bar{p},I}} + \alpha \sum_{p\in\bar{p}^1_I} r_d \cdot \epsilon\sqrt{(d+1)p} \\
&\quad + \|\hat{f}_{\bar{I}^2_{\bar{q},I}} - f\|_{\bar{I}^2_{\bar{q},I}} + \alpha \sum_{q\in\bar{q}^2_I} r_d \cdot \epsilon\sqrt{(d+1)q},
\end{aligned} \tag{9}
$$

where (a) follows since by definition, $\hat{f}_{\bar{I}_{\bar{q}}} = \hat{f}_{I,\text{INT}}$ in interval I, (b) follows since $\bar{J}_{\bar{p},I}$ being a partition of I lies in the same region as I, (c) follows since $\bar{J}_{\bar{p},I} = \bar{I}^1_{\bar{p},I} \cup \bar{I}^2_{\bar{q},I}$, and (d) follows since $\hat{f}_{\bar{I}_{\bar{p}}} = \hat{f}_{\bar{I}^1_{\bar{p},I}}$ in $\bar{I}^1_{\bar{p},I}$ as $\bar{I}^1_{\bar{p},I} \subseteq \bar{I}_{\bar{p}}$.

Now consider an interval $I' \in \bar{I}^2_{\bar{q},I}$. Since $\bar{I}^2_{\bar{q},I} \subseteq \bar{I}_{\bar{q},I}$, and since $\bar{I}_{\bar{q},I}$, by definition, are intervals that were merged to produce I, it follows that I' in turn was an interval that was merged into in some previous step of MERGE. As before, let the intervals that were merged to generate I' be denoted by $\bar{I}_{\bar{q},I'}$. Further, by definition of $\bar{I}^2_{\bar{q},I}$, all intervals in it occur as unions of those in $\bar{I}^2_{\bar{p},I}$, and so does I'. Let $\bar{I}_{\bar{p},I'} \subseteq \bar{I}^2_{\bar{p},I}$ be these intervals whose union gives I'. Repeat the same argument as above to obtain

$$
\begin{aligned}
\|\hat{f}_{\bar{I}^2_{\bar{q}}} - f\|_{I'} \leq \|\hat{f}_{\bar{I}_{\bar{p}}} - f\|_{\bar{I}^1_{\bar{p},I'}} &+ \alpha \sum_{p\in\bar{p}^1_{I'}} r_d \cdot \epsilon\sqrt{(d+1)p} \\
+ \|\hat{f}_{\bar{I}^2_{\bar{q},I'}} - f\|_{\bar{I}^2_{\bar{q},I'}} &+ \alpha \sum_{q\in\bar{q}^2_{I'}} r_d \cdot \epsilon\sqrt{(d+1)q},
\end{aligned} \tag{10}
$$

where each of $\bar{I}^1_{\bar{p},I'}$, $\bar{p}^1_{I'}$, $\bar{I}^2_{\bar{q},I'}$ and $\bar{q}^2_{I'}$ are defined in exactly the same manner as was for I, but by replacing I' in all definitions. Since $\bar{I}^1_{\bar{p},I'} \subseteq \bar{I}^2_{\bar{p},I} \subseteq \bar{I}_{\bar{p}}$, substituting Equation (10) into (9), a larger portion of I is bounded using the difference $\|\hat{f}_{\bar{I}_{\bar{p}}} - f\|$. Upon repeating the same argument for all $\|\hat{f}_{\bar{I}_{\bar{q}}} - f\|$ terms that remain, a bound on the RHS is obtained that consists exclusively of $\|\hat{f}_{\bar{I}_{\bar{p}}} - f\|$. The entire procedure is shown in Figure 7.

Further, from Lemma 14, the sum of all the ϵ-deviation terms that results on the RHS from repeating the argument is bounded by $\sqrt{2}/(\sqrt{2}-1)$ times the total ϵ-deviation in $\bar{I}^3_{\bar{p}}$. This results in

$$
\begin{aligned}
\|\hat{f}_I - f\|_I &\leq \|\hat{f}_{\bar{I}_{\bar{p}}} - f\|_I + \frac{\alpha}{\sqrt{2}-1} \sum_{p\in\bar{p}_I} r_d \cdot \epsilon\sqrt{(d+1)p} \\
&\stackrel{(a)}{\leq} (r_d+1) \cdot \|f^\star_{\bar{I}_{\bar{p}}} - f\|_I + \left(1 + \frac{\alpha\sqrt{2}}{\sqrt{2}-1}\right) \sum_{p\in\bar{p}_I} r_d \cdot \epsilon\sqrt{(d+1)p},
\end{aligned}
$$

where (a) follows from Theorem 1. Repeating across $I \in \bar{I}^3_{\bar{q}}$ gives

$$\|\hat{f}_{\bar{I}_{\bar{q}}} - f\|_{\bar{I}^3_{\bar{q}}} \le (r_d + 1) \cdot \|f^\star_{\bar{I}_{\bar{p}}} - f\|_{\bar{I}^3_{\bar{q}}} + \left(1 + \frac{\alpha\sqrt{2}}{\sqrt{2}-1}\right) \sum_{I \in \bar{I}^3_{\bar{p}}} r_d \cdot \epsilon\sqrt{(d+1)p_I}.$$

Lemma 14. *Suppose in the run of* MERGE*, a collection of consecutive intervals $\bar{I}_1$ was merged in $k-1$ steps to generate $\bar{I}_k$, and suppose $\bar{I}_2, \ldots, \bar{I}_{k-1}$ are the intermediate interval collections. Then,*

$$\sum_{i=1}^{k} \sum_{I \in \bar{I}_i} \sqrt{q_I} \le \sum_{I \in \bar{I}_1} \frac{\sqrt{2}}{\sqrt{2}-1} \sqrt{q_I}.$$

Proof WLOG assume $\bar{q}_{\bar{I}_1}$ is a distribution, which also implies $\bar{q}_{\bar{I}_i}$ is a distribution $\forall i \in \{2, \ldots, k\}$. Notice that for any $i \in \{2, \ldots, k\}$, $\bar{q}_{\bar{I}_i} \in \Delta_{\text{bin,n},\ge \bar{q}_{\bar{I}_{i-1}}}$. Thus $|\bar{I}_i| \le 1/2 \cdot |\bar{I}_{i-1}|$, where $|\bar{I}|$ denotes the number of intervals in $\bar{I}$. By concavity of $\sqrt{x}$ for $x \ge 0$, the sum $\sum_{I \in \bar{I}_i} \sqrt{q_I}$ is maximized for a given $\bar{q}_{\bar{I}_{i-1}}$, if $|\bar{I}_{i-1}| = 2 \cdot |\bar{I}_i|$. Since this equality is attained iff $\bar{q}_{\bar{I}_{i-1}}$ is the uniform distribution over $|\bar{I}_{i-1}|$ elements and $\bar{q}_{\bar{I}_i}$ is uniform over $|\bar{I}_i| = 1/2 \cdot |\bar{I}_{i-1}|$ elements,

$$\sum_{I \in \bar{I}_i} \sqrt{q_I} \le \frac{1}{\sqrt{2}} \sum_{I \in \bar{I}_{i-1}} \sqrt{q_I}.$$

This implies

$$\sum_{i=1}^{k} \sum_{I \in \bar{I}_i} \sqrt{q_I} \le \sum_{i=1}^{k} \left(\frac{1}{\sqrt{2}}\right)^{i-1} \cdot \sum_{I \in \bar{I}_1} \sqrt{q_I} \le \frac{\sqrt{2}}{\sqrt{2}-1} \sum_{I \in \bar{I}_1} \sqrt{q_I}.$$

E Additional Lemmas

Lemma 15. *Adapting [2] to achieve a factor-2 approximation for $\mathcal{P}_2$ results in $\epsilon_n = \tilde{\mathcal{O}}(n^{-1/4})$.*

Proof WLOG fix the interval be $[0, 1]$. Further if $h \in \mathcal{P}_2$ has an ℓ_1 norm > 2, it follows that for any distribution f, $\|h - f\|_1 \ge \|h\|_1 - \|f\|_1 \ge 1$. Thus WLOG restrict the $\mathcal{P}_2$ to the subset $\mathcal{Q} \stackrel{\text{def}}{=} \{h \in \mathcal{P}_2 : \|h\|_1 \le 2, h \ge 0\}$

Let $\mathcal{D}_\epsilon$ be an arbitrary $\epsilon/2$ cover of $\mathcal{Q}$. Thus for the estimate $\hat{f}_{\text{BK}}$ output by [2],

$$\mathbb{E}\|\hat{f}_{\text{BK}} - f\|_1 = 2\text{OPT}_{\mathcal{D}_\epsilon}(f) + \tilde{\mathcal{O}}\left(|\mathcal{D}_\epsilon|^{1/5}/n^{2/5}\right) \le 2\text{OPT}_{\mathcal{P}_2}(f) + \epsilon + \tilde{\mathcal{O}}\left(|\mathcal{D}_\epsilon|^{1/5}/n^{2/5}\right). \quad (11)$$

For $\bar{c} = (c_0, c_1) \in \mathbb{R}^2$, let $h_{\bar{c}} \stackrel{\text{def}}{=} c_0 + c_1 x + c_2 x^2$. Consider $\mathcal{C}_\epsilon$, a subset of $\mathcal{Q}$ defined as $\mathcal{C}_\epsilon \stackrel{\text{def}}{=} \{h_{\bar{c}} \in \mathcal{Q} : c_i = \lambda_i \epsilon,\ \lambda_i \in \mathbb{Z},\ i \in \{0, 1, 2\}\}$. It is easy to see that $|\mathcal{C}_\epsilon| \ge \Omega(1/\epsilon^3)$. Since the ℓ_1 norm between any two members in $\mathcal{C}_\epsilon$ is at least $\epsilon/2$, $|\mathcal{D}_\epsilon| \ge |\mathcal{C}_\epsilon| \ge \Omega(1/\epsilon^3)$.

Optimizing Equation (11) w.r.t. ϵ results in

$$\mathbb{E}\|\hat{f}_{\text{BK}} - f\|_1 \le 2\text{OPT}_{\mathcal{P}_1}(f) + \tilde{\mathcal{O}}(n^{-1/4}).$$

Lemma 16. *Let f be a Gaussian distribution. Then for a constant d, OPT$_{\mathcal{P}_{t,d}}(f) = \mathcal{O}(1/t^{d-1})$.*

Proof Let $t = t_1 - 2$. WLOG assume f has mean 0 and variance 1 so that $f = 1/\sqrt{2\pi} \cdot e^{-x^2/2}$. Fix an $L > 0$. Divide $[-L, L]$ into t equal sized intervals of length $l \stackrel{\text{def}}{=} 2L/t$. Let h_d be the t-piecewise order d Taylor polynomial of f on that interval. Then for any $x \in \mathbb{R}$, $|h_d(x) - f(x)| \le f^{d+1}(c) l^{d+1}/(d+1)!$, where $0 \le c \le l$.

Since $f^{d+1}(x) = H_{d+1}(x) e^{-x^2/2}$ where H_{d+1} is the $d+1$th-order Hermite polynomial, standard bounds [11] imply that there exists a constant $c_d : f^{d+1}(x) \le c_d, \forall x \in \mathbb{R}$.

Observe that on the interval, (L, ∞), from standard sub-gaussian inequalities,

$$\int_L^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \leq e^{-L^2/2}.$$

Similarly for $(-\infty, L)$, extend h_d to these intervals to obtain a $t+2$-piecewise polynomial. Then

$$\|h_d - f\|_1 \leq t \cdot c_d \frac{l^{d+1}}{(d+1)!} + e^{-L^2/2} = c_d \frac{(2L)^{d+1}}{t^d (d+1)!} + e^{-L^2/2}$$

Choosing $L = \mathcal{O}(\sqrt{2d \log t})$ gives $\|h_d - f\|_1 \leq \mathcal{O}\left(\left((\log t)^{\frac{d+1}{2}} + 1\right)/t^d\right) = \mathcal{O}(1/t^{(d-1)})$.

F MERGE and COMP Algorithms

This section provides a detailed description of MERGE and COMP, the main routines of SURF. We also restate the necessary definitions.

F.1 The MERGE Routine

MERGE receives as input, X^{n-1} and parameters d, α, ϵ. The routine operates in $i \in \{1, \ldots, D\}$ steps. Define $D(i) \stackrel{\text{def}}{=} D - i$ and let

$$\bar{u}_i \stackrel{\text{def}}{=} \left(1/2^{D(i)}, \ldots, 1/2^{D(i)}\right),\ \bar{I}_{\bar{u}_i} = (I_{\bar{u}_i,1}, \ldots, I_{\bar{u}_i,2^{D(i)}}).$$

Initialize $\bar{q}_0 \leftarrow (1/n, \ldots, 1/n)$.

Start with $i = 1$ and assign $\bar{s} \leftarrow \bar{q}_{i-1}$. In each step, the routine maintains this $\bar{s} = \bar{q}_{i-1} \in \Delta_{\text{bin,n,}\leq \bar{u}_i}$. This can be seen from the initialization above for $i = 1$ since $\bar{u}_1 = (2/n, \ldots, 2/n)$, and verified for $i > 1$. Thus, using $\bar{I}_{\bar{u}_i}$, we may separate

$$\bar{I}_{\bar{s}} = (\bar{I}_{\bar{s},1}, \ldots, \bar{I}_{\bar{s},2^{D(i)}}),\ \bar{s} = (\bar{s}_1, \ldots, \bar{s}_{2^{D(i)}}),$$

where for each $j \in \{1, \ldots, 2^{D(i)}\}$, $\bar{I}_{\bar{s},j} \subseteq \bar{I}_{\bar{s}}$ are intervals in $\bar{I}_{\bar{s}}$ whose union gives $I_{\bar{u}_i,j} \in \bar{I}_{\bar{u}_i}$. Let $\bar{s}_j \in \bar{s}$ denote the empirical probabilities in $\bar{s}$ corresponding to intervals in $\bar{I}_{\bar{s},j}$. Notice that the sum of all probabilities in $\bar{s}_j$, $\sum_{s \in \bar{s}_j} s = 1/2^{D(i)}$. Therefore the scaled $2^{D(i)} \bar{s}_j$ is an empirical distribution. For brevity, let the polynomial estimate output by INT on $I_{\bar{u}_i,j}$, be denoted by

$$\hat{f}_{I_j} \stackrel{\text{def}}{=} \hat{f}_{I_{\bar{u}_i,j},\text{INT}}.$$

Starting with $j = 1$, invoke COMP with arguments, the polynomial estimate $\hat{f}_{I_j}$, intervals $\bar{I}_{\bar{s},j}$ and the empirical distribution $2^{D(i)} \bar{s}_j$, samples $X_{i,j}^{n-1} \subseteq X^{n-1}$ that lie in $I_{\bar{s},j}$, and parameters d,

$$\gamma \stackrel{\text{def}}{=} \alpha \cdot r_d \cdot \epsilon\sqrt{d+1}.$$

This parameter, γ, is used to tune the bias-variance trade-off. As will be shown subsequently, if $\gamma \to \infty$, $\bar{I}_{\bar{s},j}$ will be merged, resulting in an estimate with a larger bias but smaller variance. A small γ has the opposite effect.

If $\text{COMP}(\hat{f}_{I_j}, \bar{I}_{\bar{s},j}, 2^{D(i)} \bar{s}_j, X_{i,j}^{n-1}, d, \gamma) \leq 0$, merge $\bar{I}_{\bar{s},j}$ into a single interval $I_{\bar{u}_i,j}$. Accomplish this by updating $\bar{s}_j$ to a unitary value, its sum, $(1/2^{D(i)})$. Otherwise, maintain $\bar{s}$ as is. Increment j within the range $\{1, \ldots, 2^{D(i)}\}$ and repeat this procedure.

After the entire run in j is complete, update $\bar{q}_i \leftarrow \bar{s}$. If $D(i) = D - i > 0$, increment i and repeat the same steps. Otherwise, if $D(i) = 0$ or in other words if $i = D$, MERGE, and in turn, SURF outputs the piecewise estimate on $\bar{I}_{\bar{q}_D}$, i.e. $\hat{f}_{\text{SURF}} = \hat{f}_{\bar{I}_{\bar{q}_D},\text{INT}}$.

At each step $i \in \{1, \ldots, D\}$, MERGE calls COMP on $2^{D(i)}$ intervals, each consisting of 2^i samples. Thus each step of MERGE takes $\mathcal{O}(2^{D(i)} \cdot (d^\tau + \log(2^i)) \cdot 2^i) = \mathcal{O}((d^\tau + \log n) 2^D)$ time. The total time complexity is therefore $\mathcal{O}((d^\tau + \log n) 2^D D) = \mathcal{O}((d^\tau + \log n) n \log n)$.

Algorithm 1 MERGE

Input: X^{n-1}, d, α, ϵ
Initialize $D = \log n, \bar{q} = (1/n, \ldots, 1/n)$, $\gamma \leftarrow \alpha \cdot r_d \epsilon \cdot \sqrt{d+1}$
for $i = 1$ **to** D **do**
  $D(i) \leftarrow D - i,\ \bar{s} \leftarrow \bar{q}$
  for $j = 1$ **to** $2^{D(i)}$ **do**
    if $\text{COMP}(\hat{f}_{I_j}, \bar{I}_{\bar{s},j}, 2^{D(i)}\bar{s}_j, X^{n-1}_{i,j}, d, \gamma) \le 0$ **then**
      $\bar{s}_j \leftarrow (1/2^{D(i)})$
    end if
  end for
  $\bar{q} \leftarrow \bar{s}$
end for
Output: $\bar{q}$

F.2 The COMP Routine

COMP receives as input, a function $\hat{f}$, an interval partition $\bar{I} \stackrel{\text{def}}{=} \bar{I}_{\bar{s}}$ and the corresponding empirical distribution $\bar{s}$, samples X^m that lie in $\bar{I}$, and parameters d, γ.

Fix a $\bar{p} \in \Delta_{\text{bin,m},\ge \bar{s}}$, and consider the piecewise polynomial estimate on $\bar{I}_{\bar{p}}$, $\hat{f}_{\bar{I}_{\bar{p}},\text{INT}}$. Define

$$\Lambda_{\bar{I}_{\bar{p}}}(\hat{f}) \stackrel{\text{def}}{=} \|\hat{f}_{\bar{I}_{\bar{p}},\text{INT}} - \hat{f}\|_{\bar{I}_{\bar{p}}},\ \lambda_{\bar{p},\gamma} \stackrel{\text{def}}{=} \sum_{p \in \bar{p}} \gamma\sqrt{p}. \tag{12}$$

$\text{COMP}(\hat{f})$ returns $\mu_{\bar{I}_{\bar{s}},\gamma}(\hat{f})$, the largest difference between $\Lambda_{I_{\bar{p}}}(\hat{f})$ and $\lambda_{\bar{p},\gamma}$ across all $\bar{p} \in \Delta_{\text{bin,m},\ge \bar{s}}$,

$$\mu_{\bar{I}_{\bar{s}},\gamma}(\hat{f}) \stackrel{\text{def}}{=} \max_{\bar{p} \in \Delta_{\text{bin,m},\ge \bar{s}}} \Lambda_{\bar{I}_{\bar{p}}}(\hat{f}) - \lambda_{\bar{p},\gamma}.$$

The quantity, $\Lambda_{\bar{I}_{\bar{p}}}(\hat{f})$ acts as a proxy for the increment in bias that results if the piecewise estimate $\hat{f}_{\bar{I}_{\bar{p}},\text{INT}}$ is merged into $\hat{f}$, while $\lambda_{\bar{p},\gamma}$ accounts for the deviation in $\hat{f}_{\bar{I}_{\bar{p}},\text{INT}}$ under $\mathcal{Q}_\epsilon$. Notice that for any $\bar{p} \in \Delta_{\text{bin,m},\ge \bar{s}}$, $\lambda_{\bar{p},\gamma} \le \lambda_{\bar{s},\gamma}$. Thus $\mu_{\bar{I}_{\bar{s}},\gamma}(\hat{f}) \le 0$ if the decrease in deviation under $\bar{I} = \bar{I}_{\bar{s}}$ is larger than the increased bias under any candidate $\bar{I}_{\bar{p}}$. This in turn signals MERGE to merge $\bar{I}$.

It may be shown that if $\bar{s} = (1/m, \ldots, 1/m)$, the cardinality, $|\Delta_{\text{bin,m},\ge \bar{s}}| = \Omega(m^c)$ for any $c > 0$. Therefore, naively evaluating $\Lambda_{I_{\bar{p}}}(\hat{f}) - \lambda_{\bar{p},\gamma}$ over each $\bar{p} \in \Delta_{\text{bin,m},\ge \bar{s}}$ incurs a worst case time complexity that is super-linear in m. Instead, COMP uses a simple divide-and-conquer procedure that computes $\mu_{\bar{I}_{\bar{s}},\gamma}(\hat{f})$ in time $\mathcal{O}((d^\tau + \log m)m)$.

To describe this, notice that if $\bar{I}_{\bar{s}}$ is a singleton (I), then $\bar{s} = (1)$, implying $\Delta_{\text{bin,m},\ge \bar{s}} = \{(1)\}$. In this case, obtain $\hat{f}_{I,\text{INT}} \in \mathcal{P}_d$ and return

$$\mu_{\bar{I}_{\bar{s}},\gamma}(\hat{f}) = \Lambda_{\bar{I}_{(1)}}(\hat{f}) - \lambda_{(1),\gamma} = \|\hat{f}_{I,\text{INT}} - \hat{f}\|_{\bar{I}_{(1)}} - \gamma\sqrt{1}.$$

If $\bar{I}_{\bar{s}}$ is non singleton or $\bar{s} \ne (1)$, any $\bar{p} \in \Delta_{\text{bin,m},\ge \bar{s}} \setminus \{(1)\}$ may be split into two sub-distributions, $\bar{p}_1$, $\bar{p}_2$ that each sum to $1/2$. For example, if the particular $\bar{p} = (1/4, 1/4, 1/8, 1/8, 1/4)$, it may be split into $\bar{p}_1 = (1/4, 1/4)$ and $\bar{p}_2 = (1/8, 1/8, 1/4)$. The corresponding interval partition is also split into $\bar{I}_{\bar{p}} = (\bar{I}_{\bar{p}_1}, \bar{I}_{\bar{p}_2})$. Since $\bar{s} \ne (1)$, this may also be similarly split into $\bar{s}_1$ and $\bar{s}_2$. As a consequence, $\bar{I}_{\bar{s}}$ is also cleaved into $(\bar{I}_{\bar{s}_1}, \bar{I}_{\bar{s}_2})$ corresponding to $\bar{s}_1$ and $\bar{s}_2$. Using this observation,

$$\begin{aligned}
\max_{\bar{p} \in \Delta_{\text{bin,m},\ge \bar{s}},\ \bar{p} \ne (1)} \Lambda_{\bar{I}_{\bar{p}}}(\hat{f}) - \lambda_{\bar{p},\gamma} &= \max_{\bar{p} \in \Delta_{\text{bin,m},\ge \bar{s}},\ \bar{p} \ne (1)} \Lambda_{\bar{I}_{\bar{p}_1}}(\hat{f}) - \lambda_{\bar{p}_1,\gamma} + \Lambda_{\bar{I}_{\bar{p}_2}}(\hat{f}) - \lambda_{\bar{p}_2,\gamma} \\
&= \max_{\bar{p}_1 \in \Delta_{\text{bin,m/2},\ge 2\bar{s}_1}} \Lambda_{\bar{I}_{\bar{p}_1}}(\hat{f}) - \lambda_{\bar{p}_1,\gamma/\sqrt{2}} \\
&\quad + \max_{\bar{p}_2 \in \Delta_{\text{bin,m/2},\ge 2\bar{s}_2}} \Lambda_{I_{\bar{p}_2}}(\hat{f}) - \lambda_{\bar{p}_2,\gamma/\sqrt{2}} \\
&= \mu_{\bar{I}_{2\bar{s}_1},\gamma/\sqrt{2}}(\hat{f}) + \mu_{\bar{I}_{2\bar{s}_2},\gamma/\sqrt{2}}(\hat{f}),
\end{aligned}$$

Algorithm 2 COMP

Input: $\hat{f}$, $\bar{I}_{\bar{s}}$, $\bar{s}$, X^m, d,γ
$I \leftarrow \cup \bar{I}_{\bar{s}},\ \mu \leftarrow \Lambda_I(\hat{f}) - \lambda_{\bar{s},\gamma}$
if $|\bar{I}| = 1$ **then**
  Return: μ
else
  Return: $\max\{\mu, \text{COMP}(\hat{f}, \bar{I}_{\bar{s}_1}, 2\bar{s}_1, X_1^m, d, \gamma/\sqrt{2})$
    $+ \text{COMP}(\hat{f}, \bar{I}_{\bar{s}_2}, 2\bar{s}_2, X_2^m, d, \gamma/\sqrt{2})\}$
end if

where $2\bar{s}_1$, $2\bar{s}_2$ are the normalized variants of $\bar{s}_1$, $\bar{s}_2$, and γ is scaled by $1/\sqrt{2}$ to accommodate for this scaling. By evaluating $\mu_{\bar{I}_{2\bar{s}_1},\gamma/\sqrt{2}}(\hat{f})$, $\mu_{\bar{I}_{2\bar{s}_1},\gamma/\sqrt{2}}(\hat{f})$ separately, and then comparing their sum with $\Lambda_{\bar{I}_{(1)}}(\hat{f}) - \lambda_{(1),\gamma}$, we allow for a recursive computation of $\mu_{\bar{I}_{\bar{s}},\gamma}(\hat{f})$.

Let X_1^m and X_2^m denote the samples in $\bar{I}_{\bar{s}_1}$ and $\bar{I}_{\bar{s}_2}$ respectively. Using these arguments, call COMP on $\bar{I}_{\bar{s}_1}$, $\bar{s}_1$ and $\bar{I}_{\bar{s}_2}$, $\bar{s}_2$, return the maximum as shown in Algorithm 2.

Now $\Lambda_{\bar{I}_{(1)}}(\hat{f}) - \lambda_{(1),\gamma}$ is calculated by obtaining $\hat{f}_{I,\text{INT}} \in \mathcal{P}_d$ from INT. Since I has m samples, from Theorem 1, this takes $\mathcal{O}(m + d^\tau)$ time. Further, notice that since both $\bar{s}_1$ and $\bar{s}_2$ sum to $1/2$, the split $\bar{I}_{\bar{s}} = (\bar{I}_{\bar{s}_1}, \bar{I}_{\bar{s}_2})$ occurs along the median of X^m. Thus $\bar{I}_{\bar{s}_1}$ and $\bar{I}_{\bar{s}_2}$ has at most half the number of samples, $m/2$, and the time complexity of COMP, $T(m)$, is captured by

$$T(m) \le 2T(m/2) + \mathcal{O}(m + d^\tau),$$

implying $T(m) = \mathcal{O}((d^\tau + \log m)m)$.

F.3 Distributed Computation of COMP and MERGE

We consider the scenario where we are provided with pre-sorted samples, a known t and $\Theta(m)$ memory for some $t \le m \le n$. Let a unit of memory be equivalent to that which is required to store the value of one sample. In this case, we may split the available memory to simulate m concurrent processors with constant processing memory. WLOG let $0 \le t \le n$ and for simplicity, let t, m be a power of 2, just like n. Define $D_t \stackrel{\text{def}}{=} \log_2 t$, $D_m \stackrel{\text{def}}{=} \log_2 m$ and recall that $D \stackrel{\text{def}}{=} \log_2 n$.

Let MERGE_t be the modified MERGE that halts in $D - D_t$ steps instead of D. The corresponding SURF_t outputs the polynomial estimate corresponding to the interval partition given by $\bar{q}_{D-D_t}$ (instead of the one corresponding to $\bar{q}_D$ output by SURF). Let this estimate be denoted by $\hat{f}_{\text{SURF}_t} \stackrel{\text{def}}{=} \hat{f}_{\text{INT},I_{\bar{q}_{D-D_t}}}$ and let $\bar{u}_t \stackrel{\text{def}}{=} (1/2^{D_t}, \ldots, 1/2^{D_t}) = (1/t, \ldots, 1/t)$ be the uniform distribution on t intervals.

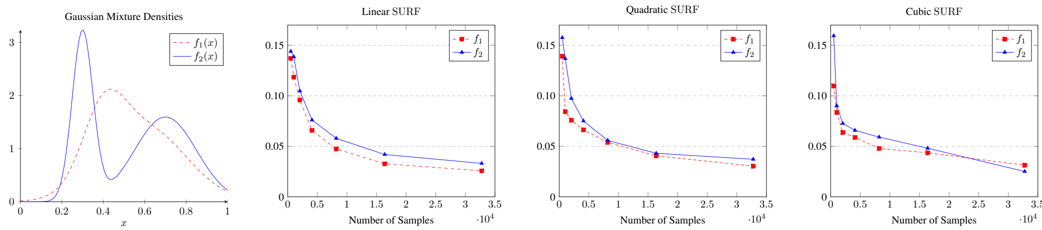
Lemma 17. *Given samples $X^{n-1} \sim f$, for some $t < n$ that are both powers of 2, degree $d \le 8$ and the threshold $\alpha > 2$, SURF_t outputs $\hat{f}_{\text{SURF}_t}$ in time $\mathcal{O}((d^\tau + \log n)n \log n)$ such that under event $\mathcal{Q}_\epsilon$,*

$$\|\hat{f}_{\text{SURF}_t} - f\|_1 \le \min_{\bar{p} \in \Delta_{\text{bin,n}}, \le \bar{u}_t} \sum_{I \in \bar{I}_{\bar{p}}} \left(\frac{(r_d+1)\alpha}{\alpha - 2} \inf_{h \in \mathcal{P}_d} \|h - f\|_I + \frac{r_d(\alpha\sqrt{2} + \sqrt{2} - 1)}{\sqrt{2} - 1} \epsilon \sqrt{(d+1)q_I} \right),$$

where q_I is the empirical mass under interval I, r_d is the constant in Theorem 1.

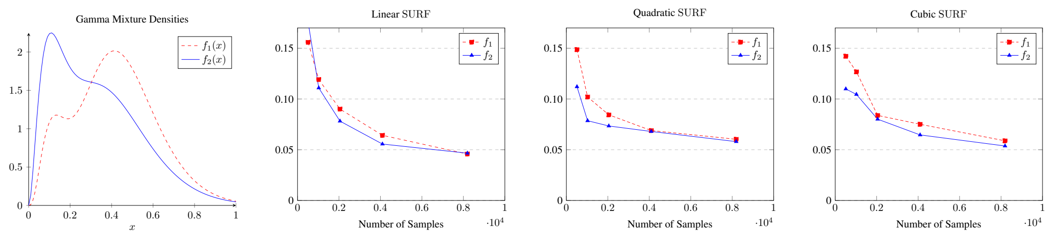
As argued in Theorem 2, Lemma 17 along with Lemma 9 implies that $\hat{f}_{\text{SURF}_t}$ is an r_d-factor approximation for $\mathcal{P}_{t,d}$.

For $1 \le i \le D$, recall that $D(i) \stackrel{\text{def}}{=} D - i$. In step i of MERGE_t, COMP is called on sub-intervals $\bar{I}_{\bar{s},j} \subseteq \bar{I}_{\bar{s}}$ for $j \in \{1, \ldots, 2^{D(i)}\}$, generated by the interval partition $\bar{I}_{(1/2^{D(i)}, \ldots, 1/2^{D(i)})}$. Each $\bar{I}_{\bar{s},j}$, for $j \in \{1, \ldots, 2^{D(i)}\}$ consists of $n/2^{D(i)} = 2^i$ samples.



(a) Densities f_1, f_2. (b) ℓ_1 error with $d = 1$. (c) ℓ_1 error with $d = 2$. (d) ℓ_1 error with $d = 3$.

Figure 8: Evaluation of the estimate output by SURF with degrees $d = 1, 2, 3$, $\alpha = 0.25$, on $f_1 = 0.3\mathcal{N}(0.4, 0.1^2) + 0.7\mathcal{N}(0.6, 0.2^2)$ and $f_2 = 0.4\mathcal{N}(0.3, 0.05^2) + 0.6\mathcal{N}(0.7, 0.15^2)$.



(a) Densities f_1, f_2. (b) ℓ_1 error with $d = 1$. (c) ℓ_1 error with $d = 2$. (d) ℓ_1 error with $d = 3$.

Figure 9: Evaluation of the estimate output by SURF with degrees $d = 1, 2, 3$, $\alpha = 0.25$, on $f_1 = 0.2\text{Gam}(4, 0.04) + 0.8\text{Gam}(8, .06)$ and $f_2 = 0.4\text{Gam}(3, 0.05) + 0.6\text{Gam}(6, .075)$.

Given presorted samples, for the steps i such that $2^{D(i)} \le m$, each call to COMP may be implemented concurrently on the m processors. This results in a time complexity of $\mathcal{O}((d^\tau + \log n)2^i)$ for that step, where $\tau \in [2, 2.4)$ is the matrix inversion constant. As $2^{D(i)} \le m$ implies $D(i) = D - i \le \log_2 m = D_m$, or $i \ge D - D_m$,. The total time taken by these steps is thus given by $\sum_{i=D-D_m}^{D-D_t} \mathcal{O}((d^\tau + \log n)2^i) = \mathcal{O}((d^\tau + \log n)2^{D-D_t}) = \mathcal{O}((d^\tau + \log n)n/t)$.

For steps i in the range $1 \le i < D - D_m$, COMP may be implemented concurrently in batches, with each batch consisting of m sub-intervals among $\bar{I}_{\bar{s},1}, \ldots, \bar{I}_{\bar{s},j}$. As there are a total of $2^{D(i)}/m$ batches, and as each interval consists of 2^i samples, step i takes time $\mathcal{O}((d^\tau + \log n)2^i) \cdot 2^{D(i)}/m$ = $\mathcal{O}((d^\tau + \log n)n/m)$. The total time taken by steps $1 \le i < D - D_m$ is given by $\mathcal{O}((d^\tau + \log n)n/m) \cdot (D - D_m) = \mathcal{O}((d^\tau + \log n)n \log n/m)$.

Thus the time complexity under distributed computation is $\mathcal{O}((d^\tau + \log n)n \max\{1/t, \log n/m\})$.

G Additional Experiments

This section shows additional experiments on the Gaussian and gamma mixtures. Just as in Section 5, SURF is run with $\alpha = 0.25$ and the results are averaged over 10 runs.

Since SURF is invariant to location-scale transformations, WLOG we run experiments on distributions such that essentially all its mass lies in the interval $[0, 1]$. Let $\mathcal{N}(\mu, \sigma^2)$ be the Gaussian distribution with parameters μ, σ. We run SURF with degrees $d = 1, 2, 3$ on the two Gaussian mixtures shown in Figure 8(a). Figures 8(b)–8(d) show the resulting ℓ_1 errors. This is repeated for the gamma mixture density shown in Figure 9, where $\text{Gam}(k, \theta)$ denotes the gamma distribution with shape, scale parameters k, θ respectively. Figures 9(b)–9(d) show the corresponding ℓ_1 errors.

Notice that the errors are similar between distributions, and that the error saturates more quickly for $d = 3$, as the higher degree allows SURF to exploit the smoothness inherent in the considered parametric families. These observations are in line with what was observed for the beta mixtures considered in Figure 2.