arXiv:0712.2869v1 [cs.LG] 18 Dec 2007

Density estimation in linear time

Satyaki Mahalanabis* Daniel Štefankovič*

Abstract

We consider the problem of choosing a density estimate from a set of distributions $\mathcal{F}$, minimizing the L_1-distance to an unknown distribution ([DL01]). Devroye and Lugosi [DL01] analyze two algorithms for the problem: Scheffé tournament winner and minimum distance estimate. The Scheffé tournament estimate requires fewer computations than the minimum distance estimate, but has strictly weaker guarantees than the latter.

We focus on the computational aspect of density estimation. We present two algorithms, both with the same guarantee as the minimum distance estimate. The first one, a modification of the minimum distance estimate, uses the same number (quadratic in $|\mathcal{F}|$) of computations as the Scheffé tournament. The second one, called "efficient minimum loss-weight estimate," uses only a linear number of computations, assuming that $\mathcal{F}$ is preprocessed.

We also give examples showing that the guarantees of the algorithms cannot be improved and explore randomized algorithms for density estimation.

1 Introduction

We study the following density estimation problem considered in [DL96, DL01, DGL02]. There is an unknown distribution g and we are given n (not necessarily independent) samples which define empirical distribution h. Given a finite class $\mathcal{F}$ of distributions, our objective is to output $f \in \mathcal{F}$ such that the error $||f - g||_1$ is minimized. The use of the L_1-norm is well justified by it has many useful properties, for example, scale invariance and the fact that approximate identification of a distribution in the L_1-norm gives an estimate for the probability of every event.

The following two parameters influence the error of a possible estimate: the distance of g from $\mathcal{F}$ and the empirical error. The first parameter is required since we have no control over $\mathcal{F}$, and hence we cannot select a distribution which is better than the "optimal" distribution in $\mathcal{F}$, that is, the one closest to g in L_1-norm. It is not obvious how to define the second parameter—the error of h with respect to g. We follow the definition of [DL01], which is inspired by [Yat85] (see Section 1.1 for a precise definition).

Devroye and Lugosi [DL01] analyze two algorithms in this setting: Scheffé tournament winner and minimum distance estimate. The minimum distance estimate, defined by Yatracos [Yat85], is a special case of the minimum distance principle, formalized by Wolfowitz in [Wol57]. The minimum distance estimate is a helpful tool, for example, it

*Department of Computer Science, University of Rochester, Rochester, NY 14627. Email: {smahalan,stefanko}@cs.rochester.edu

was used by [DL96, DL97] to obtain estimates for the smoothing factor for kernel density estimates and also by [DGL02] for hypothesis testing.

The Scheffé tournament winner algorithm requires fewer computations than the minimum distance estimate, but it has strictly weaker guarantees (in terms of the two parameters mentioned above) than the latter. Our main contribution are two procedures for selecting an estimate from $\mathcal{F}$, both of which have the same guarantees as the minimum distance estimate, but are computationally more efficient. The first has a quadratic (in $|\mathcal{F}|$) cost, matching the cost of the Scheffé tournament winner algorithm. The second one is even faster, using *linearly* many (in $|\mathcal{F}|$) computations (after preprocessing $\mathcal{F}$).

Now we outline the rest of the paper. In Section 1.1 we give the required definitions and introduce the notion of a test-function (a variant of Scheffé set). Then, in Section 1.2, we restate the previous density estimation algorithms (Scheffé tournament winner and the minimum distance estimate) using test-functions. Next, in Section 2, we present our algorithms. The first one is a modification of the minimum-distance estimate with improved (quadratic in $|\mathcal{F}|$) computational cost. The second one, which we call "efficient minimum loss-weight estimate," has only *linear* computational cost after preprocessing $\mathcal{F}$. In Section 3 we explore randomized density estimation algorithms. In the final Section 4, we give examples showing tightness of the theorems stated in the previous sections.

Throughout this paper we focus on the case when $\mathcal{F}$ is finite, in order to compare the computational costs of our estimates to previous ones. However our results generalize in a straightforward way to infinite classes as well if we ignore computational complexity.

1.1 Definitions and Notations

Throughout the paper g will be the unknown distribution and h will be the empirical distribution. Let $\mathcal{F}$ be a set of distributions. We will assume that $\mathcal{F}$ is finite (the results generalize straightforwardly to infinite sets of distributions). Let $\mathrm{d}_1(g, \mathcal{F})$ be the L_1-distance of g from $\mathcal{F}$, that is, $\min_{f \in \mathcal{F}} \|f - g\|_1$.

Given two functions f_i, f_j on Ω (in this context, distributions) we define a *test-function* $T_{ij} : \Omega \to \{-1, 0, 1\}$ to be the function $T_{ij}(x) = \mathrm{sgn}(f_i(x) - f_j(x))$. Note that $T_{ij} = -T_{j\,i}$. We also define $\mathcal{T}_{\mathcal{F}}$ to be the set of all test-functions for $\mathcal{F}$, that is,

$$\mathcal{T}_{\mathcal{F}} = \{\, T_{ij} \mid f_i, f_j \in \mathcal{F} \,\}.$$

Let $\cdot$ be the inner product for the functions on Ω. Note that

$$(f_i - f_j) \cdot T_{ij} = \|f_i - f_j\|_1.$$

We use the inner product of the empirical distribution h with the test-functions to choose an estimate, which is a distribution from $\mathcal{F}$.

In this paper we only consider algorithms which make their decisions purely on inner products of the test-functions with h and members of $\mathcal{F}$. It is reasonable to assume that the computation of the inner product will take significant time. Hence we measure the *computational cost* of an algorithm is by the number of inner products used.

We say that f_i *wins* against f_j if

$$(f_i - h) \cdot T_{ij} < (f_j - h) \cdot T_{j\,i}. \tag{1}$$

Note that either f_i wins against f_j, or f_j wins against f_i, or there is a draw (that is, there is equality in (1)).

The algorithms choose an estimate $f \in \mathcal{F}$ using the empirical distribution h. The L_1-distance of the estimates from the unknown distribution g will depend on the following measure of distance between the empirical and the unknown distribution:

$$\Delta := \max_{T \in \mathcal{T}_{\mathcal{F}}} (g - h) \cdot T. \tag{2}$$

Now we discuss how test-functions can be viewed as a reformulation of Scheffé sets, defined by Devroye and Lugosi [DL01] (inspired by [Sch47] and implicit in [Yat85]), as follows. The Scheffé set of distributions f_i, f_j is

$$A_{ij} = \{x \ ; \ f_i(x) > f_j(x)\}.$$

Devroye and Lugosi say that f_i wins against f_j if

$$\left| \int_{A_{ij}} f_i - h(A_{ij}) \right| < \left| \int_{A_{ij}} f_j - h(A_{ij}) \right|. \tag{3}$$

The advantage of using Scheffé sets is that for a concrete set $\mathcal{F}$ of distributions one can immediately use the theory of Vapnik-Chervonenkis dimension [VČ71] for the family of Scheffé sets of $\mathcal{F}$ (this family is called the *Yatracos class* of $\mathcal{F}$), to obtain a bound on the empirical error.

If h, f_i, f_j are distributions then the condition (1) is *equivalent* to (3) (to see this recall that $T_{ij} = -T_{ji}$, and add $(f_i - h) \cdot \mathbf{1} = (h - f_j) \cdot \mathbf{1}$ to (1), where $\mathbf{1}$ is the vector of all ones). Thus, in our algorithms the test-functions can be replaced by Scheffé sets and VC dimension arguments can be applied.

We chose to use test-functions for two reasons: first, they allow us to give succinct proofs of our theorems (especially Theorem 7), and second, they immediately extend to the case when the members of $\mathcal{F}$ are not distributions (cf, e. g., Exercise 6.2, in [DL01]).

Remark 1. Note that our value of Δ, defined in terms of $\mathcal{T}_{\mathcal{F}}$, is at most twice the Δ used in [DL01], which is defined in terms of Scheffé sets.

1.2 Previous Estimates

In this section we restate the two algorithms for density estimation from Chapter 6 of [DL01]) using test-functions. The first algorithm requires less computation but has worse guarantees than the second algorithm.

Algorithm 1 - SCHEFFÉ TOURNAMENT WINNER.
Output $f \in \mathcal{F}$ with the most wins (tie broken arbitrarily).

Theorem 2 ([DL01], Theorem 6.2)**.** *Let $f_1 \in \mathcal{F}$ be the distribution output by Algorithm 1. Then*

$$\|f_1 - g\|_1 \leq 9 \, \mathrm{d}_1(g, \mathcal{F}) + 8\Delta.$$

The number of inner products used by Algorithm 1 is $\Theta(|\mathcal{F}|^2)$.

Algorithm 2 - MINIMUM DISTANCE ESTIMATE.
Output $f \in \mathcal{F}$ that minimizes

$$\max\{\, |(f-h)\cdot T_{ij}| \;;\; f_i, f_j \in \mathcal{F}\,\}. \tag{4}$$

Theorem 3 ([DL01], Theorem 6.3)**.** *Let f_1 be the distribution output by Algorithm 2. Then*

$$\|f_1 - g\|_1 \le 3\,\mathrm{d}_1(g,\mathcal{F}) + 2\Delta.$$

The number of inner products used by Algorithm 2 is $\Theta(|\mathcal{F}|^3)$.

Let us point out that Theorems 6.2 and 6.3 in [DL01] require that each $f \in \mathcal{F}$ is a distribution, that is, $\int f = 1$. Since we use test-functions in the algorithms instead of Scheffé set based comparisons, the assumption $\int f = 1$ is not actually needed in the proofs of Theorems 6.2 and 6.3 (we skip the proof), and is not used in the proofs of Theorems 4, 7.

2 Our estimators

2.1 A variant of the minimum distance estimate

The following modified minimum distance estimate uses only $O(|\mathcal{F}|^2)$ computations as compared to $O(|\mathcal{F}|^3)$ computations used by Algorithm 2 (equation (5) takes minimum of $O(|\mathcal{F}|)$ terms, whereas equation (4) takes minimum of $O(|\mathcal{F}|^2)$ terms), but as we show in Theorem 4, it gives us the same guarantee as the minimum distance estimate.

Algorithm 3 - MODIFIED MINIMUM DISTANCE ESTIMATE.
Output $f_i \in \mathcal{F}$ that minimizes

$$\max\{\, |(f_i-h)\cdot T_{ij}| \;;\; f_j \in \mathcal{F}\,\}. \tag{5}$$

Theorem 4. *Let $f_1 \in \mathcal{F}$ be the distribution output by Algorithm 3. Then*

$$\|f_1 - g\|_1 \le 3\,\mathrm{d}_1(g,\mathcal{F}) + 2\Delta.$$

The number of inner products used by Algorithm 3 is $\Theta(|\mathcal{F}|^2)$.

Proof :
Let $f_1 \in \mathcal{F}$ be the function output by Algorithm 3. Let $f_2 = \mathrm{argmin}_{f\in\mathcal{F}}\|f-g\|_1$. By the triangle inequality we have

$$\|f_1 - g\|_1 \le \|f_1 - f_2\|_1 + \|f_2 - g\|_1. \tag{6}$$

We bound $\|f_1 - f_2\|_1$ as follows:

$$\begin{aligned}\|f_1 - f_2\|_1 = (f_1-f_2)\cdot T_{12} &\le |(f_1-h)\cdot T_{12}| + |(f_2-h)\cdot T_{12}| \\ &\le |(f_1-h)\cdot T_{12}| + \max_{f_j\in\mathcal{F}} |(f_2-h)\cdot T_{2,j}|,\end{aligned}$$

where in the last inequality we used the fact that $T_{12} = -T_{21}$.

By the criteria of selecting f_1 we have $|(f_1 - h) \cdot T_{12}| \leq \max_{f_j \in \mathcal{F}} |(f_2 - h) \cdot T_{2,j}|$ (since otherwise f_2 would be selected). Hence

$$\begin{aligned}\|f_1 - f_2\|_1 \leq 2 \max_{f_j \in \mathcal{F}} |(f_2 - h) \cdot T_{2,j}| &\leq 2 \max_{f_j \in \mathcal{F}} |(f_2 - g) \cdot T_{2,j}| + 2 \max_{f_j \in \mathcal{F}} |(g - h) \cdot T_{2,j}| \\ &\leq 2\|(f_2 - g)\|_1 + 2 \max_{T \in \mathcal{T}_{\mathcal{F}}} |(g - h) \cdot T| = 2\|f_2 - g\|_1 + 2\Delta.\end{aligned}$$

Combining the last inequality with (6) we obtain

$$\|f_1 - g\|_1 \leq 3\|f_2 - g\|_1 + 2\Delta.$$

■

Remark 5. Note that one can modify the Lemma to only require that g and h be "close" with respect to the test functions for the "best" function in the class, that is, only $|(g - h) \cdot T_{2,j}|$ need to be small (where f_2 is $\operatorname{argmin}_{f \in \mathcal{F}} \|f - g\|_1$).

One can ask whether the observation in Remark 5 can lead to improved density estimation algorithms for concrete sets of distributions. The bounds on Δ (which is given by (2)) are often based on the VC-dimension of the Yatracos class of $\mathcal{F}$. Recall that the Yatracos class Y is the set of $A_{ij} = \{x\,;\, f_i(x) > f_j(x)\}$ for all $f_i, f_j \in \mathcal{F}$. Remark 5 implies that instead of the Yatracos class it is enough to consider the set $Y_i = \{A_{ij}\,;\, f_j \in \mathcal{F}\}$ for $f_i \in \mathcal{F}$. Is it possible that the VC-dimension of each set Y_i is smaller the VC-dimension of the Yatracos class Y? The following (artificial) example shows that this can, indeed, be the case. Let $\Omega = \{0, \ldots, n\}$. For each $(n+1)$-bit binary string $a_0, a_1, \ldots, a_n$, let us consider the distribution

$$P(k) = \frac{1}{4n}(1 + (1/2 - a_0)(1/2 - a_k))2^{-\sum_{j=1}^{n} a_j 2^j},$$

for $k \in \{1, \ldots, n\}$ (with $P(0)$ chosen to make P into a distribution). For this family of 2^{n+1} distributions the VC-dimension of the Yatracos class is n, whereas each Y_i has VC-dimension 1 (since a pair of distributions f_i, f_j has a non-trivial set A_{ij} if and only if their binary strings differ only in the first bit).

2.2 An even more efficient estimator - minimum loss-weight

In this section we present an estimator which, after preprocessing $\mathcal{F}$, uses only $O(|\mathcal{F}|)$ inner products to obtain a density estimate. The guarantees of the estimate are the same as for Algorithms 2 and 3.

The algorithm uses the following quantity to choose the estimate:

$$\text{loss-weight}(f) = \max\{\, \|f - f'\|_1 \;;\; f \text{ does not win against } f' \in \mathcal{F} \,\}.$$

Intuitively a good estimate should have small loss-weight (ideally the loss-weight of the estimate would be $-\infty = \max\{\}$, that is, the estimate would not lose at all). Thus the following algorithm would be a natural candidate for a good density estimator (and, indeed, it has a guarantee matching Algorithms 2 and 3), but, unfortunately, we do not know how to implement it using $O(|\mathcal{F}|)$ inner products.

Algorithm 4a - MINIMUM LOSS-WEIGHT ESTIMATE.
Output $f \in \mathcal{F}$ that minimizes loss-weight(f).

The next algorithm, seems less natural than algorithm 4a, but its condition can be implemented using only $O(|\mathcal{F}|)$ inner products.

Algorithm 4b - EFFICIENT MINIMUM LOSS-WEIGHT ESTIMATE.
Output $f \in \mathcal{F}$ such that for every f' to which f loses we have

$$\|f - f'\|_1 \leq \text{loss-weight}(f'). \tag{7}$$

Before we delve into the proof of (8) let us see how Algorithm 4b can be made to use $|\mathcal{F}| - 1$ inner products. We preprocess $\mathcal{F}$ by computing L_1-distances between all pairs of distributions in $\mathcal{F}$ and store the distances in an list sorted in decreasing order. When the algorithm is presented with the empirical distribution h, all it needs to do is perform comparison between select pairs of distributions. The advantage is that we preprocess $\mathcal{F}$ only once and, for each new empirical distribution we only compute inner products necessary for the comparisons.

We will compute the estimate as follows.

input : family of distributions $\mathcal{F}$, list L of all pairs $\{f_i, f_j\}$ sorted in decreasing order by $\|f_i - f_j\|_1$, oracle for computing inner products $h \cdot T_{ij}$.
output : $f \in \mathcal{F}$ such that: $(\forall f')$ f loses to $f' \implies \|f - f'\|_1 \leq$ loss-weight(f').

```
1 S ← F
2 repeat
3     pick the first edge {f_i, f_j} in L
4     if f_i loses to f_j then f' ← f_i else f' ← f_j fi
5     remove f' from S
6     remove pairs containing f' from L
7 until |S| = 1
8 output the distribution in S
```

Algorithm 4b - using $O(|\mathcal{F}|)$ inner products.

Note that while Algorithm 4b uses only $O(|\mathcal{F}|)$ inner products its running time is actually $\Theta(|\mathcal{F}|^2)$, since it traverses a list of length $\Theta(|\mathcal{F}|^2)$. If we are willing to spend exponential time for the preprocessing then we can build the complete decision tree corresponding to Algorithm 4b and obtain a linear-time density selection procedure. Is it possible to achieve linear running time using only polynomial-time preprocessing?

Question 6 (Tournament Revelation Problem). *We are given a weighted undirected complete graph on n vertices. Assume that the edge-weights are distinct. We preprocess the weighted graph and then play the following game with an adversary until only one vertex remains: we report the edge with the largest weight and the adversary chooses one of the endpoints of the edge and removes it from the graph (together with all the adjacent edges).*

Our goal is to make the computational cost during the game linear-time (in n) in the worst-case (over the adversary's moves). Is it possible to achieve this goal with polynomial-time preprocessing?

We now show that estimate f output by algorithm 4b satisfies (7) for every f' against which f loses. We show, using induction, that the following invariant is always satisfied

on line 2. For any $f \in S$ and any $f' \in \mathcal{F} \setminus S$ we have that if f loses to f' then $\|f - f'\|_1 \leq$ loss-weight(f'). Initially, $\mathcal{F} \setminus S$ is empty and the invariant is trivially true. For the inductive step, let f' be the distribution most recently removed from S. To prove the induction step we only need to show that for every $f \in S$ we have that if f loses to f' then $\|f - f'\|_1 \leq$ loss-weight(f'). Let W be the L_1-distance between two distributions in $S \cup \{f'\}$. Then loss-weight$(f') \geq W$ (since f' lost), and $\|f - f'\|_1 \leq W$ (by the definition of W).

Theorem 7. *Let $f_1 \in \mathcal{F}$ be the distribution output by Algorithm 4a (or Algorithm 4b). Then*

$$\|f_1 - g\|_1 \leq 3\,\mathrm{d}_1(g, \mathcal{F}) + 2\Delta. \tag{8}$$

Assume that we are given L_1-distances between every pair in $\mathcal{F}$. The number of inner products used by Algorithm 4b is $\Theta(|\mathcal{F}|)$.

Proof of Theorem 7:
Let $f_4 = g$. Let f_2 be the function $f \in \mathcal{F}$ minimizing $\|g - f\|_1$. We can reformulate our goal (8) as follows:

$$(f_1 - f_4) \cdot T_{14} \leq 2\Delta + 3(f_2 - f_4) \cdot T_{24}. \tag{9}$$

Let $f_3 \in \mathcal{F}$ be the function $f' \in \mathcal{F}$ such that f_2 loses against f_3 and $\|f_2 - f'\|_1$ is maximal. Note that $f_1, f_2, f_3 \in \mathcal{F}$, but f_4 does need to be in $\mathcal{F}$.

We know that f_2 loses against f_3, that is, we have (see (1))

$$2h \cdot T_{23} \leq f_2 \cdot T_{23} + f_3 \cdot T_{23}, \tag{10}$$

and, since f_1 minimized the maximum loss, we also have

$$(f_1 - f_2) \cdot T_{12} \leq (f_2 - f_3) \cdot T_{23}. \tag{11}$$

By (2) we have

$$2(f_4 - h) \cdot T_{23} \leq 2\Delta. \tag{12}$$

Adding (10), (11), and (12) we obtain

$$2(f_2 - f_4) \cdot T_{23} + (f_2 - f_1) \cdot T_{12} + 2\Delta \geq 0. \tag{13}$$

Note that for any i, j, k, ℓ we have:

$$(f_i - f_j) \cdot (T_{ij} - T_{k\ell}) \geq 0, \tag{14}$$

since if $f_i(x) > f_j(x)$ then $T_{ij} - T_{k\ell} \geq 0$, if $f_i(x) < f_j(x)$ then $T_{ij} - T_{k\ell} \leq 0$, and if $f_i(x) = f_j(x)$ then the contribution of that x is zero. By applying (14) four times we obtain

$$(f_2 - f_4) \cdot (3T_{24} - 2T_{23} - T_{14}) + (f_1 - f_2) \cdot (T_{12} - T_{14}) \geq 0. \tag{15}$$

Finally, adding (13) and (15) yields (9). ■

Remark 8. Note that Remark 5 also applies to Algorithms 4a and 4b, since (12) is the only inequality in which Δ is used.

Remark 9. If the condition (7) of Algorithm 4b is relaxed to

$$\|f - f'\|_1 \leq C \cdot \text{loss-weight}(f'), \tag{16}$$

for some $C \geq 1$, one can prove an analogue of Theorem 7 with (8) replaced by

$$\|f_1 - g\|_1 \leq (1 + 2C)\,\mathrm{d}_1(g, \mathcal{F}) + 2C\Delta. \tag{17}$$

3 Randomized algorithm and mixtures

In this section we explore the following question: can constant 3 be improved if we allow randomized algorithms? Let f be the output of a randomized algorithm (f is a random variable with values in $\mathcal{F}$). We would like to bound the expected error $\mathrm{E}\big[\|f-g\|_1\big]$.

If instead of randomization we consider algorithms which output mixtures of distributions in $\mathcal{F}$ we obtain a related problem. Indeed, let α be the distribution on $\mathcal{F}$ produced by a randomized algorithm, and let $r=\sum_{s\in\mathcal{F}}\alpha_s s$ be the corresponding mixture. Then, by triangle inequality, we have

$$\|r-g\|_1 \le \mathrm{E}\big[\|f-g\|_1\big].$$

Hence the model in which the output is allowed to be a mixture of distributions in $\mathcal{F}$ is "easier" than the model in which the density selection algorithm is randomized.

We consider here only the special case in which $\mathcal{F}$ has only two distributions f_1, f_2, and give an randomized algorithm with a better guarantee than is possible for deterministic algorithms. Later, in Section 4, we give a matching lower bound in the mixture model.

To simplify the exposition we will, without loss of generality, assume that $\|f_1-f_2\|_1 > 0$. Thus for any h we have $(f_1-h)\cdot T_{12}+(h-f_2)\cdot T_{12}=\|f_1-f_2\|_1>0$.

Algorithm 5 - RANDOMIZED ESTIMATE**.**
Let
$$r=\frac{|(f_1-h)\cdot T_{12}|}{|(f_2-h)\cdot T_{12}|}.$$
With probability $1/(r+1)$ output f_1, otherwise output f_2.

(By convention, if $|(f_2-h)\cdot T_{12}|=0$ then we take $r=\infty$ and output f_2 with probability 1).

Theorem 10. *Let $\mathcal{F}=\{f_1,f_2\}$. Let $f\in\mathcal{F}$ be the distribution output by Algorithm 5. Then*

$$\mathrm{E}\Big[\|f-g\|_1\Big] \le 2\,\mathrm{d}_1(g,\mathcal{F})+\Delta.$$

Proof :
Without loss of generality assume that $f_2=\mathrm{argmin}_{f\in\mathcal{F}}\|f-g\|_1$. First we bound the error of f_1 and later use it to bound the error of f. We have, by triangle inequality,

$$\|f_1-g\|_1 \le \|f_1-f_2\|_1+\|f_2-g\|_1.$$

We can bound $\|f_1-f_2\|_1$ as follows

$$\|f_1-f_2\|_1=(f_1-f_2)\cdot T_{12}\le |(f_1-h)\cdot T_{12}|+|(f_2-h)\cdot T_{12}|$$
$$=(r+1)|(f_2-h)\cdot T_{12}|\le (r+1)|(f_2-g)\cdot T_{12}|+(r+1)|(g-h)\cdot T_{12}|.$$

Thus

$$\|f_1-g\|_1\le (r+2)\|f_2-g\|_1+(r+1)\Delta. \tag{18}$$

Hence

$$\mathrm{E}\Big[\|f-g\|_1\Big]=\frac{1}{r+1}\|f_1-g\|_1+\frac{r}{r+1}\|f_2-g\|_1\le 2\|f_2-g\|_1+\Delta,$$

where in the last inequality we used (18). ■

4 Lower bound examples

In this section we construct an example showing that deterministic distribution selection algorithms based on test-functions cannot improve on the constant 3, that is, Theorems 2, 3, 4, 7 are tight. For algorithms that output mixtures (and hence randomized algorithms) the example yields a lower bound of 2, matching the constant in Theorem 10.

Lemma 11. *For every $\varepsilon' > 0$ there exist distributions f_1, f_2, and $g = h$ such that*

$$\|f_1 - g\|_1 \geq (3 - \varepsilon')\|f_2 - g\|_1,$$

and $f_1 \cdot T_{12} = -f_2 \cdot T_{12}$ and $h \cdot T_{12} = 0$.

Before we prove Lemma 11 let us see how it is applied. Consider the behavior of the algorithm on empirical distribution h for $\mathcal{F} = \{f_1, f_2\}$ and $\mathcal{F}' = \{f'_1, f'_2\}$, where $f'_1 = f_2$ and $f'_2 = f_1$. Note that $T'_{12} = T_{21} = -T_{12}$ and hence

$$f'_1 \cdot T'_{12} = -f'_2 \cdot T'_{12} = f_1 \cdot T_{12} = -f_2 \cdot T_{12}.$$

Moreover, we have $h \cdot T_{12} = h \cdot T'_{12} = 0$. Note that all the test-functions have the same value for $\mathcal{F}$ and $\mathcal{F}'$. Hence a test-function based algorithm either outputs f_1 and f'_1, or it outputs f_2 and $f'_2 = f_1$. In both cases it outputs f_1 for one of the inputs and hence we obtain the following consequence.

Corollary 12. *For any $\varepsilon > 0$ and any deterministic test-function based algorithm there exist an input $\mathcal{F}$ and $h = g$ such that the output f_1 of the algorithm satisfies $\|f_1 - g\|_1 \geq (3 - \varepsilon)\mathrm{d}_1(g, \mathcal{F})$.*

Proof of Lemma 11:
Consider the following probability space consisting of of 4 atomic events A_1, A_2, A_3, A_4:

	A_1	A_2	A_3	A_4
f_1	0	$1/4 + \varepsilon$	$1/2$	$1/4 - \varepsilon$
f_2	$1/2 + \varepsilon$	$1/4 - \varepsilon$	0	$1/4$
$g = h$	$1/2$	$1/2$	0	0
T_{12}	-1	1	1	-1

Note that we have $f_1 \cdot T_{12} = -f_2 \cdot T_{12} = \frac{1}{2} + 2\varepsilon$, and $\|f_1 - g\|_1 = \frac{3}{2} - 2\varepsilon$, $\|f_2 - g\|_1 = \frac{1}{2} + \varepsilon$. The ratio $\|f_1 - g\|_1 / \|f_2 - g\|_1$ gets arbitrarily close to 3 as ε goes to zero. ■

Consider f_1 and f_2 from the proof of Lemma 11. Let $f = \alpha f_1 + (1 - \alpha) f_2$ where $\alpha \geq 1/2$. For $0 < \varepsilon < 1/4$ we have $\|f - g\|_1 = 1/2 + \alpha - 2\varepsilon\alpha \geq 1 - 2\varepsilon$. By symmetry, for one of $\mathcal{F} = \{f_1, f_2\}$ and $\mathcal{F}' = \{f'_1, f'_2\}$ (with $f'_1 = f_2$ and $f'_2 = f_1$), the algorithm outputs $\alpha f_1 + (1 - \alpha) f_2$ with $\alpha \geq 1/2$, and hence we obtain the following.

Corollary 13. *For any $\varepsilon > 0$ and any deterministic test-function based algorithm which outputs a mixture there exist an input $\mathcal{F}$ and $h = g$ such that the output f of the algorithm satisfies $\|f - g\|_1 \geq (2 - \varepsilon)\mathrm{d}_1(g, \mathcal{F})$.*

Thus for two distributions the correct constant is 2 for randomized algorithms using test-functions. For larger families of distributions we do not know what the value of the constant is (we only know that it is from the interval $[2, 3]$).

Question 14. *What is the correct constant for deterministic test-function based algorithm which output a mixture? What is the correct constant for randomized test-function based algorithms?*

Next we construct an example showing that 9 is the right constant for Algorithm 1.

Lemma 15. *For every $\varepsilon' > 0$ there exist probability distributions $f_1, f_2, f_3 = f_3'$ and g such that*

$$\|f_1 - g\|_1 \geq (9 - \varepsilon')\|f_2 - g\|_1,$$

yet the Algorithm 1, for $\mathcal{F} = \{f_1, f_2, f_3, f_3'\}$, even when given the true distribution (that is, $h = g$) outputs f_1.

Proof :
Consider the following probability space with 6 events $A_1, \ldots, A_6$ and f_1, f_2 and g with the probabilities given by the following table:

	A_1	A_2	A_3	A_4	A_5	A_6
$g = h$	$2/3 - 21\varepsilon$	$1/9 - 2\varepsilon$	9ε	0	$2/9 + 14\varepsilon$	0
f_1	0	18ε	$2/3 - 12\varepsilon$	$2/9 - 13\varepsilon$	9ε	$1/9 - 2\varepsilon$
f_2	$2/3 - 30\varepsilon$	0	0	0	$2/9 + 14\varepsilon$	$1/9 + 16\varepsilon$
f_3	$2/3 - 21\varepsilon$	9ε	9ε	$2/9 - 4\varepsilon$	0	$1/9 + 7\varepsilon$
T_{12}	-1	1	1	1	-1	-1
T_{13}	-1	1	1	-1	1	-1
T_{23}	-1	-1	-1	-1	1	1

Note that we have

$$\begin{aligned} f_1 \cdot T_{12} &= 7/9 - 14\varepsilon, \quad h \cdot T_{12} = -7/9 + 14\varepsilon, \quad f_2 \cdot T_{12} = -1, \\ f_1 \cdot T_{13} &= 1/3 + 30x, \quad h \cdot T_{13} = -1/3 + 42x, \quad f_3 \cdot T_{13} = -1 + 36x, \\ f_2 \cdot T_{23} &= -1/3 + 60x, \quad h \cdot T_{23} = -5/9 + 28x, \quad f_3 \cdot T_{23} = -7/9 + 14x. \end{aligned}$$

Hence f_1 wins over f_3, f_3 wins over f_2, and f_2 wins over f_1. Since $f_3 = f_3'$ we have that f_1 is the tournament winner. Finally, we have $\|f_1 - g\|_1 = 2 - 72\varepsilon$ and $\|f_2 - g\|_1 = 2/9 + 32\varepsilon$. As $\varepsilon \to 0$ the ratio $\|f_1 - g\|_1 / \|f_2 - g\|_1$ gets arbitrarily close to 9. ■

References

[DGL02] Luc Devroye, László Györfi, and Gábor Lugosi. A note on robust hypothesis testing. *IEEE Transactions on Information Theory*, 48(7):2111–2114, 2002.

[DL96] Luc Devroye and Gábor Lugosi. A universally acceptable smoothing factor for kernel density estimates. *Ann. Statist.*, 24(6):2499–2512, 1996.

[DL97] Luc Devroye and Gábor Lugosi. Nonasymptotic universal smoothing factors, kernel complexity and Yatracos classes. *Ann. Statist.*, 25(6):2626–2637, 1997.

[DL01] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation.* Springer Series in Statistics. Springer-Verlag, New York, 2001.

[Sch47] Henry Scheffé. A useful convergence theorem for probability distributions. *Ann. Math. Statistics*, 18:434–438, 1947.

[VČ71] Vladimir N. Vapnik and Alexey J. Červonenkis. The uniform convergence of frequencies of the appearance of events to their probabilities. *Teor. Verojatnost. i Primenen.*, 16:264–279, 1971.

[Wol57] Jacob Wolfowitz. The minimum distance method. *The Annals of Mathematical Statistics*, 28:75–88, 1957.

[Yat85] Yannis G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *Ann. Statist.*, 13(2):768–774, 1985.