Introduction

Consider the following fundamental statistical task:

> Given independent draws from an unknown probability distribution, what is the minimum sample size needed to obtain an accurate estimate of the distrubution?

k-histograms Given a sample from a pdf p, the method partitions the domain into a number of intervals $B_1, \cdots B_k$ and outputs the empirical pdf which is constant within each interval. The number and location of those intervals are selected in an ad hoc manner.

> Suppose that there exists a k-histogram that provides an accurate approximation to the unknown target distribution. Could such an approximation be efficiently found?

This paper provided an algorithm that approximates the target distrubution nearly as accurately as the best k-histogram, uses a near-optimal sample size, runs in near-linear time in its sample size, given a bound k on the number of intervals.

Vocabulary

A distrubution learning problem is defined by a class C of distributions over a domain Ω. The algorithm has access to independent draws from an unknow pdf p, and its goal is to output a hypothesis distribution h that is close to the target distribution in TV distance.

In the noiseless setting, we are promised that $p \in \mathcal{C}$ and the goal is to construct a hypothesis $h \in \mathcal{C}$ such that the TV distance $d_{TV}(h, p)$ is at most ϵ, where $\epsilon > 0$ is the accuracy parameter.

The more challenging agnostic setting captures the situation of having arbitrary noise in the data - we do not make any assumptions about the target density p and the goal is to find a hypothesis distribution h that is almost as accurate as the best approximation of p in $\mathcal{C}$ in TV distance - i.e. find h such that $d_{TV}(h, p) \leq \alpha \cdot \mathbf{opt}(p) + \epsilon$, where $\mathbf{opt}(p) = \inf_{q \in \mathcal{C}} d_{TV}(q, p)$. Algorithms like this are called α-agnostic learning algorithms.

A distribution f over a finite interval $I \subseteq \mathbb{R}$ is called k-flat if there exists a partition of I into k intervals $I_1, \cdots, I_k$ such that the pdf constant within each such interval. Obviously we could restrict ourselves to the case $I = [0, 1)$ (because pdf must be of finite measure). Let $\mathcal{C}_k$ be the class of all k-flat distributions over $[0, 1)$, and for a distribution p over $[0, 1)$, denote $\mathbf{opt}_k(p) = \inf_{f \in \mathcal{C}_k} d_{TV}(f, p)$.

Main Results

Theorem. There is an algorithm $\mathtt{A}$ that given as input $\tilde{O}(k/\epsilon^2)$ i.i.d. draws from a target distribution p and a parameter $\epsilon > 0$, runs in time $\tilde{O}(k/\epsilon^2)$, and has the following performance guarantee: If (i) p is $\frac{\epsilon/\log(1/\epsilon)}{384k}$-well-behaved, and (ii) $\mathbf{opt}_k(p) \leq \epsilon$, then with probability at least 19/20, it outputs an $O(k \cdot \log^2(1/\epsilon))$-flat distribution h such that $d_{TV}(p, h) \leq 2 \cdot \mathbf{opt}_k(p) + 3\epsilon$.

Even in the significantly easier case that $p \in \mathcal{C}_k$, it is known that a sample of size $\Omega(k/\epsilon^2)$ is information-theoretically necessary. Therefore this result is near-optimal, and has a near-optimal linear running time (since

each sample has to be accessed at least once).

Preliminaries

By the identity $d_{TV}(p,q) = \frac{1}{2}||p-q||_1$, for convenience we can work with L_1 distance only.

We refer to a non-negative function p over an interval as a *sub-distribution*. Given $\kappa > 0$, we say that a (sub-)distribution p over $[0,1)$ is *κ-well-behaved* if $\sup_{x\in[0,1)} \Pr_{x\sim p}[x] \leq \kappa$, i.e. no individual real value is assigned more than κ probability under p.

Given m independent draws $s_1, \cdots, s_m$ from a distribution p over $[0,1)$, the *empirical distribution* $\hat{p}_m$ over $[0,1)$ is the discrete distribution supported on $\{s_1, \cdots, s_m\}$, with $\Pr_{x\sim\hat{p}_m}[x = z] = |\{j \in [m] | s_j = z\}|/m$.

Let $p : [0,1) \to \mathbb{R}$ be a Lebesgue measurable function. Given a family of subsets $\mathcal{A} \subset \mathcal{L}([0,1))$ (the Lebesgue measurable subsets of $[0,1)$), define $||p||_{\mathcal{A}} = \sup_{A\in\mathcal{A}} |p(A)|$. The *VC dimension* of A is the maximum size of a subset $X \subseteq [0,1)$ that is shattered by $\mathcal{A}$ (a set X is shattered by $\mathcal{A}$ if for every $Y \subseteq X$, some $A \in \mathcal{A}$ satisfies $A \cap X = Y$).

Theorem 3 (VC inequality). *Let $p : I \to \mathbb{R}^+$ be a probability density function over $I \subseteq \mathbb{R}$ and $\hat{p}_m$ be the empirical distribution obtained after drawing m points from p. Let $\mathcal{A} \subset 2^I$ be a family of subsets with VC dimension d. Then $E[||p - \hat{p}_m||_{\mathcal{A}}] \leq O(\sqrt{d/m})$.*

As a basic primitive, given access to a sample drawn from a κ-well-behaved target distribution p over $[0,1)$, we will need to partition $[0,1)$ into $\Theta(1/k)$ intervals each of which has probability $\Theta(\kappa)$ under p. There is a simple algorithm that is able to achieve this:

Lemma 2.1. Given $\kappa \in (0,1)$ and access to points drawn from a $\kappa/64$-well-behaved distribution p over $[0,1)$, the procedure `B` draws $\tilde{O}(1/\kappa)$ points from p, runs in time $\tilde{O}(1/\kappa)$ and with probability at least $99/100$ outputs a partition of $[0,1)$ into $l = \Theta(1/\kappa)$ intervals such that $p(I_j) \in [\kappa/2, 3\kappa]$ for all $1 \leq j \leq l$.

Let r be a distribution over $[0,1)$, ane let $\mathcal{P}$ be a set of disjoint intervals that are contained in $[0,1)$. We say that the $\mathcal{P}$-flattening of r, denoted $(r)^{\mathcal{P}}$, is the sub-distribution defined as

$$r(v) = \begin{cases} r(I)/|I|, & \text{if } v \in I, I \in \mathcal{P} \\ 0, & \text{otherwise} \end{cases}$$

Obviously if $\mathcal{P}$ is a partition of $[0,1)$, then $(r)^{\mathcal{P}}$ is a distribution.

We say that two intervals are *consecutive* if $I = [a,b)$ and $I' = [b,c)$. Denote $\alpha_r(I,I') = \int_{I\cup I'} |(r)^{\{I,I'\}} - (r)^{\{I\cup I'\}}| dx$.

The algorithm and analysis

Preparation

First, we give an algorithm for agnositically learning a target distribution p that is nice in two senses: (i) p does not have any heavy atomic elements, and (ii) $\mathrm{opt}_k(p)$ is bounded from above by the error parameter ϵ. Then, we give a general efficient reduction showing how the second assumption can be removed, and lastly we briefly explain how the first assumption can be removed, thus yielding Theorem 1.

The algorithm

Step 1. construct a partition of $[0,1)$ into $z = \Theta(k/\epsilon')$ intervals $I_1, \cdots, I_z$ (where $\epsilon' = \tilde{\Theta}(\epsilon)$) such that p has weight $\Theta(\epsilon'/k)$ on each subinterval:

> Let $\epsilon' = \epsilon/\log(1/\epsilon)$. Run algorithm $\mathtt{B}$ on input parameter to partition $[0,1)$ into $z = \Theta(k/\epsilon')$ intervals $I_1, \cdots, I_z$ such that we have $p(I_i) \in [\epsilon'/12k, \epsilon'/2k]$ with probability at least 99/100.

Step 2. Draw a sample of $\tilde{O}(k/\epsilon^2)$ points from p and uses them to define an empirical distribution $\hat{p}_m$. Then we pretend that the weight $\hat{p}(I)$ assigned by $\hat{p}_m$ is the same as the true weight $p(I)$ (reasonable by Lemma 3.1).

Imagine the target distribution p is actually k-flat, then there are at most k breakpoints, which can be identified by $\alpha_{\hat{p}_m}(I_j, I_{j+1}) > 0$. ($(r)^{\{I \cup I'\}}$ is just flat).

Of course in reality, $\mathsf{opt}_k(p) > 0$. Therefore if we try to use $\alpha_{\hat{p}_m}(I_j, I_{j+1})$ criterion to identify breakpoints there might be miscatch because of the difference between p and q (optimal k-flat approximation). However, recall that we know $\mathsf{opt}_k(p) \le \epsilon$.

Step 3. Set $\mathcal{P}_0 = \{I_1, \cdots, I_z\}$ and $\mathcal{F}_0 = \emptyset$.

Step 4. Set $s = \log_2 \frac{1}{\epsilon'}$. Repeat for $t = 1$ until $t = s$:

- Intitialize $\mathcal{P}_t$ to $\emptyset$ and $\mathcal{F}_t$ to $\mathcal{F}_{t-1}$
- WLOG assume $P_{t-1} = \{I_{t-1,1}, \cdots, I_{t-1,z_{t-1}}\}$. Scan left to right accross the intervals in $\mathcal{P}_{t-1}$. If intervals $I_{t-1,i}, I_{t-1,i+1}$ are (i) both not in $\mathcal{F}_{t-1}$ and (ii) $\alpha_{\hat{p}_m}(I_{t-1,i}, I_{t-1,i+1}) > \epsilon'/(2k)$, then add both of them into $\mathcal{F}_t$.
- Initialize i to 1, and repeatedly execute one of the followig four mutually exclusive and exhaustive cases until $i > z_{t-1}$:
    - $i \le z_{t-1} - 1$ and $I_{t-1,i} = [a,b)$, $I_{t-1,i+1} = [b,c)$ are both consecutive intervals both not in $\mathcal{F}_t$. Add the merged interval into $\mathcal{P}_t$ and set $i = i + 2$
    - $i \le z_{t-1} - 1$ and $I_{t-1,i} \in \mathcal{F}_t$. Set $i = i + 1$.
    - $i \le z_{t-1} - 1$, $I_{t-1,i} \notin \mathcal{F}_t$ and $I_{t-1,i+1} \in \mathcal{F}_t$. Add $I_{t-1,i}$ into F_t and set $i = i + 2$
    - $i = z_{t-1} - 1$. Add $I_{t-1,z_{t-1}}$ into $\mathcal{F}_t$ if $I_{t-1,z_{t-1}}$ is not in $\mathcal{F}_t$ and set $i = i + 1$
    - Set $\mathcal{P}_t = \mathcal{P}_t \cup \mathcal{F}_t$
- Output $(\hat{p}_m)^{\mathcal{P}_s}$

Analysis

The running time is obviously dominated by step 2. We now show correctness.

Lemma 3.1. *With probability 99/100 over the sample drawn in Step 2, for every $0 \le a < b \le z$ we have that $|\hat{p}_m([i_a, i_b)) - p([i_a, i_b))| \le \sqrt{\epsilon'(b-a)} \cdot \epsilon'/(10k)$*

We assume this 99/100 likely event indeed takes place.

Lemma 3.2. *Fix $1 \le t \le s$. Then we have $|\alpha_{\hat{p}_m}(I_{t-1,i}, I_{t-1,i+1}) - \alpha_p(I_{t-1mi}, I_{t-1,i+1})| \le 2\epsilon'/(5k)$.*

Proof. Observe that in iteration t, $I_{t-1,i}$, $I_{t-1,i+1}$ correspond to two unions of consecutive intervals $I_a \cup \cdots \cup I_b$ and $I_{b+1} \cup \cdots \cup I_c$. Because they come from merging two intervals at a time, we have $b-a+1, c-b+1 \leq 2^{t-1} < 2^{s-1} \leq 1/(2\epsilon')$. Therefore by previous lemma,

$$|p(I_{t-1,i}) - \hat{p}_m(I_{t-1,i})| \leq \sqrt{\epsilon' \cdot 2^{s-1}} \cdot \frac{\epsilon'}{10k} \leq \frac{\epsilon'}{10\sqrt{2}k}$$

Let $I = I_{t-1,i}$ and $J = I_{t-1,i+1}$. By definition of α,

$$\begin{aligned}\alpha_p(I,J) &= \left|\frac{p(I)}{|I|} - \frac{p(I)+p(J)}{|I|+|J|}\right| |I| - \left|\frac{p(J)}{|J|} - \frac{p(I)+p(J)}{|I|+|J|}\right| |J| \\ &= \frac{2}{|I|+|J|} \left|p(I)|J| - p(J)|I|\right|\end{aligned}$$

then

$$\begin{aligned}|\alpha_p(I,J) - \alpha_{\hat{p}_m}(I,J)| &= \frac{2}{|I|+|J|} \left||p(I)|J| - p(J)|I|| - |\hat{p}_m(I)|J| - \hat{p}_m(J)|I|\right| \\ &\leq \frac{2}{|I|+|J|} \left(|p(I) - \hat{p}_m(I)||J| + |p(J) - \hat{p}_m(J)||I|\right) \\ &\leq 2\epsilon'/(5k)\end{aligned}$$

For the rest of the analysis, let q denote a fixed k-flat distribution that is closest to p, so $||p-q||_1 = \mathrm{opt}_k(p)$. Let $\mathcal{Q}$ be the partition of $[0,1)$ corresponding to the intervals on which q is piecewise constant. We say a *breakpoint* of $\mathcal{Q}$ is a value in $[0,1]$ that is an endpoint of one of the intervals.

Lemma 3.3. $\mathcal{P}_s$ contains at most $O(k \log^2(1/\epsilon))$ intervals.

Claim 3.4. Let $p, g : I \to \mathbb{R}^{\geq 0}$ be probability distributions over I (so $\int_I p(x)dx = \int_I g(x)dx = 1$), then for every $\alpha > 0$ we have that $||p-q||_1 \leq 2||p-ag||_1$

We now proceed with the proof. We first show that a total of at most $O(k \log(1/\epsilon'))$ intervals are ever added into $\mathcal{F}_t$ across all executions in the loop.

Suppose that the intervals $I_{t-1,i} \cup I_{t-1,i+1}$ are added into $\mathcal{F}_t$ in the second step of the loop. Consider the following cases:

1. $I_{t-1,i} \cup I_{t-1,i+1}$ contains at least one breakpoint of $\mathcal{Q}$. Since $\mathcal{Q}$ has at most k breakpoints, this can happen at most k times in total.
2. $I_{t-1,i} \cup I_{t-1,i+1}$ does not contain any breakpoint of $\mathcal{Q}$. Then $I_{t-1,i} \cup I_{t-1,i+1}$ is a subset of an interval in $\mathcal{Q}$. Recalling that intervals $I_{t-1,i}, I_{t-1,i+1}$ were added into $\mathcal{F}_t$, we know that $\alpha_{\hat{p}_m}(I_{t-1,i}, I_{t-1,i+1}) > \epsilon'/(2k)$. Therefore $\alpha_p(I_{t-1,i}, I_{t-1,i+1}) \geq \epsilon'/5k$ by **Lemma 3.2**. By **Claim 3.4** the contribution to the L_1 distance between p and q on $I_{t-1,i} \cup I_{t-1,i+1}$ is bounded to be at least $\frac{\epsilon'}{10k}$. Since $||p-q||_1 = \mathrm{opt}_k(p)$, there can be most

$$k + O\left(\frac{\mathrm{opt}_k(p) \cdot k}{\epsilon'}\right) = O(k \cdot \log(1/\epsilon))$$

Next, we argue that each $\mathcal{F}_t$ satisfies $|\mathcal{F}_t|$ satisfies $|\mathcal{F}_t| \leq O(k \log^2(1/\epsilon))$. In the third step, for case (4), obviously the number is bounded by $O(\log(1/\epsilon'))$. For case (3), the number is bounded by $O(k \log(1/\epsilon') \log(1/\epsilon'))$. Thus $|\mathcal{F}_s| = O(k \log^2(1/\epsilon))$.

To bound $|\mathcal{P}_t \backslash \mathcal{F}_t|$, obviously $|\mathcal{P}_t \backslash \mathcal{F}_t| < |\mathcal{P}_{t-1} \backslash \mathcal{F}_{t-1}|/2$. Since $|\mathcal{P}_0| = \Theta(k/\epsilon')$, $|\mathcal{P}_s \backslash \mathcal{F}_s| = 1/2^s |\mathcal{P}_0| = O(k)$.

With that the lemma is proved.

Let $\mathcal{P}$ denote any partition of $[0,1)$. We say that partition $\mathcal{P}$ is ϵ'-good for (p,q) if for every breakpoint v of $\mathcal{Q}$, the interval I in $\mathcal{P}$ containing v satisfies $p(I) \leq \epsilon'/(2k)$.

Lemma 3.5. If $\mathcal{P}$ is ϵ'-good for (p,q), then $||p-(p)^{\mathcal{P}}||_1 \leq 2\mathrm{opt}_k(p)+\epsilon'$

Proof. Fix an interval I in $\mathcal{P}$. If there does not exist an interval J in $\mathcal{Q}$ such that $I \subseteq J$, then I must contain a breakpoint of $\mathcal{Q}$, thus $p(I) \leq \epsilon'/(2k)$. Thus

$$\begin{aligned}\int_I |(p)^{\mathcal{P}}(x)-q(x)|dx &\leq \int_I |(p)^{\mathcal{P}}(x)-p(x)|dx + \int_I |p(x)-q(x)|dx \\ &\leq \int_I |p(x)-q(x)|dx + 2p(I)(\epsilon'/k)\end{aligned}$$

Otherwise we have that

$$\int_I |(p)^{\mathcal{P}}(x)-q(x)|dx \leq \int_I |p(x)-q(x)|dx$$

Since there are at most k intervals containing the breakpoints, summing the above inequalities we get the error ϵ'.

Lemma 3.6. There exists a partition $\mathcal{R}$ of $[0,1)$ that is ϵ'-good for (p,q) and satisfies $||(p)^{\mathcal{P}_s}-(p)^{\mathcal{R}}||_1 \leq \epsilon$.

We construct the claimed $\mathcal{R}$ based on $\mathcal{P}_s, \mathcal{P}_{s-1}, \cdots, \mathcal{P}_0$ as follows:

(i) If I is an interval in $\mathcal{P}_s$ not containing a breakpoint of $\mathcal{Q}$, then I is also in $\mathcal{R}$.

(ii) If I is an interval in $\mathcal{P}_s$ that does contain a breakpoint of $\mathcal{Q}$, then we further partition I into a set of intervals S by procedure `Refine-partition(s,I)`.

The procedure: input: (int t, interval J), output: S, a partition of J. Then:

(1) If $t=0$, then output $\{J\}$. (2) If J is an interval in $\mathcal{P}_t$, then, if J contains a breakpoint of $\mathcal{Q}$, then output `RP(t-1,J)`, otherwise output $\{J\}$. (3) Otherwise, J is a union of two intervals in $\mathcal{P}_t$. Output RP(t,J_1)$\cup$RP(t, J_2).

We claim that $|\mathcal{R}|$ is at most $|\mathcal{P}_s| + O(k \cdot \log \frac{1}{\epsilon})$. To see this, note that there are at most k intervals $I \in \mathcal{P}_s$ that contains a breakpoint - and multiply it by 2^s does the trick.

Now consider a fixed breakpoint v of $\mathcal{Q}$. Let $I_{t,v}$ denote the interval containing v in the partition $\mathcal{P}_t$. If $I_{t,v}$ merges with another interval in $\mathcal{P}_t$ in case 1 of step 4c, we denote it with $I'_{t,v}$ and $\alpha_{\hat{p}_m}(I_{t,v}, I'_{t,v}) \leq \epsilon'/2k$. It follow that $\alpha_p(I_{t,v}, I'_{t,v}) \leq 4\epsilon'/5k$. Therefore the contribution to the L_1 distance is at most $\epsilon'/k \cdot \log \frac{1}{\epsilon'} \cdot N$ where N is the number of breakpoints in J. And since after summing over all J, N is just k, we know the L_1 distance between $(p)^{\mathcal{P}_s}$ and $(p)^{\mathcal{R}}$ is at most ϵ.

Finally we give the proof of theorem 4.

By lemma 3.5 applied to $\mathcal{R}$, we have that $||p-(p)^{\mathcal{R}}||_1 \leq 2\mathrm{opt}_k(p)+\epsilon'$. By lemma 3.6, we have that $||(p)^{\mathcal{P}_s}-(p)^{\mathcal{R}}||_1 \leq \epsilon$, therefore $||p-(p)^{\mathcal{P}_s}||_1 \leq 2\mathrm{opt}_k(p)+2\epsilon$. By lemma 3.3 the partition $\mathcal{P}_s$ contains at most $O(k\log^2(1/\epsilon))$ intervals, so both $(p)^{\mathcal{P}_s}$ and $(\hat{p}_m)^{\mathcal{P}_s}$ are $O(k\log^2(1/\epsilon))$-flat distributions. Thus, $||(p)^{\mathcal{P}_s}-(\hat{p}_m)^{\mathcal{P}_s}||_1 = ||(p)^{\mathcal{P}_s}-(\hat{p}_m)^{\mathcal{P}_s}||_{\mathcal{A}_l}$, where l is the number of intervals and $\mathcal{A}_l$ is the family of all subsets of $[0,1)$ that consist of unions of up to l intervals(which has VC dimension $2l$). Consequently by VC inequality, with $m=\tilde{O}(k/\epsilon')$, we have that $E[||(p)^{\mathcal{P}_s}-(\hat{p}_m)^{\mathcal{P}_s}||] \leq 4\epsilon'/100$. By Markov's inequality, we have $||(p)^{\mathcal{P}_s}-(\hat{p}_m)^{\mathcal{P}_s}||_1 \leq \epsilon'$. Thus

$$||p-(\hat{p}_m)^{\mathcal{P}_s}||_1 \leq 2\mathrm{opt}_k(p)+3\epsilon$$