A APPENDIX

A.1 THEOREM 2 : F_H TO f

While estimating true distribution $f(x) : R^d \to R$, the integrated mean square error (IMSE) for the estimator $\hat{f_H}(x)$ using regular histogram with width h and number of samples n, is

$$IMSE(\hat{f_H}) \leq \frac{1}{nh^d} + \frac{R(f)}{n} + o(\frac{1}{n}) + \frac{h^2 d}{4} R(\|\nabla f\|_2)$$

Specifically, its

$$IV = \frac{1}{nh^d} + \frac{R(f)}{n} + o(\frac{1}{n})$$

and

$$ISB \leq \frac{h^2 d}{4} R(\|\nabla f\|_2)$$

where $R(\phi)$ is the roughness of the function ϕ defined as $R(\phi) = \int \phi(x)^2 dx$

Proof. Let $x \in S$. S is the support of the distribution. The estimator $\hat{f_H}(x)$ is defined as, where V(x) is volume of bin in which x lies. Equivalently, we can also use V(b) to denote volume of bin b. For standard histogram, $V(x) = h^d$

$$\hat{f_H}(x) = \frac{1}{nV(bin(x))} \Sigma_{i=1}^{n} \mathcal{I}(x_i \in bin(x)) \tag{1}$$

First let us consider the integrated variance.

$$IV = \int_{x \in S} Var(\hat{f_H}(x)) dx = \Sigma_{b \in bins(S)} \int_{x \in b} Var(\hat{f_H}(x)) dx \tag{2}$$

For a particular bin b, the variance is constant at all values of x. Also for a particular x in bin b, we can write the following for $Var(\hat{f_H}(x))$ using independence of samples.

$$Var(\hat{f_H}(x)) = \frac{1}{nV(bin(x))^2} Var(\mathcal{I}(x_i \in bin(x)) \tag{3}$$

Also $Var(\mathcal{I}(x_i \in b)) = p_b(1 - p_b)$ where p_b is the probability of x_i lying in bin b. That is, $p_b = \int_{x \in b} f(x) dx$

Using this in equation 2

$$IV = \Sigma_{b \in bins(S)} V(b) \frac{1}{nV(b)^2} p_b * (1 - p_b) \tag{4}$$

Simplifying,

$$IV = \Sigma_{b \in bins(S)} \frac{1}{nV(b)} p_b * (1 - p_b) \tag{5}$$

For standard histogram V(b) is same across bins,

$$IV = \frac{1}{nV(b)} (\Sigma_{b \in bins(S)} p_b - \Sigma_{b \in bins(S)} p_b^2) = \frac{1}{nV(b)} (1 - \Sigma_{b \in bins(S)} p_b^2) \tag{6}$$

Using mean value theorem, we can write, $p_b = V(b) f(\xi_b)$ for some point $\xi_b \in b$.

$$\Sigma_{b \in bins} p_b^2 = \Sigma_{b \in bins} V(b)^2 f(\xi_b)^2 = V(b) \Sigma_{b \in bins} V(b) f(\xi_b)^2 \tag{7}$$

Using Rieman Integral approximation , we can write the following as the bin size reduces,

$$\Sigma_{b \in bins} V(b) f(\xi_b)^2 = \int_{x \in S} f^2(x) dx + o(1) \tag{8}$$

$\int_{x \in S} f^2(x) dx$ is also known as the roughness of the function. Let us denote it using R(f). Hence

$$IV = \frac{1}{nV(b)} (1 - V(b)(R(f) + o(1))) \tag{9}$$

$$IV = \frac{1}{nV(b)} - \frac{R(f)}{n} + o(\frac{1}{n}))) \tag{10}$$

Putting $V(b) = h^d$

$$IV = \frac{1}{nh^d} - \frac{R(f)}{n} + o(\frac{1}{n}))) \tag{11}$$

Keeping only the leading term in the above expression,

$$IV = O(\frac{1}{nh^d}) \tag{12}$$

Now let us look at the ISB for this estimator, $ISB(\hat{f_h}(x))$

$$ISB(\hat{f_h}(x)) = \int_{x \in S} (E(\hat{f_H}(x) - f(x)))^2 dx \tag{13}$$

Let us look at the estimator,

$$\hat{f_H}(x) = \frac{1}{V(bin(x))} \int_{t \in bin(x)} f(t)dt \tag{14}$$

Just to make it clear, $x \in R^d$, we will use it as a vector in the following. Using 2nd order multivariate taylor series expansion of this $f(t)$ around x, we get :

$$f(t) = f(x) + \langle t - x, \nabla f(x) \rangle + \frac{1}{2}(t-x)^\top \mathcal{H}(f(x))(t-x) \tag{15}$$

Here $\mathcal{H}(f(t))$ is the hessian of f at t. Without loss of generality let us look at the $bin(x) = [0, h]^d$ that is the bin at the origin. Let us say it is bin_0

$$\int_{t \in bin_0} f(t)dt = f(x)h^d + h^d \langle (\frac{h}{2} - x, \nabla f(x) \rangle + O(h^{d+2}) \tag{16}$$

where $x^{(j)}$ is the j^{th} component of x. Using eq 17 in eq 15, we get

$$\hat{f_H}(x) = f(x) + \langle (\frac{h}{2} - x), \nabla f(x) \rangle + O(h^2) \tag{17}$$

Hence, just keeping the leading term , we have

$$Bias(\hat{f_H}(x)) = \langle (\frac{h}{2} - x), \nabla f(x) \rangle \tag{18}$$

Now,

$$\int_{x \in b_0} Bias(\hat{f_H}(x))^2 dx = \int_{x \in b_0} (\langle (\frac{h}{2} - x), \nabla f(x) \rangle)^2 dx \tag{19}$$

Using cauchy's inequality, we get

$$\int_{x \in b_0} Bias(\hat{f_H}(x))^2 dx \leq \int_{x \in b_0} \|(\frac{h}{2} - x)\|_2^2 \|\nabla f(x)\|_2^2 dx \tag{20}$$

As $[h/2, h/2, ...h/2]$ is a mid point of the bin. The max norm of $x - h/2$ can be $h\sqrt{d}/2$

$$\int_{x \in b_0} Bias(\hat{f_H}(x))^2 dx \leq \frac{h^2 d}{4} \int_{x \in b_0} \|\nabla f(x)\|_2^2 dx \tag{21}$$

Now looking at ISB

$$ISB = \Sigma_{b \in bins} \int_{x \in b_0} Bias(\hat{f_H}(x))^2 dx \leq \frac{h^2 d}{4} \int_{x \in S} \|\nabla f(x)\|_2^2 dx \tag{22}$$

$$ISB \leq \frac{h^2 d}{4} R(\|\nabla f\|_2) \tag{23}$$

A.2 THEOREM 3: F_C TO f_H

While estimating true distribution $f(x) : R^d \rightarrow R$, the integrated mean square error (IMSE) for the estimator $\hat{f_C}(x)$ using regular histogram with width h and number of samples n and countsketch with parameters (R:range, K:repetitions) and average-recovery, is

$$IMSE(\hat{f_C}) = IMSE(\hat{f_H}) + \frac{\#bins}{KRnh^d}$$

where n_{nzp} is the number of non-zero partitions. Specifically, its

$$IV(\hat{f_C}) = IV(\hat{f_H}) + \frac{\#bins - 1}{KRnh^d}$$

and

$$ISB(\hat{f_C}) = ISB(\hat{f_H})$$

where n_nzp is the number of non-zero bins/partitions. □

Proof. Consider a Countsketch with range = R and just one repetition. Let it be parameterized by the randomly drawn hash functions $g : bin \longrightarrow \{0, 1, 2, ..., R-1\}$ and $s : bin \longrightarrow \{-1, +1\}$. The estimate of density at point x can then be written as

$$\hat{f_C}(x) = \frac{1}{nV(bin(x))}(c(bin(x)) + \Sigma_{i=1}^{n}\mathcal{I}(x_i \notin bin(x) \wedge g(bin(x_i)) == g(bin(x)))s(bin(x_i))s(bin(x)) \tag{24}$$

We can rewrite this as ,

$$\hat{f_C}(x) = \hat{f_H}(x) + \frac{1}{nV(bin(x))}(\Sigma_{i=1}^{n}\mathcal{I}(x_i \notin bin(x) \wedge g(bin(x_i)) == g(bin(x)))s(bin(x_i))s(bin(x)) \tag{25}$$

where c(.) is count and V(.) is volume of the bins. As E(s(b)) = 0, it can be clearly seen that.

$$E(\hat{f_C}(x)) = E(\hat{f_H}(x)) \tag{26}$$

Hence, it follows that

$$ISB(\hat{f_C}(x)) = ISB(\hat{f_H}(x)) \tag{27}$$

It can be checked that each of the terms in the summation for right hand side of equation 26 including the terms in $\hat{f_H}(x)$ are independent to each other . i.e. covariance between them is 0. Hence we can write the variance of our estimator as,

$$Var(\hat{f_C}(x)) = Var(\hat{f_H}(x)) + \frac{1}{nV^2(bin(x))}Var(\mathcal{I}(x_i \notin bin(x) \wedge g(bin(x_i)) == g(bin(x)))s(bin(x_i))s(bin(x))) \tag{28}$$

$$Var(\hat{f_C}(x)) = Var(\hat{f_H}(x)) + \frac{1}{nV^2(bin(x))}E(\mathcal{I}(x_i \notin bin(x) \wedge g(bin(x_i)) == g(bin(x))))^2 \tag{29}$$

$$Var(\hat{f_C}(x)) = Var(\hat{f_H}(x)) + \frac{1}{nV^2(bin(x))}(1 - p_{bin(x)})\frac{1}{R}) \tag{30}$$

Hence, IV is

$$IV(\hat{f_C}(x)) = IV(\hat{f_H}(x)) + \int_{x \in S}\frac{1}{nV^2(bin(x))}(1 - p_{bin(x)})\frac{1}{R}) \tag{31}$$

$$IV(\hat{f_C}(x)) = IV(\hat{f_H}(x)) + \Sigma_{b \in bins}\int_{x \in b}\frac{1}{nV^2(b)}(1 - p_b)\frac{1}{R}) \tag{32}$$

$$IV(\hat{f_C}(x)) = IV(\hat{f_H}(x)) + \Sigma_{b \in bins}\frac{1}{nV(b)}(1 - p_b)\frac{1}{R}) \tag{33}$$

Assuming standard partitions. $V(b) = h^d$ for all b

$$IV(\hat{f_C}(x)) = IV(\hat{f_H}(x)) + \frac{1}{nh^d}\frac{(\#bins - 1)}{R} \tag{34}$$

With average recovery, with K repetitions, the analysis can be easily extended to get IV as

$$IV(\hat{f_C}(x)) = IV(\hat{f_h}(x)) + \frac{1}{nh^d}\frac{(\#bins - 1)}{KR} \tag{35}$$

The ISB remains same in this case. □

A.3 THEOREM 4: $f_C^* to f_C$

While estimating true distribution $f(x) : R^d \to R$, the integrated mean square error

(IMSE) for the estimator $\hat{f_C^*}(x)$ using regular histogram with width h and number of samples n and countsketch with parameters (R:range, K:repetitions), is related to the estimator $\hat{f_C}(x)$ as follows

$$IMSE(\hat{f_C}(x)) - \epsilon(N + 2M) \leq IMSE(\hat{f_C^*}(x)) \leq IMSE(\hat{f_C}(x)) + \epsilon(N + 2M)$$

Specifically,

$$IV(\hat{f_C}(x)) - 2\epsilon M \leq IV(\hat{f_C^*}(x)) \leq IV(\hat{f_C}(x)) + 2\epsilon M$$

and

$$ISB(\hat{f_C}(x)) - \epsilon N \leq ISB(\hat{f_C^*}(x)) \leq ISB(\hat{f_C}(x)) + \epsilon N$$

where

$$M \leq IV(\hat{f_C}(x)) + 2(R(f) + \frac{h^2 d}{4}R(\|\nabla f\|_2) + h\sqrt{d}\int_{x\in S}(f(x)\|\nabla f\|_2))$$

$$N = (1 + ISB(\hat{f_C}(x)))$$

with probability $(1 - \delta)$ where $\delta = \frac{\#bins}{\epsilon^2 nR}$

Proof. Let us look at the estimator

$$\hat{f_C^*}(x) = \frac{\widehat{c(bin(x))}}{V(bin(x))\widehat{\Sigma_b c(b)}} = \hat{f_C}(x) * \frac{n}{\hat{n}} \tag{36}$$

where $\hat{n} = \widehat{\Sigma_b c(b)}$ and $n = \Sigma_b c(b)$ □

$\hat{\mathbf{n}}$ and its relation to n Let us first analyse $\hat{n}$ and how it is related to n.

$$\hat{n} = \widehat{\Sigma_b c(b)} = \Sigma_b \Sigma_{i=1}^{n} \mathcal{I}(x_i \in b) + \mathcal{I}(x_i \notin b \wedge g(bin(x_i)) == g(b))s(bin(x_i))s(b) \tag{37}$$

$$\hat{n} = \Sigma_{b,i} \mathcal{I}(x_i \in b) + \mathcal{I}(x_i \notin b \wedge g(bin(x_i)) == g(b))s(bin(x_i))s(b) \tag{38}$$

Note that $E(\hat{n}) = n$. For varaince, observe that most of the terms in the summation have covariance 0, except the terms $Cov(\mathcal{I}(x_i \in b_1), \mathcal{I}(x_i \in b_2))$ which are negatively correlated. Hence

$$\begin{aligned} Var(\hat{n}) =& \Sigma_{b,i} Var(\mathcal{I}(x_i \in b)) + Var(\mathcal{I}(x_i \notin b \wedge g(bin(x_i))! = g(b))s(bin(x_i))s(b)) + \\ & 2\Sigma_{i,b_1,b_2,b_1\neq b_2} Cov(\mathcal{I}(x_i \in b_1), \mathcal{I}(x_i \in b_2)) \end{aligned} \tag{39}$$

We know that

$Var(\mathcal{I}(x_i \in b)) = p_b(1 - p_b)$

$Var(\mathcal{I}(x_i \notin b \wedge g(bin(x_i)) == g(b))s(bin(x_i))s(b)) = E(\mathcal{I}(x_i \notin b \wedge g(bin(x_i))! = g(b))^2) = \frac{1 - p_b}{R}$

$Cov(\mathcal{I}(x_i \in b_1), \mathcal{I}(x_i \in b_2)) = -p_{b_1} p_{b_2}$

Hence, we pluggin in the values in previous equation ,

$$Var(\hat{n}) = n\Sigma_b p_b(1 - p_b) + n\Sigma_b \frac{1 - p_b}{R} - 2n\Sigma_{b_1,b_2,b_1\neq b_2} p_{b_1} p_{b_2} \tag{40}$$

$$Var(\hat{n}) = n(1 - \Sigma_b p_b^2) + n\Sigma_b \frac{1 - p_b}{R} - 2n\Sigma_{b_1,b_2} p_{b_1} p_{b_2} \tag{41}$$

$$Var(\hat{n}) = n\{(1 + \Sigma_b \frac{1 - p_b}{R} - (\Sigma_b p_b^2) - 2n\Sigma_{b_1,b_2} p_{b_1} p_{b_2})\} \tag{42}$$

$$Var(\hat{n}) = n\{(1 + \Sigma_b \frac{1 - p_b}{R} - (\Sigma_b p_b)^2\} \tag{43}$$

$$Var(\hat{n}) = n\{\Sigma_b \frac{1 - p_b}{R}\} \tag{44}$$

$$Var(\hat{n}) = \frac{n(\#bins - 1)}{R} < \frac{n(\#bins)}{R} \tag{45}$$

Using Chebyshev's inequality , we have

$$P(|\hat{n} - n| > \epsilon n) \leq \frac{Var(\hat{n})}{\epsilon^2 n^2} \tag{46}$$

$$P(|\hat{n} - n| > \epsilon n) \leq \frac{\#bins}{\epsilon^2 nR} \tag{47}$$

Hence with probability $(1 - \delta)$, $\delta = \frac{\#bins}{\epsilon^2 nR}$, $\hat{n}$ is within ϵ multiplicative error.

relation of pointwise Bias and ISB With probability $1 - \delta$,

$$\frac{\hat{f_C}(x)}{1 + \epsilon} \leq \hat{f_C^*}(x) \leq \frac{\hat{f_C}(x)}{1 - \epsilon} \tag{48}$$

As expectations respect inequalities

$$\frac{E(\hat{f_C}(x))}{1 + \epsilon} \leq E(\hat{f_C^*}(x)) \leq \frac{E(\hat{f_C}(x))}{1 - \epsilon} \tag{49}$$

$$\frac{E(\hat{f_C}(x))}{1 + \epsilon} - f(x) \leq Bias(\hat{f_C^*}(x)) \leq \frac{E(\hat{f_C}(x))}{1 - \epsilon} - f(x) \tag{50}$$

$$\frac{Bias(\hat{f_C}(x)) - \epsilon f(x)}{1 + \epsilon} \leq Bias(\hat{f_C^*}(x)) \leq \frac{Bias(\hat{f_C}(x)) + \epsilon f(x)}{1 - \epsilon} \tag{51}$$

$$\frac{Bias(\hat{f_C}(x)) - \epsilon f(x)}{1 + \epsilon} \leq Bias(\hat{f_C^*}(x)) \leq \frac{Bias(\hat{f_C}(x)) + \epsilon f(x)}{1 - \epsilon} \tag{52}$$

Integrating expressions again respects inequalities

$$\frac{ISB(\hat{f_C}(x)) - \epsilon \int f(x)}{1 + \epsilon} \leq ISB(\hat{f_C^*}(x)) \leq \frac{ISB(\hat{f_C}(x)) + \epsilon \int f(x)}{1 - \epsilon} \tag{53}$$

$$\frac{ISB(\hat{f_C}(x)) - \epsilon}{1 + \epsilon} \leq ISB(\hat{f_C^*}(x)) \leq \frac{ISB(\hat{f_C}(x)) + \epsilon}{1 - \epsilon} \tag{54}$$

Using first order taylor expansion of $\frac{1}{1+\epsilon}$ and ignore square terms

$$(1 - \epsilon)ISB(\hat{f_C}(x)) - \epsilon \leq ISB(\hat{f_C^*}(x)) \leq (1 + \epsilon)ISB(\hat{f_C}(x)) + \epsilon \tag{55}$$

$$ISB(\hat{f_C}(x)) - \epsilon(1 + ISB(\hat{f_C}(x))) \leq ISB(\hat{f_C^*}(x)) \leq ISB(\hat{f_C}(x)) + \epsilon(1 + ISB(\hat{f_C}(x))) \tag{56}$$

Hence,

$$ISB(\hat{f_C}(x)) - \epsilon N \leq ISB(\hat{f_C^*}(x)) \leq ISB(\hat{f_C}(x)) + \epsilon N \tag{57}$$

where

$$N = (1 + ISB(\hat{f_C}(x)))$$

Point wise variance and IV Using the similar arguments

$$\frac{E(\hat{f_C}^2(x))}{(1+\epsilon)^2} - \frac{E^2(\hat{f_C}(x))}{(1-\epsilon)^2} \leq Var(\hat{f_C^*}(x)) \leq \frac{E(\hat{f_C}^2(x))}{(1-\epsilon)^2} - \frac{E^2(\hat{f_C}(x))}{(1+\epsilon)^2} \tag{58}$$

Again making first order taylor expansions of denominator and ignoring square terms

$$Var(\hat{f_C}(x)) - 2\epsilon(E(\hat{f_C}^2(x) + E^2(\hat{f_C}(x))) \leq Var(\hat{f_C^*}(x)) \leq Var(\hat{f_C}(x)) + 2(E(\hat{f_C}^2(x) + E^2(\hat{f_C}(x))) \tag{59}$$

Since, $Var(\hat{f_C}(x)) = E(\hat{f_C}^2(x)) - E^2(\hat{f_C}(x))$

$$Var(\hat{f_C}(x)) - 2\epsilon(Var(\hat{f_C}(x)) + 2E^2(\hat{f_C}(x))) \leq Var(\hat{f_C^*}(x)) \leq Var(\hat{f_C}(x)) + 2\epsilon(Var(\hat{f_C}(x)) + 2E^2(\hat{f_C}(x))) \tag{60}$$

$$IV(\hat{f_C}(x)) - 2\epsilon(IV(\hat{f_C}(x)) + 2\int_{x\in S} E^2(\hat{f_C}(x))) \leq IV(\hat{f_C^*}(x)) \leq IV(\hat{f_C}(x)) + 2\epsilon(IV(\hat{f_C}(x)) + 2\int_{x\in S} E^2(\hat{f_C}(x))) \tag{61}$$

Let us now figure out the $\int_{x\in S} E^2(\hat{f_C}(x))$

$$\int_{x\in S} E^2(\hat{f_C}(x)) = \int_{x\in S} E^2(\hat{f_H}(x)) \tag{62}$$

From equation 18, $E(\hat{f_H}(x))^2 = f(x)^2 + (\langle(\frac{h}{2} - x), \nabla f(x)\rangle)^2 + 2f(x)\langle(\frac{h}{2} - x), \nabla f(x)\rangle$

$$\int_{x\in S} E^2(\hat{f_H}(x)) \leq R(f) + \frac{h^2 d}{4} R(\|\nabla f\|_2) + h\sqrt{d}\int_{x\in S}(f(x)\|\nabla f\|_2) \tag{63}$$

Hence,

$$IV(\hat{f_C}(x)) - 2\epsilon M \leq IV(\hat{f_C^*}(x)) \leq IV(\hat{f_C}(x)) + 2\epsilon M \tag{64}$$

Where

$$M \leq IV(\hat{f_C}(x)) + 2(R(f) + \frac{h^2 d}{4} R(\|\nabla f\|_2) + h\sqrt{d}\int_{x\in S}(f(x)\|\nabla f\|_2)) \tag{65}$$

A.4 LEMMA 1

Estimators $\hat{f_S}(x)$ and $\hat{f_C^*}(x)$, obtained from the Density Sketch with parameters(R,K,H) using histogram of width h built over n i.i.d samples drawn from true distribution have a relation

$$\int |\hat{f_C^*}(x) - \hat{f_S}(x)|dx = 2(1 - ratio_h)$$

where $ratio_h$ is the capture ratio as defined in section 3

$$\int |\hat{f_C^*}(x) - \hat{f_S}(x)|dx = \Sigma_{b\in bins}\int_{x\in b} |\hat{f_C^*}(x) - \hat{f_S}(x)|dx \tag{66}$$

$$\int |\hat{f_C^*}(x) - \hat{f_S}(x)|dx = \Sigma_{b\in bins(H)}\int_{x\in b} |\hat{f_C^*}(x) - \hat{f_S}(x)|dx + \Sigma_{b\notin bins(H)}\int_{x\in b} |\hat{f_C^*}(x) - \hat{f_S}(x)|dx \tag{67}$$

we know that for $x \in b, b \notin bins(H)$, $\hat{f_S}(x) = 0$. Hence,

$$\int |\hat{f_C^*}(x) - \hat{f_S}(x)|dx = \Sigma_{b\in bins(H)}\int_{x\in b} |\hat{f_C^*}(x) - \hat{f_S}(x)|dx + \Sigma_{b\notin bins(H)}\int_{x\in b} \hat{f_C^*}(x)dx \tag{68}$$

$\int_{x\in b} \hat{f_C^*}(x)dx$ is the probability of a data point lying in that bucket according to $\hat{f_C^*}(x)$

$$\int |\hat{f_C^*}(x) - \hat{f_S}(x)|dx = \Sigma_{b \in bins(H)} \int_{x \in b} |\hat{f_C^*}(x) - \hat{f_S}(x)|dx + \Sigma_{b \notin bins(H)} \frac{\hat{c_b}}{\hat{n}} \tag{69}$$

For points $x \in b, b \in bins(H)$, $\hat{f_C^*}(x) * \hat{n} = \hat{f_S}(x) * \hat{n_h}$, Hence, $\hat{f_S}(x) = \frac{\hat{n}}{\hat{n_h}} \hat{f_C^*}(x)$

$$\int |\hat{f_C^*}(x) - \hat{f_S}(x)|dx = \Sigma_{b \in bins(H)} \int_{x \in b} \hat{f_C^*}(x)(\frac{\hat{n}}{\hat{n_h}} - 1)dx + \Sigma_{b \notin bins(H)} \frac{\hat{c_b}}{\hat{n}} \tag{70}$$

$$\int |\hat{f_C^*}(x) - \hat{f_S}(x)|dx = \Sigma_{b \in bins(H)} \int_{x \in b} \hat{f_C^*}(x)(\frac{\hat{n}}{\hat{n_h}} - 1)dx + \Sigma_{b \notin bins(H)} \frac{\hat{c_b}}{\hat{n}} \tag{71}$$

$$\int |\hat{f_C^*}(x) - \hat{f_S}(x)|dx = (\frac{\hat{n}}{\hat{n_h}} - 1)\Sigma_{b \in bins(H)} \frac{\hat{c_b}}{\hat{n}} + \Sigma_{b \notin bins(H)} \frac{\hat{c_b}}{\hat{n}} \tag{72}$$

$$\int |\hat{f_C^*}(x) - \hat{f_S}(x)|dx = (\frac{\hat{n}}{\hat{n_h}} - 1)(\frac{\hat{n_h}}{\hat{n}}) + \frac{\hat{n} - \hat{n_h}}{\hat{n}} \tag{73}$$

$$\int |\hat{f_C^*}(x) - \hat{f_S}(x)|dx = (1 - \frac{\hat{n_h}}{\hat{n}}) + \frac{\hat{n} - \hat{n_h}}{\hat{n}} \tag{74}$$

$$\int |\hat{f_C^*}(x) - \hat{f_S}(x)|dx = 2(1 - \frac{\hat{n_h}}{\hat{n}}) \tag{75}$$

$$\int |\hat{f_C^*}(x) - \hat{f_S}(x)|dx = 2(1 - ratio_h) \tag{76}$$

A.5 THEOREM 5

The IMSE of estimator $\hat{f_S}(x)$ obtained from the Density Sketch with parameters(R,K,H) using histogram of width h built over n i.i.d samples drawn from true distribution f(x) is

$$IMSE(\hat{f_S}(x)) \leq 12(1 - ratio_h)^2 + 3IMSE(\hat{f_C^*}(x))$$

where $ratio_h$ is the capture ratio as defined in

Proof. Giving a very loose relation between $\hat{f_S}$ and f. We can write

$$\int (\hat{f_S}(x) - f(x))^2 dx = \int ((\hat{f_S}(x) - \hat{f_C^*}(x)) - (\hat{f_C^*}(x) - f(x)))^2 dx \tag{77}$$

$$\int (\hat{f_S}(x) - f(x))^2 dx \leq 3 \int (\hat{f_S}(x) - \hat{f_C^*}(x))^2 dx + 3 \int (\hat{f_C^*}(x) - f(x))^2 dx \tag{78}$$

$$\int (\hat{f_S}(x) - f(x))^2 dx \leq 3(\int |(\hat{f_S}(x) - \hat{f_C^*}(x))|dx)^2 + 3 \int (\hat{f_C^*}(x) - f(x))^2 dx \tag{79}$$

$$\int (\hat{f_S}(x) - f(x))^2 dx \leq 12(1 - ratio_h)^2 + 3 \int (\hat{f_C^*}(x) - f(x))^2 dx \tag{80}$$

$$IMSE = MISE(\hat{f_S}(x)) \leq 12(1 - ratio_h)^2 + 3IMSE(\hat{f_C^*}(x)) \tag{81}$$

□

B THEOREM 1 (MAIN THEOREM) COMBINES ALL OTHER THEOREMS

This theorem directly relates the distribution $\hat{f_S}(x)$ to the true distribution. f(x)

$$IMSE(\hat{f_S}(x)) \leq 12(1 - ratio_h)^2 + 3IMSE(\hat{f_C^*}(x)) \tag{82}$$

$$IMSE(\hat{f_S}(x)) \leq 12(1 - ratio_h)^2 + 3(IMSE(\hat{f_C}(x) + \epsilon(N + 2M))) \tag{83}$$

$$IMSE(\hat{f_S}(x)) \leq 12(1-ratio_h)^2 + 3(IMSE(\hat{f_H}) + \frac{\#bins-1}{KRnh^d} + \epsilon(N+2M))) \tag{84}$$

$$IMSE(\hat{f_S}(x)) \leq 12(1-ratio_h)^2 + 3(\frac{1}{nh^d} + \frac{R(f)}{n} + o(\frac{1}{n}) + \frac{h^2d}{4}R(\|\nabla f\|_2) + \frac{\#bins-1}{KRnh^d} + \epsilon(N+2M))) \tag{85}$$

$$N = (1 + ISB(\hat{f_C}))$$
$$N \leq 1 + \frac{h^2d}{4}\mathcal{R}(\|\nabla f\|_2)$$

$$M \leq IV(\hat{f_C}) + 2\mathcal{R}(f) + \frac{h^2d}{4}\mathcal{R}(\|\nabla f\|_2) + h\sqrt{d}\int_{x\in S}(f(x)\|\nabla f\|_2))$$
$$M \leq IV(\hat{f_H}) + \frac{\#bins-1}{KRnh^d} + 2\mathcal{R}(f) + \frac{h^2d}{4}\mathcal{R}(\|\nabla f\|_2) + h\sqrt{d}\int_{x\in S}(f(x)\|\nabla f\|_2))$$
$$M \leq \frac{1}{nh^d} + \frac{\mathcal{R}(f)}{n} + o(\frac{1}{n}) + \frac{\#bins-1}{KRnh^d} + 2\mathcal{R}(f) + \frac{h^2d}{4}\mathcal{R}(\|\nabla f\|_2) + h\sqrt{d}\int_{x\in S}(f(x)\|\nabla f\|_2))$$

$$IMSE(\hat{f_S}(x)) \leq 12(1-ratio_h)^2 + 3(\frac{1}{nh^d} + \frac{R(f)}{n} + o(\frac{1}{n}) + (1+\epsilon)\frac{h^2d}{4}R(\|\nabla f\|_2) + \frac{\#bins-1}{KRnh^d} + 2\epsilon M + \epsilon) \tag{86}$$

$$\begin{aligned} IMSE(\hat{f_S}(x)) \leq & 12(1-ratio_h)^2 + \\ & 3(1+2\epsilon)(\frac{1}{nh^d} + \frac{R(f)}{n} + o(\frac{1}{n}) + \frac{\#bins-1}{KRnh^d}) + \\ & 3(1+3\epsilon)\frac{h^2d}{4}R(\|\nabla f\|_2) + \\ & 3\epsilon(1 + 2\mathcal{R}(f) + h\sqrt{d}\int_{x\in S}(f(x)\|\nabla f\|_2))) \end{aligned}$$

.

C Other Base lines

Coresets: We considered a comparison with sophisticated data summaries such as coresets. Briefly, a coreset is a collection of (possibly weighted) points that can be used to estimate functions over the dataset. To use coresets to generate a synthetic dataset, we would need to estimate the KDE. Unfortunately, coresets for the KDE suffer from practical issues such as a large memory cost to construct the point set. Despite recent progress toward coresets in the streaming environment Phillips & Tai (2020), coresets remain difficult to implement for real-world KDE problems Charikar & Siminelakis (2017).

Clustering and Importance Sampling: Another reasonable strategy is to represent the dataset as a collection of weighted cluster centers, which may be used to compute the KDE and sample synthetic points. Unfortunately, algorithms such as k-means clustering are inappropriate for large streaming datasets and do not have the same mergeability properties as our sketch. Furthermore, such techniques are unlikely to substantially improve over random sampling when the samples is spread sufficiently well over the support of the distribution. An alternative approach is to select points from the dataset based on importance sampling Charikar & Siminelakis (2017), geometric properties Cortes & Scott (2016), and other sampling techniques Chen et al. (2012). However, recent

experiments show that for many real-world datasets, random samples have competitive performance when compared to point sets obtained via importance sampling and cluster-based approaches Coleman & Shrivastava (2020).

Dimensionality Reduction: One can also apply sketching algorithms to compress a dataset by reducing the dimension of each data point via feature hashing, random projections or similar methods Achlioptas (2003). However, this is unlikely to perform well in our evaluation since our datasets are already relatively low-dimensional. Such algorithms also fail to address the streaming setting, where N can grow very large, because the size of the compressed representation is linear in N. Finally, most dimensionality reduction algorithms do not easily permit the generation of more synthetic data in the original metric space.

D DIFFERENTIALLY PRIVATE DENSITY SKETCHES

In order to make the density sketch differentially private, we add noise to the distribution stored by density sketch. This is achieved by adding noise to the underlying count sketch array ($K \times R$ matrix of integers). Let the function mapping histogram of the data to the density sketch (before the heap construction) be denoted as $f : N^{|X|} \longrightarrow Z^{KR}$ where X is the set of all partitions. We fill first define an discrete analog of laplacian noise.

Definition 1 (Double geometric distribution)**.** *The double geometric distribution parameterized by $p \in (0, 1)$ is defined as follows on the support of all integers.*

$$P(z|p) = \frac{1}{2-p}(1-p)^{|z|}p \tag{87}$$

Algorithm to make Density Sketches private: Each cell of sketch $(K \times R)$ matrix is added an i.i.d noise drawn from the double geometric distribution. We will prove that this noise addition makes the function $\mathcal{M} = f + noise$ differentially private. Heap construction can be considered as an post processing operation on the density sketch matrix. Hence, the sampling distribution is then differentially private. (Note that heap construction algorithm also needs to be modified in practical settings to ensure that it carries the differential privacy properties. But this is achievable)

Theorem 2 (Differential privacy)**.** *The density sketches constructed with addition of double geometric noise with $p = 1 - e^{-\epsilon/K}$ where K is the number of repetitions in the sketch is $(\epsilon, 0)$ differentially private.*

Proof. Consider the l1 metric for computing the distance between datasets. Consider any arbitrary pair x,y which satisfy $\|x - y\|_1 = 1$. In the histogram view of data, it is easy to check that a distance of 1 can exist if and only if there is an additional row in either x or y and all other data points are same. Without loss of generality we can write $x = y \cup \{d\}$ where d is the extra data point. As the constructed count sketch does not depend on the order of insertion, we can say that count sketch for x, i.e. f(x), is obtained from count sketch for y by sketching additional data point into it. Also, because of countsketch's mergeable property, we can write $f(x) = f(y) + f(\{d\})$. Hence $\|f(x) - f(y)\|_1 = \|f(\{d\})\|_1$. As sketching a single entry changes exactly one element of each row of countsketch by 1. $\|f(\{d\})\|_1 = K$. Hence sensitivity of the function f is $\Delta f = K$

We use the double geometric distribution as defined above for noise.

$$P(z|p) = \frac{1}{2-p}(1-p)^{|z|}p \tag{88}$$

Now Let us consider the privacy achieved with this error. Let $\mathcal{M}$ be the final randomized algorithm with computation of f and adding noise. We are interested in the following quantity with x,y such

that $\|x - y\|_1 = 1$.

$$\begin{aligned}\frac{P(\mathcal{M}(x) = z)}{P(\mathcal{M}(y) = z)} &= \frac{\Pi_i P(\mathcal{M}(x)_i = z_i)}{\Pi_i P(\mathcal{M}(y)_i = z_i)} \\ &= \frac{\Pi_i (1-p)^{|f(x)_i - z_i|}}{\Pi_i (1-p)^{|f(y)_i - z_i|}} \\ &= \Pi_i (1-p)^{|f(x)_i - z_i| - |f(y)_i - z_i|} \\ &= (1-p)^{\|f(x) - z\|_1 - \|f(y) - z\|_1}\end{aligned}$$

As l1-norm is a distance metric we can write

$$\begin{aligned}\frac{P(\mathcal{M}(x) = z)}{P(\mathcal{M}(y) = z)} &= (1-p)^{\|f(x) - z\|_1 - \|f(y) - z\|_1} \\ &\geq (1-p)^{\|f(x) - f(y)\|_1} \\ &= (1-p)^{\Delta(f)}\end{aligned}$$

If we put $p = 1 - e^{-\epsilon/\Delta(f)}$

$$\begin{aligned}\frac{P(\mathcal{M}(x) = z)}{P(\mathcal{M}(y) = z)} &\geq e^{-\epsilon} \\ \frac{P(\mathcal{M}(y) = z)}{P(\mathcal{M}(x) = z)} &\leq e^{\epsilon}\end{aligned} \tag{89}$$

Hence $\mathcal{M}(x)$ is $(\epsilon, 0)$- differentially private. Hence we have that the countsketch produced by the sketching algorithm with added double geometric noise is $(\epsilon, 0)$- differentially private when we have $p = 1 - e^{-\epsilon/K}$

why heaps are differentially private? If the data is bounded in R^d (d is the dimension of the data), then it is easy to check that there is a cell in R^d, which contains all the data, It follows that the number of partitions inside this cell is finite. So we can consider heap construction as iteratively going through each partition and noting down its count. Once we do that, we sort all the partitions according to the counts and keep top H elements. In this sense, we can consider heap construction as a post processing over count sketch. From the proposition 2.1 [Dwork, Roth], we know that post processing maintains differential privacy. Hence the heap we create is $(\epsilon, 0)$ differentially private □