A Proof of Lemma 4

The following lemma states that the with high probability, the ratio $\frac{N_{\mathsf{far}}(x)}{N(x)} \to 0$ as n approaches ∞. Throughout this section, $B(x,r)$ denotes the closed Euclidean r-ball about x.

Lemma 4. *Let $x \sim \mathcal{D}_{\mathcal{X}}$. Then, for all $\delta > 0$, $\frac{1}{2} < s < 1$, the hash table T calculated by Algorithm 1 satisfies:*

$$\begin{aligned}\mathbb{P}\Big(N_{\mathsf{far}}(x) >&\delta N(x)\Big) \leq \\ & 2\exp(-0.09n^{\frac{1-s}{2}}) + \Big(\frac{1.6}{n^{1-s}\sqrt{d}^d}\Big)^{\frac{1}{d+1}} \\ &+ \sqrt{\frac{1}{n^{\frac{1-s}{2}}}} + \exp(-\frac{1}{8}n^{s-\frac{1}{2}}) \\ &+ 2^{-\frac{\delta}{2}n^{s-\frac{1}{2}}},\end{aligned}$$

where the probability is over $S_n, x \sim \mathcal{D}^{n+1}$ and the choice of the function g_n.

Proof. Fix $\delta > 0$, $\varepsilon = \frac{r_n}{\sqrt{d}}$, and let $C_1, \dots, C_t$ be a partition of $[0,1]^d$ into $t = (\frac{1}{\varepsilon})^d$ boxes of length ε. Notice that for any x, x' in the same box, we have $\|x - x'\| \leq \sqrt{d}\varepsilon$. Put $k = n^s$ and define the random variable $L_{\varepsilon,k}(S_n) = \Sigma_{i:|C_i \cap S_n| < k}\, \mathbb{P}(C_i)$, and note that it is precisely the k-missing mass (defined in (2)) associated with the distribution $P = (\mathbb{P}(C_1), \dots, \mathbb{P}(C_t))$. By Theorem 2(a), we have $\mathbb{E}[L_{\varepsilon,k}(S_n)] \leq \frac{1.6kt}{n}$. By the law of total probability,

$$\begin{aligned}\mathbb{P}\Big(N_{\mathsf{far}}(x) >&\delta N(x)\Big) \\ &\leq \mathbb{P}\Big(L_{\varepsilon,m}(S_n) > \frac{1.6}{\varepsilon^d n^{1-s}} + \gamma\Big) \\ &+ \mathbb{P}\Big(N_{\mathsf{far}}(x) > \delta N(x) \;\Big|\; \\ &\qquad L_{\varepsilon,m}(S_n) \leq \frac{1.6}{\varepsilon^d n^{1-s}} + \gamma\Big). \qquad (18)\end{aligned}$$

For the first term in (18), we apply Theorem 2(b):

$$\begin{aligned}\mathbb{P}\Big(L_{\varepsilon,m}(S_n) >& \frac{1.6}{\varepsilon^d n^{1-s}} + \gamma\Big) \\ &\leq \mathbb{P}\Big(L_{\varepsilon,m}(S_n) > \mathbb{E}[L_{\varepsilon,m}(S_n)] + \gamma\Big) \\ &\leq 2\exp\big(-0.09n^{1-s}\gamma^2\big).\end{aligned}$$

For the second term in (18), we have

$$\begin{aligned}\mathbb{P}\Big(N_{\mathsf{far}}(x) > \delta N(x) \;\Big|\; & L_{\varepsilon,m}(S_n) \leq \frac{1.6}{\varepsilon^d n^{1-s}} + \gamma\Big) \\ \leq \mathbb{P}\Big(&|B(x,r_n) \cap S_n| < n^s \;\Big| \\ & L_{\varepsilon,m}(S_n) \leq \frac{1.6}{\varepsilon^d n^{1-s}} + \gamma\Big) \\ + \mathbb{P}\Big(&N_{\mathsf{far}}(x) > \delta N(x),\; |B(x,r_n) \cap S_n| \geq n^s \;\Big| \\ & L_{\varepsilon,m}(S_n) \leq \frac{1.6}{\varepsilon^d n^{1-s}} + \gamma\Big) \\ = (*) &+ (**). \qquad (19)\end{aligned}$$

Since $r_n = \sqrt{d}\varepsilon$, we have $\{|B(x,r_n) \cap S_n| < n^s\} \implies \{|C(x) \cap S_n| < n^s\}$, where $C(x)$ is the ε-length box containing x. Thus,

$$(*) \leq \frac{1.6}{\varepsilon^d n^{1-s}} + \gamma.$$

We are left to bound the second term in (19)

$$\begin{aligned}(**) \leq& \mathbb{P}\Big(N_{\mathsf{close}}(x) < \frac{1}{2}n^{s-\frac{1}{2}} \;\Big| \\ & L_{\varepsilon,m}(S_n) \leq \frac{1.6}{\varepsilon^d n^{1-s}} + \gamma,\; |B(x,r_n) \cap S_n| > n^s\Big) \\ &+ \mathbb{P}\Big(N_{\mathsf{far}}(x) > \delta N(x), N_{\mathsf{close}}(x) \geq \frac{1}{2}n^{s-\frac{1}{2}} \;\Big| \\ & L_{\varepsilon,m}(S_n) \leq \frac{1.6}{\varepsilon^d n^{1-s}} + \gamma,\; |B(x,r_n) \cap S_n| > n^s\Big) \\ &= (***) + (****). \qquad (20)\end{aligned}$$

Since the algorithm set $m_n = \lfloor \frac{\log n}{2\log(\frac{1}{p_1})} \rfloor$, we have

$$\begin{aligned}\mathbb{E}\Big[N_{\mathsf{close}}(x) \Big| &|B(x,r) \cap S_n| > n^s\Big] \\ &\geq p_1^{m_n} n^s \\ &\geq p_1^{\frac{\log n}{2\log(\frac{1}{p_1})}} n^s \\ &\geq \big(2^{\log p_1}\big)^{\frac{1}{2\log\frac{1}{p_1}}\log n} n^s \\ &\geq n^{s-\frac{1}{2}}.\end{aligned}$$

Let $Z \sim \mathrm{Bin}(n^s, p_1^{m_n})$. We have $\mathbb{E}\Big[N_{\mathsf{close}}(x) \Big| |B(x,r_n) \cap S_n| > n^s\Big] \geq \mathbb{E}[Z] = n^{s-\frac{1}{2}}$. In addition, for each $x' \in A_{\mathsf{close}}(x)$ we have $\mathbb{P}\big(g_n(x) = g_n(x')\big) \geq p_1^{m_n}$, and thus, invoking the Chernoff bound,

$$\begin{aligned}(***) \leq& \\ &\mathbb{P}(Z < \frac{1}{2}n^{s-\frac{1}{2}}) \\ &= \mathbb{P}(Z < \frac{1}{2}\mathbb{E}[Z]) \\ &\leq \exp(-\frac{1}{8}\mathbb{E}[Z]) \\ &\leq \exp(-\frac{1}{8}n^{s-\frac{1}{2}}).\end{aligned}$$

The last term we have to bound is the second term in (20). Notice that

$$\{N_{\mathsf{far}}(x) > \delta N(x), N_{\mathsf{close}}(x) \geq \frac{1}{2}n^{s-\frac{1}{2}}\} \implies \{N_{\mathsf{far}}(x) > \frac{\delta}{2}n^{s-\frac{1}{2}}\}.$$

In addition, since $p_1^2 > p_2$, we have

$$\mathbb{E}\big[N_{\mathsf{far}}(x)\big] \leq p_2^{m_n} n \leq p_1^{2m_n} n \leq p_1^{2\left(\frac{\log n}{2\log\frac{1}{p_1}}-1\right)} n = p_1^{-2} = O(1).$$

Since for each $x' \in A_{\mathsf{far}}(x)$ we have $\mathbb{P}\big(g_n(x) = g_n(x')\big) \leq p_2^{m_n}$, if we let $Z \sim \mathrm{Bin}(n, p_2^{m_n})$ then, by Chernoff's bound,

$$(****) \leq \mathbb{P}(Z > \frac{\delta}{2}n^{s-\frac{1}{2}}) \leq 2^{\frac{\delta}{2}n^{s-\frac{1}{2}}}.$$

For $s > \frac{1}{2}$ and large enough n s.t. $2e\mathbb{E}\big[N_{\mathsf{far}}(x)\big] \leq 2e\mathbb{E}[Z] \leq 2e \leq \frac{\delta}{2}n^{s-\frac{1}{2}}$.

Finally, setting $\gamma = \sqrt{\frac{1}{n^{\frac{1-s}{2}}}}$, $r_n = \left(\frac{1.6\sqrt{d}^{d+2}}{n^{1-s}}\right)^{\frac{1}{d+1}}$ we conclude our proof. □

B Proof of Lemma 5

Here we show that with high probability, the variable $N(x) \to \infty$. Namely, the number of sample points at each bucket is increasing as n goes to ∞.

Lemma 5. *Let $x \sim \mathcal{D}_{\mathcal{X}}$ be a test point. Then, for all $M > 0$, $\frac{1}{2} < s < 1$ the hash table calculated by Algorithm 1 satisfies:*

$$\mathbb{P}\Big(N(x) < M\Big) \leq \exp(-\frac{n^{s-\frac{1}{2}}}{2} + M) + 2\exp(-0.09n^{\frac{1-s}{2}}) + \Big(\frac{1.6}{n^{1-s}\sqrt{d}^d}\Big)^{\frac{1}{d+1}} + \sqrt{\frac{1}{n^{\frac{1-s}{2}}}}.$$

Where, again, the probability is over $S_n, x \sim \mathcal{D}^{n+1}$ and the choice of the function g_n.

Proof. Fix $M > 0$. Similar to Lemma 4, we have

$$\mathbb{P}\Big(N(x) < M\Big) \leq \mathbb{P}\Big(N(x) < M\Big| L_{\varepsilon,m}(S_n) \leq \frac{1.6}{\varepsilon^d n^{1-s}},\ |B(x, r_n) \cap S_n| > n^s\Big) + 2\exp(-0.09n^{\frac{1-s}{2}}) + \frac{1.6\sqrt{d}^d}{r_n^d n^{1-s}} + \sqrt{\frac{1}{n^{\frac{1-s}{2}}}}. \tag{21}$$

We only have to bound the first term in (21). Observe that

$$\{N(x) < M\} \implies \{N_{\mathsf{close}}(x) < M\}$$

and that $\mathbb{E}\big[N_{\mathsf{close}}(x)\big| |B(x, r_n \cap S_n)| > n^s\big] \geq \mathbb{E}[Z] = p_1^{m_n} n^s = n^{s-\frac{1}{2}}$. Now for $Z \sim \mathrm{Bin}(n^s, p_1^{m_n})$, if we let $\xi = 1 - \frac{M}{\mathbb{E}[Z]}$, then by Chernoff's bound we have,

$$\begin{aligned}\mathbb{P}\Big(N_{\mathsf{close}}(x) < M\Big| & L_{\varepsilon,m}(S_n) \leq \frac{1.6}{\varepsilon^d n^{1-s}}, |B(x, r_n) \cap S_n| > n^s\Big) \\ &\leq \mathbb{P}(Z < (1-\xi)\mathbb{E}[Z]) \\ &\leq \exp(-\frac{\xi^2}{2}\mathbb{E}[Z]) \\ &\leq \exp(-\frac{n^{s-\frac{1}{2}}}{2} + M).\end{aligned}$$

□