1. (a) The cardinality of $\mathcal{H}$ is d
$\Rightarrow$ the num of mistakes $< \log_2 d$.
Let $d = 2^n$, and suppose h_d is the true learning rule.
Let $x_1 = (-1, \ldots, \underset{d/2}{-1}, \underset{d/2+1}{1}, \ldots, 1)$. Then $h_{majority}(x_1) = \pm 1$

WLOG, $h^{(1)}_{majority}(x_1) = -1$. Then we get rid of predictors $h_1, \ldots, h_n$.

Let $x_K = (\underbrace{-1, \ldots, -1}_{2^{n-1}}, \underbrace{-1, \ldots, -1}_{2^{n-2}}, -1, \ldots, \underbrace{-1, \ldots, -1}_{2^{n-K}}, 1, \ldots, 1)$.

Then WLOG $h^{(n)}_{majority}(x_K) = -1$ and 2^{n-K} are kicked out.
We then require $n-1$ mistakes before we are left w/ h_d.

$$2^{n-1} + 2^{n-2} + \cdots + 2^{n-(n-1)} = 2^n - 1$$

Thus, the worst-case num of mistakes is
$\lceil \log_2 d \rceil - 1$

(b). For every $d = 2^{2^m}$, we may make $2^{2^{m-1}} + 2^{2^{m-2}} + \cdots + 2^2 + 2 + 2$ mistakes.

Base Case: Let $x_0^{(0)}, x_1^{(1)}$ be $(+1, \ldots, +1), (-1, \ldots, -1)$

Let $x_0^{(1)} = (+1, \ldots, +1, -1, \ldots, -1)$
$x_1^{(1)} = (-1, \ldots, -1, +1, \ldots, +1)$

Inductive Step: Sps we have $X^{(0)} = (x_0^{(0)}, x_1^{(0)})$, $X^{(1)} = (x_0^{(1)}, x_1^{(1)})$, $\ldots$, $X^{(n)} = (x_0^{(n)}, x_1^{(n)}, \ldots, x^{(n)}_{2^{2^{n-1}}-1})$

define $X^{(n+1)}$ as follows: divide interval $[0,1]$ into 2^{n+1} subintervals. If an interval is colored, the corresponding coordinate is 1. If it is uncolored, it is -1.

Pair the sub-intervals. A sub-interval is on if it is colored on then off $[\longmapsto \quad]$, off if off, then on: $[\quad \longmapsto]$. Interpret the elements of $X^{(0)}, \ldots, X^{(n)}$ as binary, and generate $|X^{(0)}| + \cdots + |X^{(n)}|$ elements By turning on the corresponding sub interval if that element, in order is on.

We have generated the binary sequences of len 2^n where half are on. Now we generate the rest.
For each element we have just generated, take its binary sequence of switches. Flip the last switch, add these in sequence to our list. Then do the same but for the second to last. Continue until we've done this to all.
We have now generated the elements whose switches are 2^n+1 on, and the elements where 2^n-1 are on.

Taking this new sublist, we generate the 2^n+2 on elements by turning on an off switch for each 2^n+1 ele. and generate the 2^n-2 on elements by turning off an on switch in the 2^n-1 on elements.

Continuing in this manner, we have successfully defined $X^{(n+1)}$, of which there are

$$\binom{2^n}{2^{n-1}} + \binom{2^n}{2^{n-1}-1} + \binom{2^n}{2^{n-1}+1} + \cdots = 2^{2^n}$$

elements.

Now, we must see how the nearest neighbor algorithm makes a mistake on each new element added.
The recursive elements are trivial to see this for, since the distance of every element in $X^{(n+1)}$ to one in the previous is always $\sqrt{2^n}$, and the distance to each element added is a constant times the distance when the corresponding element was added. Therefore, there will be a tie between the correct and incorrect neighbors.

For the non-recursive ones, each added also ties and may make a mistake.

(c) Let $\mathcal{D}$ be the dist over $\mathbb{R}^d$ of the spherical Gaussian. Labels are $y = \text{sgn}(x[i])$ for a fixed i.
Let h_m be a nearest neighbor predictor based on $S \sim \mathcal{D}^m$ of m iid samples.
Show that $\exists\, c \in (0,1)$ s.t. for $m = 2^{cd}$
$E_S(L(h_m)) > 0.4$.
In fact, $L(h_m) > 0.4$ with high probability.

<u>Proof</u>: <u>Recall</u>: If $x \sim \mathcal{D}$ spherical gaussian on $\mathbb{R}^d$ mean 0, unit variance. Then for all $\beta \le \sqrt{d}$

$$P(\sqrt{d} - \beta \le \|x\| \le \sqrt{d} + \beta) \ge 1 - 3e^{-\beta^2/96}$$

for some $c \ge 0$.
Thus for a, b drawn from a spherical gaussian,
$\|a\| \approx \sqrt{d} \pm O(1)$
$\|b\| \approx \sqrt{d} \pm O(1)$
Additionally points drawn at random are nearly orthogonal; so, by Pythagoras
$\|a-b\|^2 \approx 2(\sqrt{d} \pm O(1))^2 = 2d + O(\sqrt{d})$.
Thus vectors picked spherical gaussian are also approximately orthogonal.

2. Parzen Window Predictor

Let $S = \{(x_i, y_i)\}_{i=1}^{m}$. The Parzen Window (Kernel) Density Estimate $\hat{f}(x|Y=y)$ of $f(x|Y=y)$, for each $y \in \mathcal{Y}$ is.

$$\hat{f}(x|y) := Z_y \sum_{i:y_i=y} K(x, x_i)$$

where $Z_y = \left(\int_{\mathcal{X}} \sum_{i:y_i=y} K(x, x_i)\, dx\right)^{-1}$, and.

$$K(x, x_i) := \exp(-\rho(x, x_i)^2/\sigma^2)$$

where ρ is a translation invariant metric on $\mathcal{X}$, σ is a hyper-parameter.

The Parzen Window estimator $\hat{\mathcal{D}}(\mathcal{X}, \mathcal{Y})$ is given by $\hat{f}(x|y)$ combined with $\hat{p}(y) := \frac{1}{m}|\{i \mid y_i = y\}|$, or the count of num of diff labels among the m samples.

The Parzen Predictor is the Bayes' Optimal Predictor for $\hat{\mathcal{D}}(\mathcal{X}, \mathcal{Y})$. It is specified by choice of $\rho(x, x')$ and σ.

(a) Show that for $\mathcal{Y} = \{-1, +1\}$, the Parzen Predictor is given by:

$$h(x) := \text{sign}\left(\sum_{i=1}^{m} y_i K(x, x_i)\right)$$

<u>proof</u>: Recall:

$$h_{Bayes(\mathcal{D})}(x) = \text{sign}(\eta_{\mathcal{D}}(x) - \tfrac{1}{2}).$$

and the Parzen Predictor is given by

$$h_{Bayes(\hat{\mathcal{D}})}(x) = \text{sign}(\eta_{\hat{\mathcal{D}}}(x) - \tfrac{1}{2}).$$

We have, by def'n of the estimator of the joint density.

$$\hat{f}(x, y) := \hat{f}(x|y) \cdot \hat{p}(y).$$

Recall:

$$\eta_{\mathcal{D}}(x) := P(Y=1 \mid X=x)$$

$$\therefore \eta_{\hat{\mathcal{D}}}(x) = P(Y=1 \mid X=x) = \frac{\hat{f}(x, 1)}{E_y[\hat{f}(X|y)]\,|_{X=x}}$$

$$E_y[\hat{f}(x|y)] = \hat{f}_X(x).$$

$$E_y[\hat{f}(X|y)] = \hat{f}(X|+1) \cdot \underbrace{\hat{p}(+1)}_{=p} + \hat{f}(X|-1)\hat{p}(-1)$$

$$= p\, Z_1 \sum_{i:y_i=1} K(x, x_i) + (1-p)\, Z_{-1} \sum_{j:y_j=-1} K(x, x_j)$$

<u>Recall</u>: ρ shift-invariant implies

$$\int_{\mathcal{X}} K(x, x_i)\, dx = C \quad \forall i.$$

$$\therefore \int_{\mathcal{X}} Z_1 \sum_{i:y_i=1} K(x, x_i)\, dx = Z_y \cdot C \cdot \underbrace{|\{i : y_i = y\}|}_{m \cdot \hat{p}(y)} = 1.$$

$$\Rightarrow Z_y = \frac{1}{m \cdot \hat{p}(y) \cdot C}$$

Thus.

$$\hat{f}_X(x) = \frac{\cancel{p}}{m\cancel{p}C} \sum_{i:y_i=1} K(x, x_i) + \frac{\cancel{(1-p)}}{m \cdot \cancel{(1-p)}C} \sum_{j:y_j=-1} K(x, x_j).$$

$$= \frac{1}{mC} \sum_{i=1}^{m} K(x, x_i)$$

$$\therefore \eta_{\hat{\mathcal{D}}}(x) = \frac{\frac{\cancel{p}}{m\cancel{p}\cancel{C}} \sum_{i:y_i=1} \overbrace{y_i}^{=1} K(x, x_i)}{\frac{1}{\cancel{mC}} \sum_{i=1}^{m} K(x, x_i)} = \frac{\sum_{i:y_i=1} y_i K(x, x_i)}{\sum_{i=1}^{m} K(x, x_i)}$$

$$\therefore \eta_{\hat{\mathcal{D}}}(x) - \tfrac{1}{2} = \frac{2\sum_{i:y_i=1} y_i K(x, x_i) - \sum_{i=1}^{m} K(x, x_i)}{2\sum_{i=1}^{m} K(x, x_i)}$$

$$= \frac{\sum_{i:y_i=1} y_i K(x, x_i) + \sum_{i:y_i=-1} y_i K(x, x_i)}{2\sum_{i=1}^{m} K(x, x_i)}$$

$$= \frac{\sum_{i=1}^{m} y_i K(x, x_i)}{\underbrace{2\sum_{i=1}^{m} K(x, x_i)}_{\geq 0}}$$

$$\Rightarrow h(x) = \text{sgn}(\eta_{\hat{\mathcal{D}}}(x) - \tfrac{1}{2}) = \text{sgn}\left(\sum_{i=1}^{m} y_i K(x, x_i)\right) \square$$

(b) Using $K(x, x') = \exp(-\rho(x, x')^2/\sigma^2)$, how does the Parzen predictor h behave in the limit as $\sigma \to \infty$?

<u>Soln</u>: as $\sigma \to \infty$, we have

$\exp(-\rho(x, x')/\sigma^2) \to \exp(0) = 1$, so for every x

$$h(x) = \text{sign}\left(\sum_{i=1}^{m} y_i K(x, x_i)\right) \to \text{sign}\left(\sum_{i=1}^{m} y_i\right).$$

i.e. Constant majority predictor.

(c) how does h behave as $\sigma \to 0$?

<u>Soln</u>: $\exp(-\rho(x, x')^2/\sigma^2) \to \exp(-\infty) = 0$, if $x \neq x'$

if $x = x'$, $\rho(x, x') = 0$, so $\exp(-\rho(x, x')^2/\sigma^2) \to \exp(0) = 1$.

Thus $h(x) \to \text{sgn}\left(\sum_{i=1}^{m} y_i \mathbb{1}\{x_i = x\}\right)$

Memorization.

(c) When σ is small, $\rho(x, x')^2/\sigma^2$ is large, except when x' is near to x. Thus, when predicting on x, we bias the labels s.t. x_i is near to x.

When σ grows larger, this effect is less pronounced.

In effect, this is a weighted m-NN predictor with weights $K(x, x_i)$.

Whenever one ties, we pick arbitrarily, in accordance with

$$\text{sign}(0) = 1.$$

3. Nearest Neighbor in The Statistical Setting.

(a) $X \sim \mathcal{U}nif[-1,1]$, $y = \pm 1$, $P_D(Y=+1|x) = 0.5 + 0.3\,sgn(x)$

What is the Bayes Optimal Predictor h^* and the Bayes Error $L_D(h^*)$?

Sol'n: $P_D(x)$ is precisely our $\eta(x)$, our posterior dist.

Therefore

$$h^*(x) = Sgn\left(P_D(x) - \tfrac{1}{2}\right) = sign(0.3\, sign(x)) = sign(x)$$

$$\begin{aligned} L_D(h^*) &= P_D(h^*(x) \neq y^*) \\ &= P(y \neq +1 | X \geq 0) \cdot P(X \geq 0) + P(y \neq -1 | X < 0) P(X < 0) \\ &= P(y = -1 | X \geq 0) \cdot \tfrac{1}{2} + P(y = 1 | X < 0) \cdot \tfrac{1}{2} \\ &= \tfrac{1}{2}(1 - 0.5 - 0.3) + \tfrac{1}{2}(0.5 - 0.3) \\ &= \tfrac{1}{2} \cdot 0.8 + \tfrac{1}{2} \cdot 0.2 = \tfrac{1}{2}. \end{aligned}$$

(b) How does the error of the nearest neighbor predictor behave, $L(h_m)$ as $m \to \infty$?

Soln: Given $x \in [-1,1]$, Let $Z = |X - x|$, $X \sim \mathcal{U}nif(-1,1)$.

The CDF of Z is:

$$P(|X - x| \leq t) = \frac{\mu([-1,1] \cap [x-t, x+t])}{\mu([-1,1])} = \frac{\mu([\max\{x-t,-1\}, \min\{1, x+t\}])}{\mu([-1,1])}$$

$$= \frac{\min\{1, x+t\} - \max\{-1, x-t\}}{2}$$

$$= \frac{1 + (x+t) - |1 - (x+t)| + 1 - (x-t) + |x - t + 1|}{4}$$

$$= \frac{1}{2}\left(t + 1 - \frac{|1 - (x+t)| + |x - t + 1|}{2}\right)$$

Let $Z_1, \ldots, Z_m \overset{iid}{\sim} Z$, $Z_{(1)} := \min_i Z_i$

Then, the CDF of $Z_{(1)}$ is:

$$P(Z_{(1)} \leq t) = 1 - [1 - P(Z \leq t)]^m$$

$$= 1 - \left[1 - \frac{1}{2}\left(t + 1 - \frac{|x + t - 1| + |x - t + 1|}{2}\right)\right]^m$$

$$= 1 - \left[\frac{1}{2} - \frac{1}{2}t - \frac{|x + (t-1)| + |x - (t-1)|}{2}\right]^m$$

$$= 1 - 2^{-m}\left[\frac{1}{2} - \frac{t}{2} - \frac{1}{2}|x + (t-1)| + |x - (t-1)|\right]^m$$

$$\approx 1 - (1-t)^m$$

$$\Rightarrow P(Z_{(1)} \geq t) \approx (1-t)^m.$$

Finally,

$$P(h_m(x) \neq sgn(x)) \leq P(Z_{(1)} \geq |x|) \approx (1 - |x|)^m.$$

$\therefore$ as $m \to \infty$, $h_m(x)$ converges to $h_{Bayes}(x)$ almost surely since the set of points on which h_m does not converge to $sgn(x)$ is $\emptyset$.