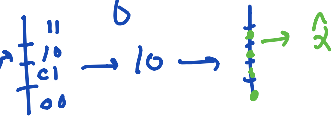
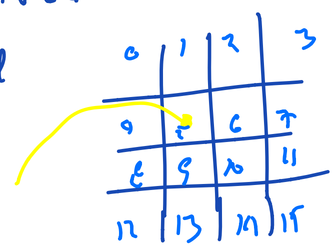
Empirical vector quantization (Chap. 13)

scalar quantizer $x \mapsto$ 

Vector quantization

$z \in \mathbb{R}^d$ 

$\xrightarrow{5}$ $(-1,-1) = \hat{z}$

Given a probability dist^n P on $\mathbb{R}^d$ there exists an optimal quantizer (for MSE loss: $\underset{P}{E}\left[\|z-\hat{z}\|^2\right] = L(q)$

$(q, \mathcal{C})$

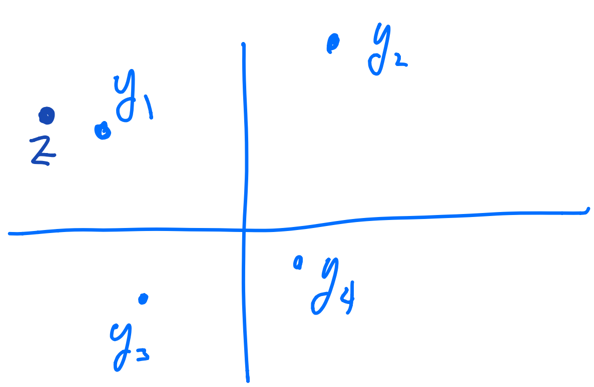
$$q : \mathbb{R}^d \to \{1, \cdots, k\}$$

Codebook $\mathcal{C}$ $\quad y_1, \cdots, y_k \in \mathbb{R}^d$

$$z \longrightarrow \boxed{q(\cdot)} \xrightarrow{\;k\;} \boxed{} \longrightarrow y_k$$

$q(z) = i$

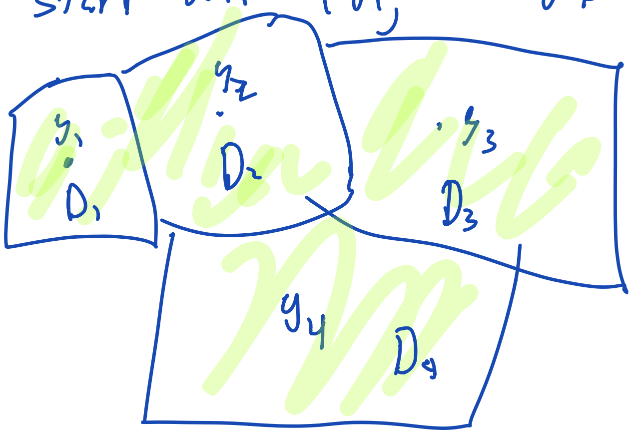
$\mathbb{R}^d$ $d=2$
$k=4$



Selecting q to minimize distance between input z and output y_i is optimal for a given codebook.

Algorithm for finding good code book

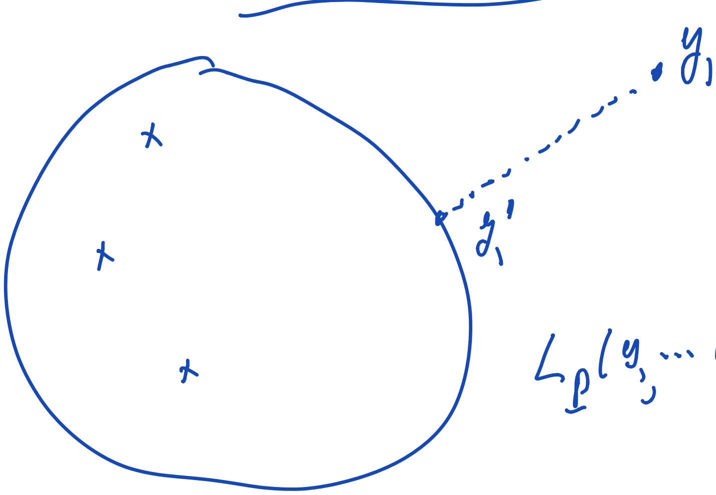
start with $\{y_1, \cdots y_k\} = \mathcal{C}^t$



$\mathcal{C}^{t+1}$ - use $y_i^{(t+1)} = E[Z \mid Z \in D_i]$

Theorem Optimal quantizers exist.

special case suppose Z's all take values in a ball of radius r in $\mathbb{R}^d$.



without loss of generality assume $y_1, \cdots y_k$ are in ball.

$$L_P(y_1, \cdots y_k) = E\left[\min_{1 \le i \le k} \|y_i - Z\|^2\right]$$

$g_1 \cdots g_k \to L_P(g_1, \cdots g_k)$ is a continuous function over compact set $\underbrace{B(r) \times \cdots \times B(r)}_{k \text{ times}}$

so minimum exists.

To connect to course, think of Z_i as a feature, with label equal to Z_i

$$Z_i \longrightarrow \boxed{q(\cdot)} \longrightarrow \boxed{} \longrightarrow \hat{Z}_i$$

If $\hat{Z}_i = Z_i$ then loss is zero.

Suppose data $Z_1, \cdots Z_n$ is given in $B(r) \subset \mathbb{Z}^d$, independent, distⁿ $\underline{P}$

(From now on, let $q_k : \mathbb{R}^d \to \mathbb{R}^d$ with k possible values in $\mathbb{R}^d$, and NN (nearest neighbor) property.

$$D(P_Z, q_k) = \int_{\mathbb{R}^d} \| z - q_k(z) \|^2 \, \underline{P}(dz)$$

risk for fresh sample.

P_n = empirical dist^n

$$\hat{q}_n = \arg\min_{q \in Q_k} D(P_n, q) = \frac{1}{n}\sum_{i=1}^{n} \|z_i - q(z_i)\|^2$$

empirically optimal quantizer

Theorem There exist an absolute constant C so that

$$\sup_{P_Z \in \mathcal{P}(r)} E\left[D(P_Z, \hat{q}_n) - D_k^*(P_Z)\right] \le Cr^2 \sqrt{\frac{k(d+1)\log k(d+1)}{n}}$$

Let $f_q(z) = \|z - q(z)\|^2$

Then $D(P, q) = P(f_q) = \int_{\mathbb{R}^d} f_q(z)\, P(dz)$

$\mathcal{F} = \{ f_q : q$ is a NN quantizer for some codebook $\{y_1, \ldots, y_k\} \subset B(r) \subset \mathbb{R}^d \}$

By the miss-matched minimization lemma

empirical dist^n

$$D(P_Z, \hat{q}_n) - D_k^*(P_Z) \le 2 \sup_{f \in \mathcal{F}} \underbrace{\left| D(P_Z, f) - D(P_n, f) \right|}_{\Delta(Z^n)}$$

$\mathcal{F}, P$

$$f_q = \| Z - q(Z) \|^2 \le (2r)^2 = 4r^2$$

(Note f_q's are not binary valued)

Idea $\quad EU = \int_0^\infty P(U \ge u)\, du$

$$E_P f_q = \int_0^{4r^2} P\{f_q \ge u\}\, du$$

$$E_{P_n} f_q = \int_0^{4r^2} P_n\{f_q \ge u\}\, du$$

suppose different by at most ε

$$\Delta_n(Z^n) = \sup_q \int_0^{4r^2} P\{f_q \ge u\} - P_n\{f_q \ge u\}\, du$$

$$\le \int_0^{4r^2} \varepsilon\, du = 4r^2 \varepsilon$$

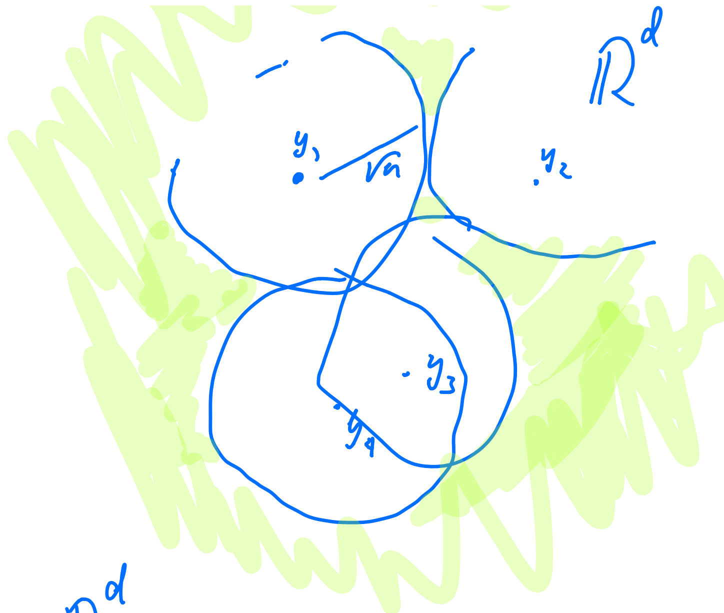
So lets look at $\mathcal{C} = \left\{ \{z : f_q(z) > u\} : \begin{array}{l} q \in Q_k \\ 0 \le u \le 4r^2 \end{array} \right\}$

Then $E[\Delta_n(Z^n)] \le 4r^2 C \sqrt{\dfrac{V(\mathcal{C})}{n}}$

$C = \{ f_\theta(z) > u \}$

i.e. $\|z - y_i\|^2 \geq u$

for $1 \leq i \leq k$

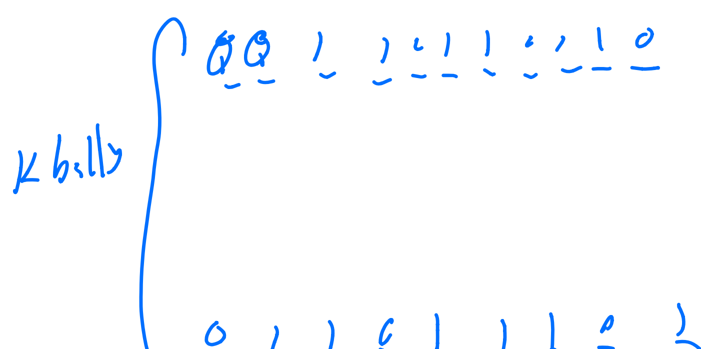


- set of balls in $\mathbb{R}^d$ has VC dimension $d+1$ (Dudley class)
- If $\mathcal{C}$ is a family of sets $\tilde{\mathcal{C}} = \{ C^c : C \in \mathcal{C} \}$ has same VC dimension as $\mathcal{C}$.

So we need upper bound on set of all subsets of $\mathbb{R}^d$ that are unions of k balls (k fixed).

n-th shatter coefficient for sets of all balls

$\leq (n+1)^d$



max $\rightarrow$ 0 1 1 0 1 1 1 0 1

So shatter coefficient for union of k balls

is $\leq (n+1)^{kd}$ //

So if $(n+1)^{kd} < 2^n$ then $n \leq VC(\mathcal{C})$

$n = 4k(d+1)\log(k(d+1))$ does the job.

$\square$

Dimensionality Reduction in Hilbert spaces

Chapter 14

find mapping from high dimensional space to a k dimensional subspace of same space.

$Z \longrightarrow \boxed{} \longrightarrow \hat{Z}$