CASE STUDY: EMPIRICAL QUANTIZER DESIGN

MAXIM RAGINSKY

Now that we have safely made our way through the combinatorial forests of Vapnik–Chervonenkis classes, we will look at an interesting application of the VC theory to a problem in communications engineering: empirical design of vector quantizers. Vector quantization is a technique for *lossy data compression* (or *source coding*), so we will first review, at a very brisk pace, the basics of source coding, and then get to business. The presentation will closely follow an excellent survey by Tamás Linder [Lin01].

1. A brief introduction to source coding

It's trite but true: we live in a digital world. We store, exchange, and manipulate vast quantities of binary data. While a lot of the data are inherently discrete (e.g., text), most are *compressed representations* of continuous-valued (analog) sources, such as audio, speech, images, or video. The process of mapping source data from their "native" format to binary representations and back is known in the information theory and the communications engineering communities as *source coding*.

There are two types of source coding: *lossless* and *lossy*. The former pertains to constructing compact binary representations of discrete data, such as text, and the objective is to map any sequence of symbols emitted by the source of interest into a binary file which is as short as possible and which will permit *exact* (i.e., error-free) reconstruction (decompression) of the data. The latter, on the other hand, deals with continuous-valued sources (such as images), and the objective is to map any source realization to a compact binary representation that would, upon decompression, differ from the original source as little as possible. We will focus on lossy source coding. Needless to say, we will only be able to give a very superficial overview of this rich subject. A survey article by Gray and Neuhoff [GN98] does a wonderful job of tracing both the historical development and the state of the art in lossy source coding; for an encyclopedic treatment I recommend the book by Gersho and Gray [GG92].

One of the simpler models of an analog source is a stationary stochastic process $Z_1, Z_2, \dots$ with values in $\mathbb{R}^d$. For example, if d is a perfect square, then each Z_i could represent a $\sqrt{d} \times \sqrt{d}$ image patch. The compression process consists of two stages. First, each Z_i is mapped to a binary string b_i. Thus, the entire data stream $\{Z_i\}_{i=1}^{\infty}$ is represented by the sequence of binary strings $\{b_i\}_{i=1}^{\infty}$. The source data are reconstructed by mapping each b_i into a vector $\widehat{Z}_i \in \mathbb{R}^d$. Since each Z_i takes on a continuum of values, the mapping $Z_i \mapsto b_i$ is inherently *many-to-one*, i.e., noninvertible. This is the reason why this process is called *lossy* source coding — in going from the analog data $\{Z_i\}$ to the digital representation $\{b_i\}$ and then to the reconstruction $\widehat{Z}_i$, we lose information needed to recover each Z_i exactly. The overall mapping $Z_i \mapsto b_i \mapsto \widehat{Z}_i$ is called a *vector quantizer*, where the term "vector" refers to the vector-valued nature of the source $\{Z_i\}$, while the term "quantizer" indicates the process of representing a continuum by a discrete set. We assume that the mappings comprising the quantizer are time-invariant, i.e., do not depend on the time index $i \in \mathbb{N}$.

There are two figures of merit for a given quantizer: the compactness of the binary representation $Z_i \mapsto b_i$ and the accuracy of the reconstruction $b_i \mapsto \widehat{Z}_i$. Te former is given by the *rate* of the

Date: February 28, 2011.

quantizer, i.e., the expected length of b_i in bits. Since the source $\{Z_i\}$ is assumed to be stationary and the quantizer is assumed to be time-invariant, we have

$$\mathbb{E}[\mathrm{len}(b_i)] = \mathbb{E}[\mathrm{len}(b_1)], \qquad \forall i \in \mathbb{N},$$

where, for a binary string b, $\mathrm{len}(b)$ denotes its length in bits. If the length of $b_i \equiv b_i(Z_i)$ depends on Z_i, then we say that the quantizer is *variable-rate*; otherwise, we say that the quantizer is *fixed-rate*. The latter measures how well the reconstruction $\widehat{Z}_i$ approximates the source Z_i on average. In order to do that, we pick a nonnegative *distortion measure* $d : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty)$, so that $d(z, \widehat{z}) \geq 0$ quantifies how well one vector $z \in \mathbb{R}^d$ is approximated by another $\widehat{z} \in \mathbb{R}^d$. Then we look at the expected value $\mathbb{E}[d(Z_i, \widehat{Z}_i)]$, which is the same for all i, again owing to the stationarity of $\{Z_i\}$ and the time invariance of the quantizer. A typical distortion measure is the squared Euclidean norm

$$d(z, \widehat{z}) = \|z - \widehat{z}\|^2 = \sum_{j=1}^{d} |z(j) - \widehat{z}(j)|^2,$$

where $z(j)$ denotes the jth coordinate of z. We will focus only on this distortion measure from now on.

Now, using the fact that the rate and the expected distortion of a quantizer do not depend on the time index i, we can just consider the problem of quantizing a single $\mathbb{R}^d$-valued random variable Z with the same distribution as that of Z_1. From now on, we will refer to such a Z as the source. Thus, the rate of a given quantizer $Z \mapsto b \mapsto \widehat{Z}$ is given by $\mathbb{E}[\mathrm{len}(b)]$ and the expected distortion $\mathbb{E}\|Z - \widehat{Z}\|^2$. Naturally, one would like to keep both of these as low as possible: low rate means that it will take less memory space to store the compressed data and that it will be possible to transmit the compressed data over low-fidelity (i.e., low-capacity) digital channels; low expected distortion means that the reconstructed source will be a very accurate approximation of the true source. However, these two quantities are in conflict: if we make the rate too low, we will be certain to incur a lot of loss in reconstructing the data; on the other hand, insisting on very accurate reconstruction will mean that the binary representation must use a large number of bits. For this reason, the natural question is as follows: what is the smallest distortion achievable on a given source by any quantizer with a given rate?

2. Fixed-rate vector quantization

Let $\mathsf{Z} = \mathbb{R}^d$.

Definition 1. *Let $k \in \mathbb{N}$. A (d-dimensional) k*-point vector quantizer *is a (measurable) mapping $q : \mathsf{Z} \to \mathcal{C} = \{y_1, \ldots, y_k\} \subset \mathsf{Z}$, where the set $\mathcal{C}$ is called the* codebook *and its elements are called the* codevectors.

The source is a random vector $Z \in \mathbb{R}^d$ with some probability distribution P_Z. A given k-point quantizer q represents Z by the quantized output $\widehat{Z} = q(Z)$. Since $q(Z)$ can take only k possible values, it is possible to represent it uniquely by a binary string of $\lceil \log_2 k \rceil$ bits. The number

$$R(q) \triangleq \lceil \log_2 k \rceil$$

is called the *rate* of q (in bits), where we follow standard practice and ignore the integer constraint on the length of the binary representation. The rate is often normalized by the dimension d to give $r(q) = d^{-1} R(q)$ (measured in bits per coordinate); however, since we assume d fixed, there is no need to worry about the normalization. The fidelity of q in representing $Z \sim P_Z$ is measured by the *expected distortion*

$$D(P_Z, q) \triangleq \mathbb{E}\|Z - q(Z)\|^2 = \int_{\mathbb{R}^d} \|z - q(z)\|^2 P_Z(dz).$$

We will assume throughout that Z has finite second moment, $\mathbb{E}\|Z\|^2 < \infty$, so that $D(P_Z, q) < \infty$.

The main objective in vector quantization is to minimize the expected distortion subject to a constraint on the rate (or, equivalently, on the codebook size). Thus, if we denote by $\mathcal{Q}_k$ the set of all k-point vector quantizers, then the optimal performance on a given source distribution P_Z is defined by

$$D_k^*(P_Z) \triangleq \inf_{q \in \mathcal{Q}_k} D(P_Z, q) \equiv \inf_{q \in \mathcal{Q}_k} \mathbb{E}\|Z - q(Z)\|^2. \tag{1}$$

Definition 2. *We say that a quantizer $q^* \in \mathcal{Q}_k$ is* optimal *for P_Z if*

$$D(P_Z, q^*) = D_k^*(P_Z).$$

As we will soon see, it turns out that an optimal quantizer always exists — in other words, the infimum in (1) is actually a minimum — and it can always be chosen to have a particularly useful structural property:

Definition 3. *A quantizer $q \in \mathcal{Q}_k$ with codebook $\mathcal{C} = \{y_1, \ldots, y_k\}$ is called* nearest-neighbor *if, for all $z \in \mathsf{Z}$,*

$$\|z - q(z)\|^2 = \min_{1 \le j \le k} \|z - y_j\|^2.$$

Let $\mathcal{Q}_k^{\mathrm{NN}}$ denote the set of all k-point nearest-neighbor quantizers. We have the following simple but important result:

Lemma 1. *For any $q \in \mathcal{Q}_k$ we can always find some $q' \in \mathcal{Q}_k^{\mathrm{NN}}$, such that $D(P_Z, q') \le D(P_Z, q)$.*

Proof. Given a quantizer $q \in \mathcal{Q}_k$ with codebook $\mathcal{C} = \{y_1, \ldots, y_k\}$, define q' by

$$q'(z) \triangleq \operatorname*{arg\,min}_{y_j \in \mathcal{C}} \|z - y_j\|^2,$$

where ties are broken by going with the lowest index. Then q' is clearly a nearest-neighbor quantizer, and

$$\begin{aligned} D(P_Z, q') &= \mathbb{E}\|Z - q'(Z)\|^2 \\ &= \mathbb{E}\left[\min_{1 \le j \le k} \|Z - y_j\|^2\right] \\ &\le \mathbb{E}\|Z - q(Z)\|^2 \\ &\equiv D(P_Z, q). \end{aligned}$$

The lemma is proved. □

In light of this lemma, we can rewrite (1) as

$$D_k^*(P_Z) = \inf_{q \in \mathcal{Q}_k^{\mathrm{NN}}} \mathbb{E}\|Z - q(Z)\|^2 = \inf_{\mathcal{C} = \{y_1, \ldots, y_k\} \subset \mathsf{Z}} \mathbb{E}\left[\min_{1 \le j \le k} \|Z - y_j\|^2\right]. \tag{2}$$

An important result due to Pollard [Pol82], which we state here without proof, then says the following:

Theorem 1. *If Z has a finite second moment, $\mathbb{E}\|Z\|^2 < \infty$, then there exists a nearest-neighbor quantizer $q^* \in \mathcal{Q}_k^{\mathrm{NN}}$ such that $D(P_Z, q^*) = D_k^*(P_Z)$.*

3. Learning an optimal quantizer

Unfortunately, finding an optimal q^* is a very difficult problem. Indeed, the optimization problem in (2) has a *combinatorial search* component to it, since we have to optimize over all k-point sets $\mathcal{C}$ in $\mathbb{R}^d$. Moreover, the source distribution P_Z is often not known exactly, especially for very complex sources, such as natural images. For these reasons, we have to resort to *empirical* methods for quantizer design, which rely on the availability of a large number of independent samples from the source distribution of interest.

Assuming that such samples are easily available, we can formulate the empirical quantizer design problem as follows. Let us fix the desired codebook size k. For each $n \in \mathbb{N}$, let $Z^n = (Z_1, \dots, Z_n)$ be an i.i.d. sample from P_Z. We seek an algorithm that would take Z^n and produce a quantizer $\widehat{q}_n \in \mathcal{Q}_k$ that would approximate, as closely as possible, an optimal quantizer $q^* \in \mathcal{Q}_k$ that achieves $D^*_k(P_Z)$. In other words, we hope to *learn* an (approximately) optimal quantizer for P_Z based on a sufficiently long training sample.

The first thing to note is that the theory of quantization outlined in the preceding section applies to the *empirical distribution* of the training sample Z^n,

$$P_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}.$$

In particular, given a quantizer $q \in \mathcal{Q}_k$, we can compute its expected distortion

$$D(P_n, q) = \mathbb{E}_{P_n}\|Z - q(Z)\|^2 = \frac{1}{n}\sum_{i=1}^n \|Z_i - q(Z_i)\|^2.$$

Moreover, the minimum achievable distortion is given by

$$D^*_k(P_n) = \min_{q \in \mathcal{Q}_k} \frac{1}{n}\sum_{i=1}^n \|Z_i - q(Z_i)\|^2 = \min_{q \in \mathcal{Q}_k^{\mathrm{NN}}} \|Z_i - q(Z_i)\|^2.$$

Note that we have replaced the infimum with the minimum, since an optimal quantizer always exists and can be assumed to have the nearest-neighbor property. Moreover, since P_n is a discrete distribution, the existence of an optimal nearest-neighbor quantizer can be proved directly, without recourse to Pollard's theorem. Thus, we can restrict our attention to nearest-neighbor k-point quantizers.

Definition 4. *We say that a quantizer $\widehat{q}_n \in \mathcal{Q}_k^{\mathrm{NN}}$ is* empirically optimal *for Z^n if*

$$D(P_n, \widehat{q}_n) = D^*_k(P_n) = \min_{q \in \mathcal{Q}_k^{\mathrm{NN}}} D(P_n, q) = \min_{q \in \mathcal{Q}_k^{\mathrm{NN}}} \frac{1}{n}\sum_{i=1}^n \|Z_i - q(Z_i)\|^2.$$

Note that, by the nearest-neighbor property,

$$D^*_k(P_n) = \min_{\mathcal{C} = \{y_1, \dots, y_k\} \subset \mathsf{Z}} \frac{1}{n}\sum_{i=1}^n \min_{1 \le j \le k} \|Z_i - y_j\|^2.$$

Thus, let $\widehat{q}_n \in \mathcal{Q}_k^{\mathrm{NN}}$ be an empirically optimal nearest-neighbor quantizer. Let $Z \sim P_Z$ be a new source realization, independent of the training data Z^n. If we apply $\widehat{q}_n$ to Z, the resulting quantized output $\widehat{Z} = \widehat{q}_n(Z)$ will depend on both the input Z and on the training data Z^n. Moreover, the expected distortion of $\widehat{q}_n$, given by

$$D(P_Z, \widehat{q}_n) = \mathbb{E}\Big[\|Z - \widehat{q}_n(Z)\|^2 \Big| Z^n\Big] = \int_{\mathsf{Z}} \|z - \widehat{q}_n(z)\|^2 P_Z(dz),$$

is a random variable, since it depends (through $\widehat{q}_n$) on the training data Z^n. In the next section we will show that, under certain assumptions on the source P_Z, the empirically optimal quantizer $\widehat{q}_n$ is nearly optimal on P_Z as well, in the sense that

$$\mathbb{E}\left[D(P_Z, \widehat{f}_n) - D_k^*(P_Z)\right] \le \frac{C}{\sqrt{n}}, \tag{3}$$

where the expectation is w.r.t. the distribution Z^n and $C > 0$ is some constant that depends on d, k, and a certain characteristic of P_Z. More generally, it is possible to show that empirically optimal quantizers are *strongly consistent* in the sense that

$$D(P_Z, \widehat{q}_n) - D_k^*(P_Z) \xrightarrow{n \to \infty} 0 \qquad \text{almost surely}$$

provided the source P_Z has a finite second moment (see Linder's survey [Lin01] for details).

Remark 1. It should be pointed out that the problem of finding an *exact* minimizer of $D(P_n, q)$ over $q \in \mathcal{Q}_k^{\mathrm{NN}}$ is NP-complete. Instead, various approximation techniques are used. The most popular one is the Lloyd algorithm, known in the computer science community as the method of k-means. There, one starts with an initial codebook $\mathcal{C}^{(0)} = \{y_1^{(0)}, \dots, y_k^{(0)}\}$ and then iteratively recomputes the quantizer partition and the new codevectors until convergence.

4. Finite sample bound for empirically optimal quantizers

In this section, we will show how the VC theory can be used to establish (3) for any source supported on a ball of finite radius. This result was proved by Linder, Lugosi and Zeger [LLZ94], and since then refined and extended by multiple authors. Some recent works even remove the requirement that Z be finite-dimensional and consider more general coding schemes in Hilbert spaces [MP10].

For a given $r > 0$ and $z \in \mathbb{R}^d$, let $B_r(z)$ denote the ℓ_2 ball of radius r centered at z:

$$B_r(z) \triangleq \left\{ y \in \mathbb{R}^d : \|y - z\| \le r \right\}.$$

Let $\mathcal{P}(r)$ denote the set of all probability distributions P_Z on $\mathsf{Z} = \mathbb{R}^d$, such that

$$P_Z\left(B_r(0)\right) = 1.$$

Here is the main result we will prove in this section:

Theorem 2. *There exists some absolute constant $C > 0$, such that*

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

Here, as before, $\widehat{q}_n$ denotes an empirically optimal quantizer based on an i.i.d. sample Z^n.

Before launching into the proof, we state and prove a useful lemma:

Lemma 2. *Let $\mathcal{Q}_k^{\mathrm{NN}}(r)$ denote the set of all nearest-neighbor k-point quantizers whose codewords lie in $B_r(0)$. Then for any $P_Z \in \mathcal{P}(r)$,*

$$D(P_Z, \widehat{q}_n) - D_k^*(P_Z) \le 2 \sup_{q \in \mathcal{Q}_k^{\mathrm{NN}}(r)} \left| D(P_n, q) - D(P_Z, q) \right|.$$

Proof. Fix P_Z and let $q^* \in \mathcal{Q}_k^{\mathrm{NN}}$ denote an optimal quantizer, i.e., $D(P_Z, q^*) = D_k^*(P_Z)$. Then, using our old trick of adding and subtracting the right empirical quantities, we can write

$$D(P_Z, \widehat{q}_n) - D_k^*(P_Z) = D(P_Z, \widehat{q}_n) - D(P_n, \widehat{q}_n) + D(P_n, \widehat{q}_n) - D(P_n, q^*) + D(P_n, q^*) - D(P_Z, q^*).$$

Since $\widehat{q}_n$ minimizes the empirical distortion $D(P_n, q)$ over all $q \in \mathcal{Q}_k^{\rm NN}$, we have $D(P_n, \widehat{q}_n) \le D(P_n, q^*)$, which leads to

$$D(P_Z, \widehat{q}_n) - D_k^*(P_Z) \le D(P_Z, \widehat{q}_n) - D(P_n, \widehat{q}_n) + D(P_n, q^*) - D(P_Z, q^*). \tag{4}$$

Now, since $B_r(0)$ is a convex set, for any point $y \notin B_r(0)$ we can compute its *projection* y' onto $B_r(0)$, namely $y' = ry/\|y\|$. Then y' is strictly closer to all $z \in B_r(0)$ than y, i.e.,

$$\|z - y'\| < \|z - y\|, \qquad \forall z \in B_r(0).$$

Thus, if we take an arbitrary quantizer $q \in \mathcal{Q}_k$ and replace all of its codevectors outside $B_r(0)$ by their projections, we will obtain another quantizer q', such that $\|z - q'(z)\| \le \|z - q(z)\|$ for all $z \in B_r(0)$. (The $\le$ sign is due to the fact that some of the codevectors of q may already be in $B_r(0)$, so the projection will not affect them). But then for any $P_Z \in \mathcal{P}(r)$ we will have $D(P_Z, q') \le D(P_Z, q)$. Moreover, if Z^n is an i.i.d. sample from P_Z and P_n is the corresponding empirical distribution, then $P_n \in \mathcal{P}(r)$ with probability one. Hence, we can assume that both $\widehat{q}_n$ and q^* have all their codevectors in $B_r(0)$, and therefore from (4) we obtain

$$\begin{aligned} D(P_Z, \widehat{q}_n) - D_k^*(P_Z) &\le D(P_Z, \widehat{q}_n) - D(P_n, \widehat{q}_n) + D(P_n, q^*) - D(P_Z, q^*) \\ &\le |D(P_n, \widehat{q}_n) - D(P_Z, \widehat{q}_n)| + |D(P_Z, q^*) - D(P_n, q^*)| \\ &\le 2 \sup_{q \in \mathcal{Q}_k^{\rm NN}(r)} |D(P_n, q) - D(P_Z, q)|. \end{aligned}$$

This finishes the proof. $\square$

Now we can get down to business:

Proof (of Theorem 2). For a given quantizer $q \in \mathcal{Q}_k^{\rm NN}(r)$, define the function

$$f_q(z) \triangleq \|z - q(z)\|^2,$$

which is just the squared Euclidean distortion between z and $q(z)$. In particular, for any $P \in \mathcal{P}(r)$ the expected distortion $D(P, q)$ is equal to $P(f_q)$. Since $q \in \mathcal{Q}_k^{\rm NN}(r)$, we have $\|q(z)\| \le r$ for all z. Therefore, for any $z \in B_r(0)$ we will have

$$0 \le f_q(z) \le 2\|z\|^2 + 2\|q(z)\|^2 \le 4r^2.$$

Therefore, using the fact that the expectation of any nonnegative random variable U can be written as

$$\mathbb{E}U = \int_0^\infty \mathbb{P}(U > u)du,$$

we can write

$$D(P_Z, q) = P_Z(f_q) = \int_0^{4r^2} P_Z(f_q(Z) > u)du$$

and

$$D(P_n, q) = P_n(f_q) = \int_0^{4r^2} P_n(f_q(Z) > u)du = \int_0^{4r^2} \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{f_q(Z_i) > u\}} du \qquad \text{a.s.}$$

Therefore

$$
\begin{aligned}
&\sup_{q\in\mathcal{Q}_k^{\mathrm{NN}}(r)} |D(P_n,q)-D(P_Z,q)| \\
&\quad = \sup_{q\in\mathcal{Q}_k^{\mathrm{NN}}(r)} |P_n(q)-P_Z(q)| \\
&\quad = \sup_{q\in\mathcal{Q}_k^{\mathrm{NN}}(r)} \left| \int_0^{4r^2} \left(\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{f_q(Z_i)>u\}} - P_Z(f_q(Z)>u) \right) du \right| \\
&\quad \le 4r^2 \sup_{q\in\mathcal{Q}_k^{\mathrm{NN}}(r)} \sup_{0\le u\le 4r^2} \left| \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{f_q(Z_i)>u\}} - P_Z(f_q(Z)>u) \right| \qquad \text{a.s.}
\end{aligned} \tag{5}
$$

where the last step uses the fact that

$$
\int_a^b h(u)du \le |b-a| \sup_{a\le u\le b} |h(u)|.
$$

Now, for a given $q\in\mathcal{Q}_k^{\mathrm{NN}}(r)$ and a given $u>0$ let us define the set

$$
A_{u,q} \triangleq \Big\{ z\in\mathbb{R}^d : f_q(z)>u \Big\},
$$

and let $\mathcal{A}$ denote the class of all such sets: $\mathcal{A} \triangleq \{A_{u,q} : u>0, q\in\mathcal{Q}_k^{\mathrm{NN}}(r)\}$. Then $\mathbf{1}_{\{f_q(z)>u\}} = \mathbf{1}_{\{z\in A_{u,q}\}}$, so from (5) we can write

$$
\sup_{q\in\mathcal{Q}_k^{\mathrm{NN}}(r)} |D(P_n,q)-D(P_Z,q)| \le 4r^2 \sup_{A\in\mathcal{A}} |P_n(A)-P_Z(A)| . \tag{6}
$$

Therefore,

$$
\begin{aligned}
\mathbb{E}\left[D(P_Z,\widehat{q}_n)-D_k^*(P_Z)\right] &\le 2\mathbb{E}\left[\sup_{q\in\mathcal{Q}_k^{\mathrm{NN}}(r)} |D(P_n,q)-D(P_Z,q)| \right] \\
&\le 8r^2\mathbb{E}\left[\sup_{A\in\mathcal{A}} |P_n(A)-P_Z(A)| \right],
\end{aligned}
$$

where the first step follows from Lemma 2 and the second step follows from (6). To finish the proof, we will show that $\mathcal{A}$ is a VC class with $V(\mathcal{A})\le 4k(d+1)\log(k(d+1))$, so that

$$
\mathbb{E}\left[\sup_{A\in\mathcal{A}} |P_n(A)-P_Z(A)| \right] \le C\sqrt{\frac{V(\mathcal{A})}{n}} \le 2C\sqrt{\frac{k(d+1)\log(k(d+1))}{n}}.
$$

In order to bound the VC dimension of $\mathcal{A}$, let us consider a typical set $A_{u,q}$. Let $\{y_1,\ldots,y_k\}$ denote the codevectors of q. Since q is a nearest-neighbor quantizer, a point z will be in $A_{u,q}$ if and only if

$$
f_q(z) = \min_{1\le j\le k} \|z-y_j\|^2 > u,
$$

which is equivalent to

$$
\|z-y_j\| > \sqrt{u}, \qquad \forall 1\le j\le k.
$$

In other words, we can write

$$
A_{u,q} = \bigcap_{j=1}^k B_{\sqrt{u}}(y_j)^c.
$$

Since this can be done for every $u > 0$ and every $q \in \mathcal{Q}_k^{\mathrm{NN}}(r)$, we conclude that the class $\mathcal{A}$ is contained in another class $\tilde{\mathcal{A}}$, defined by

$$\tilde{\mathcal{A}} \triangleq \left\{ \bigcap_{j=1}^{k} B_j^c : B_j \in \mathcal{B}, \forall j \right\},$$

where $\mathcal{B}$ denotes the class of all closed balls in $\mathbb{R}^d$. Therefore, $V(\mathcal{A}) \le V(\tilde{\mathcal{A}})$. To bound $V(\tilde{\mathcal{A}})$, we must examine its shatter coefficients. We will need the following facts[1]:

(1) For any class of sets $\mathcal{M}$, let $\overline{\mathcal{M}}$ denote the class $\{M^c : M \in \mathcal{M}\}$ formed by taking the complements of all sets in $\mathcal{M}$. Then for any n

$$\mathbb{S}_n(\overline{\mathcal{M}}) = \mathbb{S}_n(\mathcal{M}).$$

(2) For any class of sets $\mathcal{N}$, let $\mathcal{N}_k$ denote the class $\{N_1 \cap N_2 \cap \ldots \cap N_k : N_j \in \mathcal{N}, 1 \le j \le k\}$, formed by taking intersections of all possible choices of k sets from $\mathcal{N}$. Then

$$\mathbb{S}_n(\mathcal{N}_k) \le \mathbb{S}_n^k(\mathcal{N}).$$

In the above notation, $\tilde{\mathcal{A}} = (\overline{\mathcal{B}})_k$, so

$$\mathbb{S}_n(\tilde{\mathcal{A}}) \le \mathbb{S}_n^k(\mathcal{B}),$$

where $\mathcal{B}$ is the class of all closed balls in $\mathbb{R}^d$. From the previous lecture we know that $V(\mathcal{B}) = d+1$, and so the Sauer–Shelah lemma gives

$$\mathbb{S}_n(\tilde{\mathcal{A}}) \le \left(\frac{ne}{d+1}\right)^{k(d+1)}, \qquad \text{for } n \ge d+1. \tag{7}$$

We can now upper-bound $V(\tilde{\mathcal{A}})$ by finding an n for which the right-hand side of (7) is less than 2^n. It is easy to check that, for $d \ge 2$, $n = 4k(d+1)\log(k(d+1))$ does the job; for $d = 1$ it's clear that $V(\tilde{\mathcal{A}}) \le 2k$. Thus,

$$V(\mathcal{A}) \le V(\tilde{\mathcal{A}}) \le 4k(d+1)\log(k(d+1)),$$

as claimed. The proof is finished. □

References

[GG92] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression.* Kluwer, 1992.

[GN98] R. M. Gray and D. L. Neuhoff. Quantization. *IEEE Transactions on Information Theory*, 44(6):2325–2383, October 1998.

[Lin01] T. Linder. Learning-theoretic methods in vector quantization. In L. Györfi, editor, *Principles of Nonparametric Learning.* Springer, 2001.

[LLZ94] T. Linder, G. Lugosi, and K. Zeger. Rates of convergence in the source coding theorem, in empirical quantizer design, and in universal lossy source coding. *IEEE Transactions on Information Theory*, 40(6):1728–1740, November 1994.

[MP10] A. Maurer and M. Pontil. K-dimensional coding schemes in Hilbert spaces. *IEEE Transactions on Information Theory*, 56(11):5839–5846, November 2010.

[Pol82] D. Pollard. Quantization and the method of k-means. *IEEE Transactions on Information Theory*, IT-28(2):199–205, March 1982.

[1] Exercise: prove them!