Published in final edited form as:
IEEE Trans Inf Theory. 2016 May ; 62(5): 2737–2747. doi:10.1109/TIT.2016.2535172.

Compression for Quadratic Similarity Queries: Finite Blocklength and Practical Schemes

Fabian Steiner,
Dept. of Electrical Engineering of the Technische Universität München, Munich, Germany

Steffen Dempfle,
Dept. of Electrical Engineering of the Technische Universität München, Munich, Germany

Amir Ingber, and
Dept. of Electrical Engineering, Stanford University, Stanford, CA 94305. He is now with Yahoo! Labs, Sunnyvale, CA 94089

Tsachy Weissman
Dept. of Electrical Engineering, Stanford University, Stanford, CA 94305

Abstract

We study the problem of compression for the purpose of similarity identification, where similarity is measured by the mean square Euclidean distance between vectors. While the asymptotical fundamental limits of the problem – the minimal compression rate and the error exponent – were found in a previous work, in this paper we focus on the nonasymptotic domain and on practical, implementable schemes.

We first present a finite blocklength achievability bound based on shape-gain quantization: The gain (amplitude) of the vector is compressed via scalar quantization and the shape (the projection on the unit sphere) is quantized using a spherical code. The results are numerically evaluated and they converge to the asymptotic values as predicted by the error exponent. We then give a nonasymptotic lower bound on the performance of any compression scheme, and compare to the upper (achievability) bound. For a practical implementation of such a scheme, we use wrapped spherical codes, studied by Hamkins and Zeger, and use the Leech lattice as an example for an underlying lattice. As a side result, we obtain a bound on the covering angle of any wrapped spherical code, as a function of the covering radius of the underlying lattice.

I. INTRODUCTION

The number of applications dealing with a huge amount of data has increased significantly in recent years. Many of those applications do not only deal with storage of the data, but also with its retrieval and querying. In many cases, the query is whether a given database contains sequences that are similar to a given sequence of interest. The notion of "similarity" depends on the kind of application involved, where notable examples include the Hamming and

---

Parts of this work will be presented at the 2014 IEEE International Symposium on Information Theory (ISIT).

Euclidean distances. The size of the big database motivates the question of how to construct a much smaller (compressed) version of it that will allow to answer queries reliably.

In [1] Ingber et al. develop a general framework for the case of a Gaussian source and Euclidean distance measure and they also provide asymptotic results for the identification rate, i.e. the rate above which any query can be made arbitrarily reliable, and a characterization of the identification exponent associated with it. Results for the case of discrete memoryless sources are given in [2].

In the present work, we follow the framework described in [1] and extend it to the finite blocklength case. We begin by deriving a nonasymptotic achievability bound on the reliability, using shape-gain quantizers [3]. In such systems, the gain (amplitude) of the vector is compressed via scalar quantization and the shape (the projection on the unit sphere) is quantized using a spherical code. While in [1] the asymptotics of the setting allow crude scalar quantizers, here we optimize the quantizers for the distribution of the source. Combined with a (nonconstructive) result on the covering of spherical shells [4], the performance of the system can be evaluated numerically at any finite blocklength n. The numerical result validates the asymptotic approximations for the performance predicted by the error exponent of [1]. The achievability result is complemented by a lower bound on the performance at finite blocklength. The lower bound is derived following the approach in [1], but with greater attention to detail and optimization of the different parameters involved in the derivation.

In addition to the (non-constructive) achievability result, we develop a general method of constructing implementable compression schemes, which are also based on the shape-gain framework. While the gain quantizer of the achievability bound can be easily implemented, the shape quantizer (the spherical code) is not. For that purpose, we utilize *wrapped* spherical codes which were previously introduced in [5]. The shape codebook is obtained by considering a mapping which wraps an $n - 1$-dimensional lattice around the shell of the n-dimensional unit sphere. Any lattice can be used for this process and its covering radius defines the performance of the scheme. As part of the analysis of the scheme, we derive a bound on the covering angle of any wrapped spherical code (as a function of the properties of the underlying lattice), a result that may be of independent interest.

The rest of this paper is organized as follows. The next subsections introduce terms and definitions that are used throughout the paper. Section II presents the achievabilty results, whereas Section III is dedicated to the converse result. Section IV describes an actual, implementable scheme that can be used in practice, along with numerical results. We will provide some concluding remarks and possible further research objectives in Section V.

A. Problem Setting

The goal of the framework presented in [1] is to answer similarity queries from a compressed representation of the data. More specifically, for each sequence $\mathbf{x}$ in the database, we only keep a compressed signature $Q(\mathbf{x})$. The final goal is to be able to detect whether $\mathbf{x}$ is similar to a query sequence $\mathbf{y}$, given only $Q(\mathbf{x})$ and $\mathbf{y}$.

Concerning the nature of the answer "yes/no" of the setup depicted in Figure 1, the possible errors are either false positives or false negatives. While the first event is not considered catastrophic, as it only results in additional efforts when the answer of the original query has to be confirmed with the actual database entry in addition to its compressed version, the incident of false negatives can not be detected: Many practical applications, e.g. querying a criminal forensic database, obviously need to exclude this kind of error.

Therefore, we impose the restriction to our model that *false negatives are not permitted*. Basically, this means that the result of the query function is either " `no`" or " `maybe`", where the latter pertains to the cases of being either actually similar or false positive.

We focus on a similarity measure defined by the normalized squared Euclidean distance. To this end, for any length-n real sequences $\mathbf{x} = (x_1, x_2, \ldots, x_n)^T$, $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T \in \mathbb{R}^n$ define

$$d(\mathbf{x},\mathbf{y}) \triangleq \frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 = \frac{1}{n}\|\mathbf{x} - \mathbf{y}\|^2, \qquad (1)$$

where $\|\cdot\|$ denotes the standard Euclidean norm. We say that $\mathbf{x}$ and $\mathbf{y}$ are *D-similar* when $d(\mathbf{x}, \mathbf{y}) \le D$, or simply *similar* when D is clear from the context.

To formalize the previously described problem setting, we define the following (see [1]):

Definition 1—*A rate-R identification system* (Q, g) *consists of a* signature assignment

$$Q : \mathbb{R}^n \to \left\{1, 2, \ldots, 2^{nR}\right\} \qquad (2)$$

and a query function

$$g : \left\{1, 2, \ldots, 2^{nR}\right\} \times \mathbb{R}^n \to \{\mathrm{no}, \ \ \mathrm{maybe}\}. \qquad (3)$$

Definition 2—*A system* (Q, g) is said to be D-admissible, if for any $\mathbf{x}$, $\mathbf{y}$ satisfying $d(\mathbf{x}, \mathbf{y}) \le D$, we have

$$g(Q(\mathbf{x}), \mathbf{y}) = \mathrm{maybe}. \qquad (4)$$

Note that, by definition, any D-admissible system (Q, g) can not produce false negatives.

At this point, it is worthwhile to think about the figure of merit that should be considered in our system design. In the spirit of source and channel coding scenarios, where one generally aims at driving the error probability to zero for long blocklengths, we pursue the same idea with the probability of a false positive event $\mathscr{E} = \{g(Q(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}, d(\mathbf{X}, \mathbf{Y}) > D\}$.

Assuming a D-admissible system (Q, g) we can relate this probability to $\Pr\{g(Q(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}\}$ and concern ourselves with the latter quantity instead:

$$\begin{aligned}\Pr\{g(Q(\mathbf{X}), \boldsymbol{Y})&=\text{maybe}=\Pr\{g(Q(\mathbf{X}), \boldsymbol{Y})\\&=\text{maybe}|d(\mathbf{X}, \boldsymbol{Y}) \le D\Pr\{d(\mathbf{X}, \boldsymbol{Y}) \le D\}+\Pr\{g(Q(\mathbf{X}), \boldsymbol{Y})\\&\quad=\text{maybe}, d(\mathbf{X}, \boldsymbol{Y})>D\}\\&=\Pr\{d(\mathbf{X}, \boldsymbol{Y}) \le D\}+\Pr\{\mathscr{E}\},\end{aligned} \tag{5}$$

where (5) follows since $\Pr\{g(Q(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}|d(\mathbf{X}, \mathbf{Y}) \le D\} = 1$ by the D-admissibility of (Q, g). Since $\Pr\{d(\mathbf{X}, \mathbf{Y}) \le D\}$ does not depend on what scheme is employed, minimizing the false positive probability $\Pr\{\mathscr{E}\}$ over all D-admissible schemes (Q, g) is equivalent to minimizing $\Pr\{g(Q(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}\}$. Also note that the only interesting case is when $\Pr\{d(\mathbf{X}, \mathbf{Y}) \le D\} \to 0$ as n grows, since otherwise almost all the sequences in the database will be similar to the query sequence, making the problem degenerate (since almost all the database needs to be retrieved, regardless of the compression). In this case, it is easy to see that $\Pr\{\mathscr{E}\}$ vanishes if and only if the conditional probability

$$\Pr\{g(Q(\mathbf{X}), \boldsymbol{Y})=\text{maybe}|d(\mathbf{X}, \boldsymbol{Y})>D\} \tag{6}$$

vanishes. In view of the above, we henceforth restrict our attention to the behavior of $\Pr\{g(Q(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}\}$.

In analogy to the classical rate-distortion setting [6], [7], we also define:

Definition 3—*For given distributions P_X, P_Y and a similarity threshold D, a rate R is said to be* D-achievable *if there exists a sequence of rate-R admissible schemes* $(Q^{(n)}, g^{(n)})$ *satisfying*

$$\lim_{n\to\infty} \Pr\left\{g^{(n)}\left(Q^{(n)}(\mathbf{X}), \boldsymbol{Y}\right)=\text{maybe}\right\}=0, \tag{7}$$

where $\mathbf{X}$ and $\mathbf{Y}$ are independent i.i.d. sequences with respective marginals P_X and P_Y

Definition 4—*For given distributions P_X, P_Y and a similarity threshold D, the* identification rate $R_{\mathrm{ID}}(D, P_X, P_Y)$ *is the infimum of D-achievable rates. That is,*

$$R_{\mathrm{ID}}(D, P_X, P_Y) \triangleq \inf\{R: R\ is\ D\text{-}\ achievable\}, \tag{8}$$

where an infimum over the empty set is equal to ∞.

One can also define the identification exponent, i.e. the asymptotic slope of the exponential decay of $\Pr\{g(Q(\mathbf{X}), \mathbf{Y}) = \texttt{maybe}\}$:

Definition 5—*Fix* $R \geq R_{\mathrm{ID}}(D, P_X, P_Y)$. *The* identification exponent *is defined as*

$$\mathbf{E}_{\mathrm{ID}}(R, D, P_X, P_Y) \triangleq \lim_{n\to\infty} ; \sup -\frac{1}{n}\log \inf_{g^{(n)},Q^{(n)}} \Pr\left\{g^{(n)}\left(Q^{(n)}(\mathbf{X}), \boldsymbol{Y}\right) = \mathrm{maybe}\right\}, \quad (9)$$

where the infimum is over all D-admissible systems $(g^{(n)}, Q^{(n)})$ *of rate R and blocklength n.*

Note that this quantity gives rise to the approximation $\Pr\{\mathrm{maybe}\} \approx e^{-n\mathbf{E}_{\mathrm{ID}}(R)}$, assuming an approximately optimal scheme is employed, which is valid for large n.

In the following, we will focus on the standard Gaussian case, meaning that the components $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ of the length-n vectors $\mathbf{X}$ and $\mathbf{Y}$ are independent and identically distributed Gaussian random variables with zero mean and unit variance. For this special (but important) case, the identification rate is given by [1, Corollary 1]

$$R_{\mathrm{ID}}(D) = \begin{cases} \log\left(\frac{2}{2-D}\right) & \text{for } 0 \leq D < 2 \\ \infty & \text{for } D \geq 2. \end{cases} \quad (10)$$

The exponent for this case is given by [1, Corollary 2]

$$\mathbf{E}_{\mathrm{ID}}(R, D) = \min_{\rho} \frac{1}{\ln 2}(\rho - 1 - \ln\rho) - \log ; \sin ; \min\left[\frac{\pi}{2}, \left(\arcsin(2^{-R}) + \arccos\frac{2\rho - D}{2\rho}\right)\right] \text{ s. t } 2 \geq 2\rho \geq D. \quad (11)$$

B. Geometry Basics Revisited

For the derivations in the next sections some results from Euclidean geometry are needed. The reason for this is mainly due to the fact that we concentrate on Gaussian random vectors and the distribution of the shape $\mathbf{X}/\|\mathbf{X}\|$ of a Gaussian vector $\mathbf{X}$ is uniform on the shell of the unit sphere in n dimensions. More generally, we define the spherical shell with arbitrary radius $r > 0$ as

$$\mathscr{S}_r^n \triangleq \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\| = r\}. \quad (12)$$

If the index r is omitted for notational brevity, we shall refer to the spherical unit shell. In case the interior should be part of the set as well, we speak of a ball that is usually centered around a point $\mathbf{u} \in \mathbb{R}^n$:

$$\mathscr{B}_r(\mathbf{u}) \triangleq \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{u} - \mathbf{x}\| \leq r\}. \quad (13)$$

The definition of a spherical shell can be extended to a "thick" spherical shell in n dimensions by

$$\mathscr{S}_{r_1,r_2}^n \triangleq \{\mathbf{x} \in \mathbb{R}^n : r_1 \leq \|\mathbf{x}\| \leq r_2\}. \quad (14)$$

The angle between two elements $\mathbf{x}_1$, $\mathbf{x}_2$ can be expressed as:

$$\angle(\mathbf{x}_1, \mathbf{x}_2) \triangleq \arccos\left(\frac{\mathbf{x}_1^T \mathbf{x}_2}{\|\mathbf{x}_1\|\|\mathbf{x}_2\|}\right) \in [0, \pi]. \quad (15)$$

Given a point $\mathbf{u} \in \mathbb{R}^n \setminus \{0\}$ and half angle $\theta \in [0, \pi]$, define:

$$\mathrm{CONE}(\mathbf{u}, \theta) \triangleq \{\mathbf{x} \in \mathbb{R}^n : \angle(\mathbf{x}, \mathbf{u}) \leq \theta\}. \quad (16)$$

The definitions of (12) and (16) now become vital as the intersection of the two describes a spherical cap denoted by $\mathrm{CAP}_r(\mathbf{u}, \theta)$:

$$\mathrm{CAP}_r(\mathbf{u}, \theta) \triangleq \mathscr{S}_r^n \cap \mathrm{CONE}(\mathbf{u}, \theta). \quad (17)$$

Employing the notion of a thick shell in (14), we can also define a thick cap given as:

$$\mathrm{CAP}_{r_1,r_2}(\mathbf{u}, \theta) \triangleq \mathscr{S}_{r_1,r_2}^n \cap \mathrm{CONE}(\mathbf{u}, \theta). \quad (18)$$

When talking about coverings of spherical shells as needed for quantization purposes, we will need to compute their $n - 1$ and n-dimensional contents. According to [8], these are calculated as

$$|\mathscr{S}_r^n| = \frac{2\pi^{\frac{n}{2}} r^{n-1}}{\Gamma\left(\frac{n}{2}\right)}, \quad (19)$$

$$V(\mathscr{S}_r^n) = \frac{\pi^{\frac{n}{2}} r^n}{\Gamma\left(\frac{n+2}{2}\right)}. \quad (20)$$

Note that the fraction of a spherical shell $\mathscr{S}_r^n$ that is covered by a spherical cap $\mathrm{CAP}_r(\mathbf{u}, \theta)$ can be expressed as

$$\Omega(\theta, n) \triangleq \frac{|\mathrm{CAP}_r(\mathbf{u}, \theta)|}{|\mathscr{S}_r^n|} = \frac{1}{2} I_{\sin^2(\theta)}\left(\frac{n-1}{2}, \frac{1}{2}\right), \qquad (21)$$

where the function $I_x(a, b)$ denotes the regularized, incomplete beta function:

$$I_x(a, b) \triangleq \frac{\int_0^x t^a (1-t)^b \mathrm{d}t}{\int_0^1 t^a (1-t)^b \mathrm{d}t}. \qquad (22)$$

We emphasize that equation (21) is solely dependent on the angle θ and the dimension n, but not on the point **u** or the radius r. This fact will facilitate the calculation of quantities of interest in Section II. If the dimension n is clear from the context, we omit the second parameter and simply write $\Omega(\theta)$.

Finally, for $A \subseteq \mathbb{R}^n$ and $D > 0$, we define the D-expansion of A as

$$\Gamma^D(A) \triangleq \left\{\mathbf{y} \in \mathbb{R}^n : \exists_{\mathbf{x} \in A} d(\mathbf{x}, \mathbf{y}) \le D\right\}, \qquad (23)$$

where $d(\mathbf{x}, \mathbf{y})$ was defined in (1).

C. Coverings and Lattices

In this subsection, we introduce the general notion of coverings of a set and then show how these definitons directly apply to coverings of lattices and spherical shells.

1) Coverings—Let $A \subseteq \mathbb{R}^n$, then we say that a set B ρ-covers the set A, if

$$A \subseteq \bigcup_{\mathbf{x} \in B} \mathscr{B}_\rho(\mathbf{x}). \qquad (24)$$

We denote the collection of all sets that ρ-cover the set A by COV(A, ρ). A convenient measure which allows for comparison between different coverings B is provided by their covering density:

$$\zeta(A, B) \triangleq \sum_{\mathbf{x} \in B} \frac{|A \cap \mathscr{B}_\rho(\mathbf{x})|}{|A|}. \qquad (25)$$

A classical task is to look for a covering $B \in$ COV(A, ρ) which results in the smallest density. Formally, it can be found when (25) is minimized over all coverings in COV(A, ρ):

$$\vartheta(A) \triangleq \min_{B \in \mathrm{COV}(A, \rho)} \zeta(A, B). \qquad (26)$$

As we have to quantize the shape of a Gaussian vector, which lies on the shell of the unit sphere, the set A can be replaced by $\mathscr{S}^n$ for our purposes. In this case, the intersection $\mathscr{S}^n \cap \mathscr{B}_\rho(\mathbf{x})$ results in a spherical cap, namely $\mathrm{CAP}_1(\mathbf{u}, \theta)$. Using this for the evaluation of (26), the quantity turns out to read as

$$\vartheta(\mathscr{S}^n) = |B^*| \cdot \Omega(\theta), \quad (27)$$

which states the covering density of a spherical code with covering angle θ and B^* is the minimizer of (26).

In [4, Theorem 1], Dumer showed that $\vartheta(S^n)$ is bounded by

$$\vartheta(\mathscr{S}^n) \le (n-1)\log_2(n-1)\left(\frac{1}{2}+\frac{2\log_2\log_2(n-1)+5}{\log_2(n-1)}\right). \quad (28)$$

We use this result in Section II in order to retrieve an upper bound on the rate R_S of the spherical code.

2) Lattices—A lattice Λ is a set of vectors that is closed under addition, i.e. forms an additive group. It can be defined by a set of basis vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in \mathbb{R}^n$, i.e.

$$\Lambda \triangleq \left\{\mathbf{v} \in \mathbb{R}^n : \mathbf{v} = \sum_{i=1}^{n} c_i \mathbf{v}_i, c_i \in \mathbb{Z}\right\} \quad (29)$$

Generally, these vectors are combined in the generator matrix $\mathbf{M}$ of the lattice Λ:

$$\mathbf{M} = \begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_n \end{pmatrix}. \quad (30)$$

Apart from the generator matrix, another important property is given by its minimum distance d_Λ

$$d_\Lambda = \min_{\substack{\mathbf{v},\mathbf{u}\in\Lambda \\ \mathbf{v}\neq\mathbf{u}}} \|\mathbf{v}-\mathbf{u}\| \quad (31)$$

and covering radius r_Λ^{cov}

$$r_\Lambda^{\mathrm{cov}} = \sup_{\mathbf{x}\in\mathbb{R}^n} \inf_{\mathbf{v}\in\Lambda} \|\mathbf{x}-\mathbf{v}\|. \quad (32)$$

While the minimum distance d_Λ is more important for channel coding applications (as it always ensures a certain distance between two points in the lattice), the latter quantity is of special interest for source coding problems. It is reasonable to define the packing radius as $r_\Lambda^{\text{pack}} = d_\Lambda/2$, i.e. the largest radius balls that are centered around any lattice point can have so that they do not overlap. Obviously, it always holds that $r_\Lambda^{\text{pack}} \leq r_\Lambda^{\text{cov}}$.

For a general lattice the last two quantities can not be easily determined, but require a deep geometrical understanding of the lattice. However, Sloane [9] has compiled a detailed comparison of the many important lattices used in practice, and some approximative, numerical approaches can be found in the literature [10], [11, Chapter 2, 1.4] as well.

Using **M** as a description of a lattice has many advantages, as several important properties can be calculated easily. One important operation that can be performed on a lattice involves the rescaling of its basis vectors $\mathbf{v}_i$, $i \in [1 : n]$. A comprehensive survey of the effects of such a scaling is given in [9]. Assuming that $\Lambda' = s\Lambda$, $s \in \mathbb{R}$ is the scaled lattice, we particularly note that because of (31) the minimum distance of the new lattice scales with the same factor s, i.e. $d_{\Lambda'} = s d_\Lambda$ We exploit this property in Section IV to adapt the rate of the wrapped spherical code quantizer.

II. ACHIEVABILITY

A. Proposed Achievability Scheme

We have pointed out in Section I-A that we wish to minimize Pr$\{g(Q(\mathbf{X}), \mathbf{Y}) =$ `maybe`$\}$, or Pr $\{$ `maybe`$\}$ for short, as it can be regarded as a performance measure of our scheme. In order for a scheme to be admissible according to Definition 2, we must answer `maybe` whenever $\mathbf{Y} \in \Gamma^D(Q^{-1}(Q(\mathbf{X})))$, where we define the set of all the points that have a signature equal to i as $Q^{-1}(i) \triangleq \{\mathbf{x} \in \mathbb{R}^n : Q(\mathbf{x}) = i\}$. Evidently, the corresponding probability can be written formally as

$$\Pr\{\text{maybe}\} = \sum_{i \in [1:2^{nR}]} \Pr\{Q(\mathbf{X}) = i\} \Pr\left\{\mathbf{Y} \in \Gamma^D\left(Q^{-1}(i)\right)\right\}. \tag{33}$$

Analyzing this quantity turns out to be a difficult task, when no further structure or knowledge about the compression scheme $Q(\cdot)$ is available. For that purpose, we shall construct $Q(\cdot)$ as a shape-gain quantizer [3].

Shape-gain vector quantizers can be understood as a special implementation of product quantizers. The decomposition of the random vector **X** is obtained by splitting it into its shape $\mathbf{S} = \mathbf{X}/\|\mathbf{X}\|$ and gain (amplitude) $G = \|\mathbf{X}\|$

For our case of a Gaussian random vector **X**, the random variable G is a scalar value that follows a χ-distribution [12] with n degrees of freedom and possesses the probability density function

$$f_G(r)=f_{\|\mathbf{x}\|}(r)=\frac{2^{1-\frac{n}{2}}r^{n-1}\mathrm{e}^{-\frac{r^2}{2}}}{\Gamma\left(\frac{n}{2}\right)}, \qquad (34)$$

where $\Gamma(n)$ denotes the usual gamma function: $\Gamma(n)=\int_0^\infty t^{(n-1)}\mathrm{e}^{-t}\mathrm{d}t$. It is quantized via $Q_G:r\in\mathbb{R}_0^+\to[1:2^{nR_G}]$, which can efficiently be realized by the Lloyd-Max algorithm [13], [14][1]. We denote the boundaries of the quantization intervals as $[r_{k-1}, r_k]$, $k\in[1:2^{nR_G}]$.

We quantize the shape $\mathbf{S}$ independently from the gain via a spherical code $Q_S:\mathscr{S}^n\to[1:2^{nR_S}]$ and obtain the shape codebook $\mathscr{C}_S$: As $\mathbf{S}$ is now an element of the shell of the unit sphere, it is easier to quantize than the original random variable. In particular, because of the Gaussian assumption on $\mathbf{X}$, the shape $\mathbf{S}$ is uniformly distributed on the shell of the unit sphere (cf. (65)), which motivates its quantization with a spherical code, which can be implemented, for example, by wrapping lattices in $\mathbb{R}^{n-1}$ around the spherical shell in n dimensions [15], a path we pursue in Section IV.

It is important to stress that we do not care about the exact reconstruction $\hat{\mathbf{X}}=\hat{G}\hat{\mathbf{S}}$ for which the knowledge about the associated codebooks is essential, but are only interested in how close our query is to any point in the quantization cell. Nevertheless, we use the notation $\hat{\mathbf{s}}$ to refer to the center of a spherical cap that contains the quantization cell.

In order to prove the asymptotic achievability results, involving the identification rate, [1] shows that it is sufficient to neglect the gain quantizer and to concentrate on the "typical gain", as for high dimension n the probability density function of $\mathbf{X}$ concentrates near a hyperspherical shell $\mathscr{S}^n_{r_X^-,r_X^+}$ with $r_X^{\pm}=\sqrt{n(\sigma_X^2\pm\eta)}$ and η being arbitrarily small. In the nonasymptotic domain, we can no longer rely on this fact.

B. Analysis of the Proposed Scheme

In (33) we pointed out that the definition of our achievability scheme requires to answer `maybe` whenever $\mathbf{Y}\in\Gamma^D(Q^{-1}(Q(\mathbf{X})))$. We can now find an upper bound for this probability by taking the structure of a shape-gain quantizer into account. As before, we define the suggestive sets $Q_G^{-1}(k)\triangleq\left\{r\in\mathbb{R}_0^+:Q_G(r)=k\right\}=[r_{k-1},r_k]$ and $Q_S^{-1}(l)\triangleq\{\mathbf{s}\in\mathscr{S}^n:Q_S(\mathbf{s})=l\}$.

Note that it is trivial to embed both k and l in a single integer i. Therefore, for a shape-gain quantizer, we have

$$Q^{-1}(i)\triangleq\left\{\mathbb{R}^n\ni\mathbf{x}=r\cdot\mathbf{s}\,\middle|\,r\in Q_G^{-1}(k),\mathbf{s}\in Q_S^{-1}(l)\right\}. \qquad (35)$$

---

[1] While the Lloyd-Max is known to be optimal in some cases for MSE quantization, we have no optimality guarantee when used for compression for similarity identification. Nevertheless, as shown in the next sections, the performance is very good.

At this point, we only allow shape codebooks $\mathscr{C}_S$ that come with a guaranteed upper bound on the covering angle $\angle(\mathbf{S}, \hat{\mathbf{S}}) \leq \theta$. Consequently, we conclude that the set $Q^{-1}(i)$ is contained within a thick cap, i.e. $Q^{-1}(i) \subseteq \mathrm{CAP}_{r_{k-1},r_k}(\hat{\mathbf{s}}_l, \theta)$. This fact simplifies the analysis and gives rise to an easy implementation of an admissible decision rule, i.e. the second factor in (33) can be written more explicitly by checking for $\mathbf{y} \in \Gamma^D(\mathrm{CAP}_{r_{k-1},r_k}(\hat{\mathbf{s}}_l, \theta))$.

Hence, the upper bound on Pr {maybe} is given by

$$\Pr\{\text{maybe}\} \leq \sum_{k\in[1:2^{nR_G}]} \sum_{l\in[1:2^{nR_S}]} \Pr\{Q_G(\|\mathbf{X}\|)=k\} \Pr\{Q_S(\mathbf{S})=l\} \Pr\left\{\boldsymbol{Y} \in \Gamma^D(\mathrm{CAP}_{r_{k-1},r_k}(\hat{\mathbf{s}}_l, \theta))\right\}. \tag{36}$$

The propositions that follow are geared toward obtaining simpler expressions for the last factor in (36). The expression will turn out not to depend on a specific codepoint $\hat{\mathbf{s}}_l$, so that we can also write with any $\hat{\mathbf{s}} \in \mathscr{C}_S$:

$$\Pr\{\text{maybe}\} \leq \sum_{k\in[1:2^{nR_G}]} \Pr\{\|\mathbf{X}\| \in [r_{k-1}, r_k]\} \Pr\left\{\boldsymbol{Y} \in \Gamma^D(\mathrm{CAP}_{r_{k-1},r_k}(\hat{\mathbf{s}}, \theta))\right\}. \tag{37}$$

Before calculating the probability of **Y** falling into the expansion of a thick cap as suggested by (37), we approach this problem by first assuming that the gain quantization is a trivial mapping that maps the gain to one single value *r*.

Proposition 1—*The probability of the random variable* $\mathbf{Y}$ *falling into the D-expansion of a* thin *spherical cap* $\mathrm{CAP}_r(\hat{\mathbf{s}}, \theta)$ *is given by*

$$\Pr\{\boldsymbol{Y} \in \Gamma^D(\mathrm{CAP}_r(\hat{\mathbf{s}}, \theta))\}=\int_0^\infty \Pr\left\{\boldsymbol{Y} \in \Gamma^D(\mathrm{CAP}_r(\hat{\mathbf{s}}, \theta)) | \|\boldsymbol{Y}\|=r_Y\right\} \cdot f_{\|\boldsymbol{Y}\|}(r_Y) \mathrm{d}r_Y \tag{38}$$

and

$$\Pr\left\{\boldsymbol{Y} \in \Gamma^D(\mathrm{CAP}_r(\hat{\mathbf{s}}, \theta)) | \|\boldsymbol{Y}\|=r_Y\right\}=\begin{cases} 0, & |r-r_Y|>\sqrt{nD} \\ 1, & r_Y \leq r_{Y,deg}(r,\theta) \\ \Omega(\theta+\theta'(r_Y)), & otherwise. \end{cases} \tag{39}$$

The quantity $r_{Y,deg}(r, \theta)$ *is given by*

$$r_{Y,deg}(r,\theta)=\sqrt{(r\,;\cos(\theta))^2-r^2+nD}-r\,;\cos(\theta) \tag{40}$$

and $\theta^{'}(r_Y)$ *by*

$$\theta^{'}(r_Y)=\arccos\left(\frac{r^2+r_Y^2-nD}{2\cdot r\cdot r_Y}\right). \qquad (41)$$

Proof: See Appendix A.

Using similar techniques, the analysis can be extended to a thick spherical cap as follows.

Proposition 2—*The probability of the random variable* $\mathbf{Y}$ *falling into the D-expansion of a* thick *spherical cap* $\mathrm{CAP}_{r_1,r_2}(\hat{\mathbf{s}},\theta)$ *is given by*

$$\Pr\{\boldsymbol{Y}\in\Gamma^D(\mathrm{CAP}_{r_1,r_2}(\hat{\mathbf{s}},\theta))\}=\int_0^\infty \Pr\left\{\boldsymbol{Y}\in\Gamma^D(\mathrm{CAP}_{r_1,r_2}(\hat{\mathbf{s}},\theta))|\|\boldsymbol{Y}\|=r_Y\right\}\cdot f_{\|\boldsymbol{Y}\|}(r_Y)\mathrm{d}r_Y \qquad (42)$$

and

$$\Pr\{\boldsymbol{Y}\in\Gamma^D(\mathrm{CAP}_{r_1,r_2}(\hat{\mathbf{s}},\theta))|\|\boldsymbol{Y}\|=r_Y\}=\begin{cases}0, & r_Y<r_1-\sqrt{nD}\ or\\ & r_Y>r_2+\sqrt{nD}\\ 1, & r_Y\le r^{'}_{Y,deg}(r_1,\theta)\\ \Omega(\theta+\theta^{''}(r_1,r_2,r_Y)), & otherwise.\end{cases} \qquad (43)$$

The quantity $r^{'}_{Y,deg}(r_1,\theta)$ *is given by*

$$r^{'}_{Y,deg}(r_1,\theta)=\sqrt{(r_1\cos(\theta))^2-r_1^2+nD}-r_1\cos(\theta) \qquad (44)$$

and $\theta^{''}(r_1, r_2, r_Y)$ *by*

$$\theta^{''}(r_1,r_2,r_Y)=\begin{cases}\arccos\left(\frac{r_1^2+r_Y^2-nD}{2\cdot r_1\cdot r_Y}\right), & r_Y\le\sqrt{r_1^2+nD}\\ \arccos\left(\frac{r_2^2+r_Y^2-nD}{2\cdot r_2\cdot r_Y}\right), & r_Y\ge\sqrt{r_2^2+nD}\\ \arcsin\left(\frac{\sqrt{nD}}{r_Y}\right), & otherwise.\end{cases} \qquad (45)$$

Proof: See Appendix B.

Theorem 3 (Finite Blocklength Achievability)—*Let* $R = R_G + R_S$*, where* R_G *and* R_S *denote the rates of the employed gain and shape quantizers. Further, assume that the shape* *quantizer is a spherical code that has a guaranteed covering angle θ at rate R_S. At rate R, the achieved error probability is upper bounded by*

$$\Pr\{\text{maybe}\} \leq \sum_{k\in[1:2^{\text{n}R_G}]} \int_{r_{k-1}}^{r_k} f_{\|\mathbf{X}\|}(r_X)\mathrm{d}r_X \cdot \Pr\left\{\boldsymbol{Y} \in \Gamma^D(\mathrm{CAP}_{r_{k-1},r_k}(\hat{\mathbf{s}},\theta))\right\}. \tag{46}$$

Proof: Theorem 3 directly follows from Proposition 2 and the explanations leading to (37).

Averaging expression (38) with respect to r will turn out to become a vital quantity to evaluate any signature assignment scheme $Q(\mathbf{X})$ based on a shape-gain quantizer, as it allows to evaluate the effect of the shape quantization alone. This is due to the fact that by averaging over all r we assume a genie-aided scenario where the decoder knows $\|\mathbf{X}\|$ exactly.

Theorem 4 (Genie-Aided Finite Blocklength Achievability)—*If the above setting is employed and a genie-aided knowledge about the exact value of the gain is available at the decoder, the probability of the query function returning* `maybe` *is upper bounded by*

$$\Pr\{\text{maybe}\} \leq \int_0^\infty \Pr\left\{\boldsymbol{Y} \in \Gamma^D(\mathrm{CAP}_{r_x}(\hat{\mathbf{s}},\theta))\right\} f_{\|\mathbf{X}\|}(r_X)\mathrm{d}r_X. \tag{47}$$

Proof: Theorem 4 directly follows from Proposition 1 and the introductory explanations to this theorem.

Theorem 4, while not pertaining to a directly implementable scheme, gives a bound on how much we can expect the bound (46) to improve by employing the best possible scalar quantizer for this scenario.

C. Numerical Evaluation of the Integrals

As a prerequisite of Theorem 5, we assume the existence of a spherical code $\mathscr{C}_S$ that provides a guarantee on the covering angle θ at a given rate R_S. As pointed out in Section I-C, we can use Dumer's non-constructive achievability result on the covering density for spherical codes in [4, Theorem 1] for $n \geq 4$, in order to relate a given rate R_S to a covering angle. Combining (27) and (28) we can establish the following relation:

$$R_S = \frac{1}{n}\log_2\left(\frac{\vartheta(\mathscr{S}^n)}{\Omega(\theta)}\right). \tag{48}$$

The overall rate is given as $R = R_S + R_G$, where the rate allocation is performed such that for a given R_S, we search for the best R_G within a discrete set of reasonable values.

In Figure 4, the solid curves depict the numerical evaluation of the 3D-integral of (46) for different dimensions n and a desired similarity threshold of $D = 0.1$. Besides, we added the dotted curves that can be obtained by using the identification exponent $\mathbf{E}_{\text{ID}}(R)$ of [1, Theorem 2]. As expected, we see a convergence for both curves for increasing *n*, as $\mathbf{E}_{\text{ID}}(R)$ was derived for the asymptotic case of infinite blocklengths. Surprisingly, a good approximation can already be achieved for relatively small blocklengths of $n = 500$ or $n = 1000$.

Another remark should be spent on the comparison of the solid and dashed curves of the genie-aided scheme imposed by Theorem 4: As pointed out before, the genied-aided curves depict the best performance which can be hoped for with an optimal gain quantizer within our shape-gain quantization framework and provide an impression how much would be gained if such a perfect gain quantizer were found. Since the gap between our MSE-cost criterion based scalar quantizer and the genie-aided curve is negligible, we stick to the Lloyd-Max approach.

III. CONVERSE

A. Derivation of A Lower Bound

Beyond the general achievability that has been shown in the previous section, we are interested in a converse that provides a lower bound to the probability of `maybe`. The derivations in this section closely follow the spirit of the converse in [1, Section IV.C], but put special emphasis on the details of the involved optimization. Theorem 5 summarizes the main result.

Theorem 5—*Let (Q, g) be a rate R compression scheme for a similarity threshold D. For* $\eta > 0$, *define* $D' \triangleq \left(\sqrt{D}+\sqrt{1-\eta}-1\right)^2$ *and* $D'' \triangleq \left(\sqrt{D'}+\sqrt{1-\eta}-1\right)^2$. *Then, for any such* η *that ensures* $D > D' > D''$, *we have the following lower bound on* $\Pr\{\text{maybe}\}$:

$$\Pr\{\text{maybe}\} \geq \max_{\substack{c,\Omega^*,\eta \\ \mathbf{s.t.}\ 0<c<1}} c \cdot \Omega^* \int_{\sqrt{n(1-\eta)}}^{\sqrt{n(1+\eta)}} f_{\|X\|}(r_X)\mathrm{d}r_X \cdot \int_{\sqrt{n(1-\eta)}}^{\sqrt{n(1+\eta)}} f_{\|Y\|}(r_Y)\mathrm{d}r_Y \quad (49)$$

$$0<\Omega^*<1 \quad (50)$$

$$\Omega\left(\theta_{D''}+\Omega^{-1}(p^*)\right)=\Omega^* \quad (51)$$

$$R \leq \frac{1}{n}\log_2\left(\frac{1-c}{p^*}\right) \quad (52)$$

where

$$\theta_{D''} \triangleq \arccos\left(\frac{2 - D''}{2}\right). \quad (53)$$

Proof: See Appendix D.

B. Numerical Evaluation

Figure 5 depicts the numerical results of Theorem 5 (dotted curves) and thereby allows for a comparison to the achievability results (solid curves) already presented. Obviously, both bounds are not tight, but clearly become closer with increasing blocklength: For a blocklength of $n = 25$ and Pr {maybe} = 10^{-5}, the corresponding rate lies within the interval of [0.75; 1.8], whereas the interval shrinks to [0.3; 0.6] for the blocklength $n = 100$.

IV. PRACTICAL IMPLEMENTATION OF A SPHERICAL CODE

A. Introduction

The proposed achievability scheme in Sec. II relies on the existence of a spherical code with an upper bound on the covering angle. However, the previous numerical results in Figure 4 were all based on a non-constructive covering density result by Dumer (28) which provided little insight into how such a spherical code can actually be implemented practically. For this purpose, we look for implementable spherical codes, such as those that were introduced in 1996 by Hamkins in his PhD thesis [15] and subsequent work [5] [16]. In this work, he surveys several different approaches based on lattices and establishes their optimality[2] in an asymptotic sense.

For our purpose, we focus on wrapped spherical codes that are based on the fact that for any given dimension n we can find a lattice Λ in $\mathbb{R}^{n-1}$ which is then “wrapped” onto a spherical shell $\mathscr{S}^n$ in $\mathbb{R}^n$ to obtain the codepoints. The properties of the underlying lattice, particularly its covering density (26), then determine the performance of the shape quantization. For each dimension n, usually several different lattices can be found that exhibit different covering densities. One of the densest lattices is given for $n = 24$ and known as the *Leech Lattice* (cf. Appendix C). Because of its excellent properties, it serves as our model lattice in the following section. At the same time, we emphasize that all arguments hold irrespective of the dimension or employed lattice.

So far, wrapped spherical codes have primarily been used for Gaussian source coding. In [16], Hamkins and Zeger put special emphasis on the asymptotic case for high rates (which is equivalent to a small covering radius r_Λ^{cov}) and could show that the distance between two points on the lattice will be preserved on the spherical shell for asymptotically small d_Λ [15, Lemma 4.2].

---

[2] Optimality hereby refers to the fact that the covering density of the spherical code approaches the covering density of the underlying lattice if d_Λ approaches zero.

In order for a scheme to be practical, we require that both the encoding stage (the mapping $Q(\cdot)$) and the decision functions be implementable. In light of this, any spherical code that is also implementable serves as a candidate for a practical compression-for-queries scheme. The non-trivial part is the decision function $g(\cdot)$. In Sec. II, we used a spherical code with a global guarantee for the covering angle. Therefore the scheme could be admissible since the expanded quantization cells are contained in expanded spherical caps, which are highly structured. Wrapped spherical codes, while implementable rather easily, do not come with a guarantee on the resulting covering angle. For this purpose, we also provide a new derivation of an upper bound on the covering angle $\angle(\mathbf{S}, \hat{\mathbf{S}})$ of any wrapped spherical code, a result that may be of independent interest.

B. Shape Quantization

The implementation of the spherical code that is used for the shape quantization uses a wrapped spherical code and consists of three main steps. First, the input vector, which resides on an n-dimensional shell of the unit sphere $\mathscr{S}^n$ in $\mathbb{R}^n$, is mapped to the Euclidean space in $\mathbb{R}^{n-1}$ by using an appropriate mapping function $h(\cdot)$. Second, the obtained point is quantized to the nearest point of a lattice Λ in $\mathbb{R}^{n-1}$ using a nearest neighbor search algorithm. Lastly, the quantized point is mapped back onto the spherical shell in $\mathbb{R}^n$ by using an inverse mapping function $h^{-1}(\cdot)$ which results in the quantized output vector shown in Figure 6.

As a prerequisite, we partition the surface of $\mathscr{S}^n$ in several regions which are called *annuli*. Further, let the latitude of a point $\mathbf{s}=(s_1, s_2, \ldots, s_n)^T \in \mathscr{S}^n$ be defined as $\arcsin(s_n)$ and $-\pi/2 = \alpha_0 < \ldots < 0 < \ldots < \alpha_N = \pi/2$ is a sequence of latitudes, where $N=\lceil \frac{\pi}{\sqrt{d_\Lambda}} \rceil + \left(\lceil \frac{\pi}{\sqrt{d_\Lambda}} \rceil \pmod 2 \right)$ and $\alpha_i=\pi(\frac{i}{N}-\frac{1}{2})$. The modulo operation is used to ensure that one annulus boundary has latitude 0. Then, the i-th annulus is defined as

$$A_i=\left\{\mathbf{s}\in\mathscr{S}^n : \alpha_i \le \arcsin(s_n) < \alpha_{i+1}\right\}, \qquad (54)$$

for $0 \le i < N$.

1) Mapping Function—The function $h(\cdot)$ is defined in a different manner for each annulus. For $\mathbf{s} \in A_i \subseteq \mathscr{S}^n$, the mapping, now denoted $h_i(s)$, is defined as

$$h_i(\mathbf{s})=\frac{\mathbf{s}'}{\|\mathbf{s}'\|}\left(\|P_i(\mathbf{s}')\| - \|P_i(\mathbf{s})-\mathbf{s}\|\right)_+, \qquad (55)$$

where $(\mathbf{x})_+ \equiv \max(\mathbf{0}, \mathbf{x})$ and the prime notation simply means the deletion of the last coordinate, i.e. $\mathbf{s}' = (s_1, \ldots, s_{n-1})$. The point $P_i(\mathbf{s})$ is given as the solution to the optimization problem

$$P_i(\mathbf{s}){=}\arg\min_{\mathbf{z}}\|\mathbf{s}{-}\mathbf{z}\|,\quad \text{s.t. } \mathbf{z}{=}\begin{cases}(z_1,z_2,\ldots,z_n{=}\sin(\alpha_i))\in\mathscr{S}^n & \text{if } \alpha_i\geq 0\\ (z_1,z_2,\ldots,z_n{=}\sin(\alpha_{i+1}))\in\mathscr{S}^n & \text{if } \alpha_i{<}0\end{cases} \tag{56}$$

Remark 1: For facilitating the calculation of the upper bound on the angle, we further adapt the mapping in (55) such that

$$h_i(\mathbf{s}){=}\mathbf{0},\ \text{ if } \arcsin(s_n)\geq\alpha_{N-1}{+}2\arcsin\left(\frac{\cos(\alpha_{N-1})}{2}\right),\arcsin(s_n)\leq\alpha_1{-}2\arcsin\left(\frac{\cos(\alpha_1)}{2}\right).$$

The above condition holds for vectors close to the north or south pole of a spherical shell.

Remark 2: Analyzing the mapping, it becomes obvious that the image of each annulus under $h_i(A_i)$ is given by a thick spherical shell. We exploit this property to count the number of possible codewords with the help of the theta function (64) in Section IV-D1.

2) Lattice Quantizer—The lattice quantizer $Q_{\mathrm{NN}}(\cdot)$ returns the lattice point which has the shortest distance to the mapped point $h_i(\mathbf{S})$ and is therefore implemented as a nearest neighbor quantizer. This step can naively be seen as an integer least-squares (ILS) problems: Mathematically speaking we assume a point $h_i(\mathbf{s})\in\mathbb{R}^{n-1}$ and solve the optimization problem

$$\min_{\mathbf{b}\in\mathbb{Z}^{n-1}}\|h_i(\mathbf{s})-\mathbf{M}\mathbf{b}\|^2, \tag{57}$$

where the generator matrix of the lattice is $\mathbf{M}\in\mathbb{R}^{(n-1)\times(n-1)}$ Having gained insight into special geometric properties of $\mathbf{M}$, other approaches than ILS are usually prefered, as they yield solutions in a more efficient manner.

Remark 3: The process of the lattice quantization may return a point outside of the image of the original annulus. This is principally no problem for the algorithm since we keep the original annulus i still in mind. However, in case the returned point is in the outside of $\mathscr{B}^{n-1}$, i.e. if $\|Q_{\mathrm{NN}}(h_i(\mathbf{s}))\| > 1$, we scale the quantized vector back to the shell of the unit sphere. This step is necessary to ensure that the conditions under which the inverse mapping in Sec. IV-B3 is derived still hold.

3) Inverse Mapping—The inverse mapping $h_i^{-1}(\cdot)$ of a point in $\mathscr{B}^{n-1}$ back to the spherical shell $\mathscr{S}^n$ is performed as follows. Let $h_i(\mathbf{s})=\mathbf{y}$, then

$$\|\mathbf{y}\|{=}\cos(\alpha_i)-2\sin\left(\frac{|\arcsin(s_n)-\alpha_i|}{2}\right). \tag{58}$$

By reordering the previous equation we obtain

$$\arcsin(s_n)=\alpha_i \pm 2\arcsin\left(\frac{\cos\alpha_i - \|\mathbf{y}\|}{2}\right), \quad (59)$$

where the ±-operator corresponds to the northern and southern hemisphere, respectively.

We use (59) and $\|\mathbf{s}^{'}\|=\sqrt{1-s_n^2}$ in the original mapping function (55) to come up with

$$\mathbf{s}^{'}=\mathbf{y}\cdot\frac{\|\mathbf{s}^{'}\|}{\|\mathbf{y}\|} \quad (60)$$

$$=\mathbf{y}\cdot\left(\frac{\cos\left(\alpha_i \pm 2\arcsin\left(\frac{\cos\alpha_i-\|\mathbf{y}\|}{2}\right)\right)}{\|\mathbf{y}\|}\right) \quad (61)$$

and can finally state the inverse mapped point by adding the last coordinate s_n (59) to $\mathbf{s}'$ such that

$$\mathbf{s}=(\mathbf{s}^{'},s_n)^T. \quad (62)$$

The mapping, quantization and inverse mapping are illustrated graphically in Figure 7.

Remark 4: The derivation of the inverse mapping $h_i^{-1}(\cdot)$ is based on the mapping function $h_i(\cdot)$ and the assumption that there is no quantization. If we consider cases where $\|\mathbf{y}\| > h_i(P_i(\mathbf{s}))$ (due to the quantization process), we observe a negative argument in the arcsine function of equation (59). However, all formulas remain valid since the resulting point in equation (62) lies in the original (extended) annulus.

4) Encoding Process—Based on its quantized gain, annulus and index within its annulus, a vector is assigned an integer value which allows a unique encoding process. In view of this we need to know in advance how many possible codepoints exist in one unit shell and how many possible codepoints there are in each annulus. Having obtained the amount of codepoints within a unit shell and the gain partition of the original vector allows us to determine the number of codepoints in the unit shells with a lower index. This integer value then represents the quantized gain. Knowing how many codepoints there are in each annulus is necessary to determine all codepoints lying in the annuli with a lower index than the original annulus. Adding that part to the index of the codepoint within its own annulus represents the quantized shape. Then, adding both integer values of quantized gain and shape will define the message. From this message, one can easily learn both the annulus index and the codepoint number.

Remark 5: The counting of codepoints within a lattice is done by using the theta function which is described in detail in Sec. IV-D1.

C. Maximum Covering Angle

If the previously described scheme is employed for designing a spherical code, Theorem 6 will state that the angle between any vector on the shell of the unit sphere and its quantized version is always upper bounded by an angle θ such that the concept of wrapped spherical codes provides a constructive method for our proposed achievability scheme of Section II-B.

Theorem 6 (Upper Bound on the Covering Angle)—*The maximum covering angle* θ *between a vector* $\mathbf{s} \in A_i \subset \mathscr{S}^n$ *and its quantized version* $\hat{\mathbf{s}} \in \mathscr{S}^n$ *constructed by wrapped spherical codes as described before is bounded for any dimension by*

$$\theta \leq \begin{cases} \left(\frac{\pi}{2}\right) - 2\arcsin\left(\frac{\cos(\alpha_i)-r_\Lambda^{\mathrm{cov}}}{2}\right) & if\ Q_{NN}(h_i(\boldsymbol{s}))=\mathbf{0} \\ 2\cdot\arcsin\left(\frac{r_\Lambda^{\mathrm{cov}}}{2\cdot\|h_i(\hat{\mathbf{s}})\|}\right) + 2\arcsin\left(\frac{\|\hat{\mathbf{s}}-P_i(\hat{\mathbf{s}})\|+r_\Lambda^{\mathrm{cov}}}{2}\right) - 2\arcsin\left(\frac{\|\hat{\mathbf{s}}-P_i(\hat{\mathbf{s}})\|}{2}\right) & else. \end{cases} \tag{63}$$

Proof: See Appendix E.

We stress that the obtained upper bound is solely dependent on the vector $\hat{\mathbf{s}}$ and the index i of the annulus and r_Λ^{cov}. Those quantities are both available at the decoder, which then uses this knowledge in order to detect similarity, i.e. to know whether the query sequence **y** is in the expanded spherical cap.

Remark 6: As mentioned in remark 3, a vector that is quantized in the outside of $\mathscr{B}^{n-1}$ requires a projection to the unit sphere. However, the derived bound on the covering angle still holds, since the scaling of that vector does not change the angle between original and (scaled) quantized vector. Furthermore, if $\hat{\mathbf{s}}$ is outside $h_i(A_i)$, the bound holds due to the properties of the lattice (the maximum distance from the original vector to its quantized version is r_Λ^{cov}).

D. Numerical Performance Analysis

1) Introductory Remarks—In the following, we repeat the numerical simulations of Figure 4 and compare those for dimension $n = 25$ to an achievability scheme that is based on actually implementable spherical codes. For that purpose, some further explanations are necessary: The following performance analysis employs a Semi-Monte-Carlo simulation, where we first generate standard Gaussian vectors of dimension $n = 25$ and quantize the gain (Lloyd-Max) and shape (wrapped spherical code).

As before, we have to address the issue of an optimal rate allocation, as it is generally not known beforehand, how much of a given rate R should be allocated to the shape or gain quantization. Optimally, the following optimization problem has to be solved:

$$\min_{R_G, R_S} \Pr\{\text{maybe}\} \;\; \mathbf{s}.t.\; R{=}R_G{+}R_S.$$

Since this problem is computationally cumbersome, we stick to the same method as in Section II-C and search for the best R_G within a discrete set of reasonable values for a fixed R_S. At this point, we have not yet addressed the issue of how to obtain a spherical code for a desired rate R_S. As was described in the code construction, the performance of the wrapped spherical code depends on the geometrical properties of the underlying lattice, in particular its covering radius r_Λ^{cov}.

In view of this, we make use of another property of a lattice, namely its theta function, which counts the number of lattice points on a spherical shell with a given discrete radius. A common way to define this property is by expressing it as a complex polynomial in $q = e^{j\pi z}$. For our counting purposes, one particular form of the theta function is especially insightful: If N_m denotes the number of lattice points $\mathbf{x} \in \Lambda$ with $\|\mathbf{x}\|^2 = m$, then the theta function can be written as

$$\Theta_\Lambda(z) \triangleq \sum_{m=0}^{\infty} N_m q^m. \qquad (64)$$

Thus, the polynomial coefficients N_m of the theta function can be used for counting the number of lattice points lying on a spherical shell $\mathscr{S}^n_{\sqrt{m}}$. In order to relate this number of points to the actual number of codepoints of a wrapped spherical code, we have to take into account that the image of an annulus $A_i \subset \mathscr{S}^n$ under $h_i(\cdot)$ is a thick spherical shell $\mathscr{S}_{\tilde{r}_-, \tilde{r}_+}$ (14). If additionally the nearest neighbor quantization is considered, we see that all codewords for any random vector $\mathbf{S} \in A_i$ must also lie within a thick shell $\mathscr{S}_{r_-, r_+}$ where r_- and r_+ are defined as in algorithm 1. Evidently, due to the discrete nature of the coefficients of the theta function, also the rate of the shape quantizer is of discrete nature. By scaling the lattice accordingly (cf. Section I-C2), we can therefore indirectly adjust the rate as well.

Recalling that the points on the boundaries of an annulus A_i, $P_i(\mathbf{s})$, are mapped to $\mathbb{R}^{n-1}$ by simply deleting their last coordinate, we arrive at $\|P_i(\mathbf{s})'\| = \cos(\alpha_i)$ and can summarize the entire procedure in Algorithm 1.

Algorithm 1

Relating a given lattice scaling to the rate R_S of the shape quantizer.

```
**Data**: desired lattice scaling (here defined by $r^{cov}$
$M = 0$
**for** $i = 1$ *to* $N$ **do**
```

```
        $r_- \leftarrow \max(0, \cos(\alpha_i) - 2\sin(\pi/(2N)) - r_\Lambda^{\text{cov}})$
        $r_+ \leftarrow \min(1+r_\Lambda^{\text{cov}}, \cos(\alpha_i)+r_\Lambda^{\text{cov}})$
        $M = M + \text{CountCodepoints}(\Lambda, r_-, r_+)$
$R_S = \frac{1}{n}\log_2(M)$
```

Algorithm 1 refers to the function `CountCodepoints()` which basically implements the theta function of the lattice and returns $\sum_{m: r_- \leq \sqrt{m} \leq r_+} N_m$, i.e. the sum of all codepoints within the specified thick spherical shell.

2) Results of the Comparison—Figure 8 exhibits a comparison of the non-constructive result discussed earlier with the implementable scheme presented in this section. For this purpose, we have conducted a Monte-Carlo simulation that evaluates Pr {maybe} based on 1000 samples for the input vector.

The computation is facilitated by first conditioning on the value of **X**, i.e. Pr {maybe} = E [Pr {maybe|**X**}], where the expectation is calculated with Monte-Carlo. The inner expression is evaluated as follows: for each **x**, we quantize its shape using the wrapped spherical code and obtain a bound to the covering angle according to Theorem 6 and also keep the real angle for reference. With the limits on the gain of **x** (obtained using the gain quantization), we invoke Proposition 2 and calculate the conditional probability Pr {maybe|**X**}.

In addition to the new achievability and genie-aided achievability curves that are based on the upper bound of the covering angle given by Theorem 6, we include two further curves that use the actual angle $\angle(\mathbf{S}, \hat{\mathbf{S}})$ so that the quality of the upper bound can be assessed as well.

We observe that the practical implementation achieves a performance that – for a given Pr {maybe} – is only approx. 1 Bit worse than the non-constructive achievability results from Sec. II. Further, having perfect knowledge about the covering angle, the performance gap of the proposed scheme compared to the non-constructive achievability decreases steadily with higher rates and is smaller than 0.2 Bit for $R > 2$ Bit.

V. CONCLUSION

We studied the rate-reliability trade-off for the finite blocklength regime, when similarity queries with Gaussian input data and a quadratic similarity measure are performed. We provided expressions that allow for a numerically computable characterization of a lower and upper bound on the error probability and also compared it to formulas for the asymptotic case.

Throughout the paper, we emphasized the possibility of applying the derived framework to practical problems. As a matter of fact, we concentrated on implementable spherical codes, namely those based on wrapped spherical codes, and proved an upper bound on the covering angle so that this spherical code can indeed be employed for the shape quantization.

To sum up, our practical scheme can always be used if the following conditions are met: First, the *covering radius* r_Λ for the desired lattice in $\mathbb{R}^n$ - which upper bounds the distance between any arbitrary point in this dimension and a lattice point - has to be known. Second, an efficient *nearest neighbour search* must exist that guarantees the nearest neighbour in the lattice can be found. Since our bound on the covering angle is a function of the covering radius of the underlying lattice, the above mentioned two points are all that is needed to perform similarity queries with our scheme.

While the method described in the paper is implementable, as opposed to previous work, we do not yet have any optimality guarantees regarding the performance of the scheme (such as, for example, the optimality of shape-gain quantizers for source coding at high rates [16]). This is left for future work.

Another direction for future work is testing the usefulness of our schemes in real life scenarios. Applications where queries would be performed in a feature domain, where the Gaussian assumption is natural, seem particularly promising although the robustness property of the Gaussian distribution [1, Sec. III-D] suggests schemes that are designed assuming Gaussianity may be much more widely applicable.

APPENDIX A

Proof of Proposition 1

The cruical part for proving this proposition involves the probability density of the random variable $\mathbf{Y}| \, \|\mathbf{Y}\| = r_Y$ Using (34) and (19), one obtains

$$f_{\boldsymbol{Y}|\|\boldsymbol{Y}\|}(\boldsymbol{Y}|r_Y) = \frac{(2\pi)^{-\frac{n}{2}} \mathrm{e}^{-r_Y^2/2}}{\frac{2^{1-\frac{n}{2}} r^{n-1} \mathrm{e}^{-r_Y^2/2}}{\Gamma(\frac{n}{2})}} = \frac{1}{|\mathscr{S}^n_{r_Y}|}, \tag{65}$$

which shows that a Gaussian random variable $\mathbf{Y}|\|\mathbf{Y}\| = r_Y$ is uniformly distributed over sphere with radius r_Y in n dimensions.

Consequently, for calculating the probability $\Pr\{\mathbf{Y} \in \Gamma^D(\mathrm{CAP}_r(\hat{\mathbf{s}}, \theta))| \, \|\mathbf{Y}\| = r_Y\}$ it only matters what fraction of the sphere $\mathscr{S}^n_{r_Y}$ is included within in the D-expansion of the respective cap. Hence, we can make use of the Omega function described in (21) to calculate this particular fraction. As noted before, this function only depends on the angle of corresponding cap. After all, the entire problem boils down to the calculation of the angle that is passed to the Omega function. Given a certain relation of the given scheme parameters and r_Y, several different cases have to be distinguished for that and are summarized for reference in the following:

$$\Pr\{\boldsymbol{Y} \in \Gamma^D(\mathrm{CAP}_r(\hat{\mathbf{s}}, \theta)) | \|\boldsymbol{Y}\| = r_Y\} = \begin{cases} 0, & |r - r_Y| > \sqrt{nD} \\ 1, & r_Y \le r_{Y,\mathrm{deg}}(r, \theta) \\ \Omega(\theta + \theta'(r_Y, r)), & \text{otherwise.} \end{cases} \quad (66)$$

The first case (cf. situations (1a), (1b) in Figure 9) can easily be determined: If r_Y is too small or too large, **Y** can not lie inside the D-expansion of the thin cap.

Regarding the third case (cf. situation ③ in Figure 9), **Y** may lie inside the D-expansion and we account for the possible fraction of $\mathscr{S}^n_{r_Y}$ that is part of this set by introducing the expansion angle θ' (r_Y; r). Applying the law of cosines to the triangle (**0**, **x**, **y**), one obtains

$$\theta'(r_Y, r) \triangleq \arccos\left(\frac{r^2 + r_Y^2 - nD}{2 \cdot r \cdot r_Y}\right). \quad (67)$$

The second case (cf. situation ② in Figure 11) describes the degenerate situation, when the radius r_Y is so small such that the sphere $\mathscr{S}^n_{r_Y}$ is contained in the expanded cap. As Figure 10 reveals, this is the case for $r_Y \le r_{Y,\mathrm{deg}}(r, \theta)$ as given in (68).

$$r_{Y,\mathrm{deg}}(r, \theta) = \sqrt{(r\cos(\theta))^2 - r^2 + nD} - r\cos(\theta) \quad (68)$$

APPENDIX B

Proof of Proposition 2

We follow the same argumentation as in the previous proof. The calculation of the probability of the random variable $\mathbf{Y} | \, \|\mathbf{Y}\| = r_Y$ falling into a thick cap, i.e. $\Pr\left\{\boldsymbol{Y} \in \Gamma^D(\mathrm{CAP}_{r_1, r_2}(\hat{\mathbf{s}}, \theta)) | \|\boldsymbol{Y}\| = r_Y\right\}$, boils down to the computation of the covered fraction of the sphere $\mathscr{S}^n_{r_Y}$. However, the conditions for the different cases now have to be adapted:

$$\Pr\{\boldsymbol{Y} \in \Gamma^D(\mathrm{CAP}_{r_1, r_2}(\hat{\mathbf{s}}, \theta)) | \|\boldsymbol{Y}\| = r_Y\} = \begin{cases} 0, & r_Y < r_1 - \sqrt{nD} \\ & r_Y > r_2 + \sqrt{nD} \\ 1, & r_Y \le r'_{Y,\mathrm{deg}}(r_1, \theta) \\ \Omega(\theta + \theta''(r_1, r_2, r_Y)), & \text{otherwise.} \end{cases} \quad (69)$$

Cases (1a) $(r_Y < r_1 - \sqrt{nD})$ and (1b) $(r_Y < r_2 - \sqrt{nD})$ (cf. Figure 11) denote those situations when no part of the sphere $\mathscr{S}^n_{r_Y}$ is included within the Γ^D-expansion of the thick cap.

Case three (cf. ③ in Figure 11) turns out to be slightly more involved as in Proposition 1, as the thickness of cap has be taken into account. Having said this, one is able to make the distinction between three additional regions which are seperated by the boundaries $r_1^{'} = \sqrt{r_1^2 + nD}$ and $r_2^{'} = \sqrt{r_2^2 + nD}$. Those can be derived with the Pythagorean theorem for the respective triangles drawn in Figure 11. Applying the law of cosines eventually to one of the appropriate triangles $(\mathbf{0}, \mathbf{x}_1, \mathbf{y})$, $(\mathbf{0}, \mathbf{x}_2, \mathbf{y})$ and $(\mathbf{0}, \mathbf{x}_3, \mathbf{y})$ yields the following distinction for the expansion angle $\theta''(r_1, r_2, r_Y)$:

$$\theta''(r_1, r_2, r_Y) = \begin{cases} \arccos\left(\frac{r_1^2 + r_Y^2 - nD}{2 \cdot r_1 \cdot r_Y}\right), & r_Y \leq r_1^{'} \\ \arccos\left(\frac{r_2^2 + r_Y^2 - nD}{2 \cdot r_2 \cdot r_Y}\right), & r_Y \geq r_2^{'} \\ \arcsin\left(\frac{\sqrt{nD}}{r_Y}\right), & \text{otherwise.} \end{cases} \quad (70)$$

Concerning the second case in (69), the remarks of Proposition 1 apply analogously with the quantity $r_{Y,\deg}^{'}(r_1, \theta)$ being defined in the same way as in (68) and r replaced by r_1.

APPENDIX C

The Leech Lattice [11, Chapter 4.11] is a well-known lattice in $n = 24$ dimensions that is widely used in practice as it provides the densest lattice covering in this dimension.

The Leech Lattice was chosen a model lattice for this paper, as most of its properties are explored in detail and its geometry exhibits features that allow for convenient computations. It can be constructed in a large variety of ways, such as Golay codes and laminated lattices. We rely on its generation via its generator matrix $\mathbf{M} \in \mathbb{R}^{24 \times 24}$ that is given for reference in [11, Chapter 4, Figure 4.12]. Its determinant and hence the volume of the fundamental region evaluates as $|\det(\mathbf{M})| = 1$. Further, the minimal distance is $d_{\Lambda_{24}} = 2$ and relates directly to the covering radius as $r_{\Lambda_{24}}^{\text{cov}} = \frac{1}{2} d_{\Lambda_{24}} \sqrt{2} = \sqrt{2}$. Consequently, the packing and covering densities are given as

$$\vartheta^{\text{cov}}(\Lambda) \triangleq \frac{V\left(\mathscr{S}_{r_{\Lambda_{24}}^{\text{cov}}}\right)}{|\det(\mathbf{M})|} = \frac{\pi^{12}}{12!} \approx 0.001930, \quad (71)$$

$$\vartheta^{\text{pack}}(\Lambda) \triangleq \frac{V\left(\mathscr{S}_{r_{\Lambda_{24}}^{\text{pack}}}\right)}{|\det(\mathbf{M})|} = \frac{(2\pi)^{12}}{12!} \approx 7.9035. \quad (72)$$

Its theta function is given by

$$\Theta_{\Lambda_{24}}(z)=\sum_{n=0}^{\infty}N_m q^m=1+196560q^4+16773120q^6+398034000q^8+\ldots \quad (73)$$

APPENDIX D

For proving Theorem 5 we start with inequality (94) of [1] which reads as

$$\Pr\{\text{maybe}\} \geq \int_{\sqrt{n(1-\eta)}}^{\sqrt{n(1+\eta)}} f_{\|X\|}(r_X)\mathrm{d}r_X \cdot \int_{\sqrt{n(1-\eta)}}^{\sqrt{n(1+\eta)}} f_{\|Y\|}(r_Y)\mathrm{d}r_Y \cdot \sum_{i=1}^{2^{nR}} p_i \cdot \Omega\left(\theta_{D''}+\Omega^{-1}(p_i)\right), \quad (74)$$

where its full derivation can be traced back in the aforementioned paper.

The set of probabilities $\{p_i\}_{i=1}^{2^{nR}}$ expresses the probability of a Gaussian variable **X** being an element of the set of all points that have been mapped to one of the $i \in [1: 2^{nR}]$ possible signatures by a particular choice of the signature function $Q(\cdot)$ (cf. (73) in [1]). By construction, the elements p_i of the respective set sum up to one.

At that point Lemma 4 of [1] comes into play, which we state here for reference:

Lemma (Lemma 4, [1])

Let $0 < \Omega^* < 1$ *and* $0 < c < 1$ *be given constants. Define* p^* *to be the solution to* $\Omega(\theta_{D''}+\Omega^{-1}(p))=\Omega^*$. *Then if*

$$\sum_{i=1}^{2^{nR}} p_i \cdot \Omega\left(\theta_{D''}+\Omega^{-1}(p_i)\right) \leq c \cdot \Omega^*, \quad (75)$$

then

$$R \geq \frac{1}{n}\log\frac{1-c}{p^*}. \quad (76)$$

This Lemma now becomes vital as it allows for a reformulation of (74), when Lemma 4 is used the other way around:

$$\Pr\{\text{maybe}\} \geq \int_{\sqrt{n(1-\eta)}}^{\sqrt{n(1+\eta)}} f_{\|X\|}(r_X)\mathrm{d}r_X \cdot \int_{\sqrt{n(1-\eta)}}^{\sqrt{n(1+\eta)}} f_{\|Y\|}(r_Y)\mathrm{d}r_Y \cdot c \cdot \Omega^*, \quad (77)$$

if $R \leq \frac{1}{n}\log\frac{1-c}{p^*}$.

The best lower bound obviously can then be gained if we try to optimize the above expression with regard to c, Ω^* and η:

$$\Pr\{\text{maybe}\} \geq \max_{c,\Omega^*,\eta} c \cdot \Omega * \cdot \int_{\sqrt{n(1-\eta)}}^{\sqrt{n(1+\eta)}} f_{\|X\|}(r_X)\mathrm{d}r_X \cdot \int_{\sqrt{n(1-\eta)}}^{\sqrt{n(1+\eta)}} f_{\|Y\|}(r_Y)\mathrm{d}r_Y \quad (78)$$

$$\text{s.t. } 0<c<1 \quad (79)$$

$$0<\Omega^*<1 \quad (80)$$

$$0<\eta<1 \quad (81)$$

$$\Omega\left(\theta_{D''} + \Omega^{-1}(p^*)\right) = \Omega^* \quad (82)$$

$$R \leq \frac{1}{n}\log_2\left(\frac{1-c}{p^*}\right). \quad (83)$$

Appendix E

Proof of Theorem 6

First of all, we introduce an auxiliary variable $\hat{\hat{\mathbf{S}}}$ and bound the covering angle with the help of the triangle inequality, such that

$$\theta = \angle(\mathbf{X}, \hat{\mathbf{X}}) = \angle(\mathbf{S}, \hat{\mathbf{S}}) \leq \angle(\mathbf{S}, \hat{\hat{\mathbf{S}}}) + \angle(\hat{\hat{\mathbf{S}}}, \hat{\mathbf{S}}). \quad (84)$$

Without loss of generality we can assume a setting in $\mathbb{R}^{n-1}$ as shown in Figure 12, where $\hat{\hat{\mathbf{S}}}$ was chosen such that it is on the same line as $h_i(\mathbf{S})$ and $P_i'(\mathbf{S})$ and has distance $\|h_i(\hat{\mathbf{S}})\|$ from the origin. In general $\tilde{r}_\Lambda$ is bounded by $\tilde{r}_\Lambda = \|h_i(\mathbf{S}) - Q_{\text{NN}}(h_i(\mathbf{S}))\| \leq r_\Lambda$ because of the properties of the lattice Λ and the implementation of $Q_{\text{NN}}(\cdot)$ as a nearest neighbor quantizer.

We argue that $\angle(P_i(\mathbf{S}), P_i(\hat{\mathbf{S}}))=\angle(P_i(\hat{\hat{\mathbf{S}}}), P_i(\hat{\mathbf{S}}))=\angle(\hat{\hat{\mathbf{S}}}, \hat{\mathbf{S}})$ as follows. Since $h_i(\hat{\hat{\mathbf{S}}})$ is on the same line as $h_i(\mathbf{S})$ and $P_i^{'}(\mathbf{S})$ by construction, the definition of (56) implies $P_i^{'}(\mathbf{S})=P_i^{'}(\hat{\hat{\mathbf{S}}})$.

Due to the special property that the mapping for all points that lie exactly on the boundary between two annuli is just the deletion of the last coordinate, the distance between $P_i^{'}(\mathbf{S})=P_i^{'}(\hat{\hat{\mathbf{S}}})$ and $P_i^{'}(\hat{\mathbf{S}})$ does not change when applying the inverse mapping, so that $\|P_i^{'}(\mathbf{S})-P_i^{'}(\hat{\mathbf{S}})\|=\|P_i(\mathbf{S})-P_i(\hat{\mathbf{S}})\|=\|P_i(\hat{\hat{\mathbf{S}}})-P_i(\hat{\mathbf{S}})\|$.

We have then found a quantity to describe an upper bound of the first part of (84) because $\hat{\hat{\mathbf{S}}}$ and $\hat{\mathbf{S}}$ have the same latitude and $\angle(\hat{\hat{\mathbf{S}}}, \hat{\mathbf{S}})=\angle(P_i(\hat{\hat{\mathbf{S}}}), P_i(\hat{\mathbf{S}}))$.

More precise, the distance between $h_i(\hat{\hat{\mathbf{S}}})$ and $h_i(\mathbf{S})$ can be calculated as

$$\alpha=|\|h_i(\mathbf{S})\|-\|h_i(\hat{\hat{\mathbf{S}}}))\||\le|\tilde{r}_\Lambda+\|h_i(\hat{\mathbf{S}})\|-\|h_i(\hat{\hat{\mathbf{S}}})\||=\tilde{r}_\Lambda\le r_\Lambda, \quad (85)$$

where the inequality follows from $\|h_i(\mathbf{S})\|\le\tilde{r}_\Lambda+\|h_i(\hat{\mathbf{S}})\|$.

Furthermore the distance b between $h_i(\hat{\hat{\mathbf{S}}})$ and $h_i(\hat{\mathbf{S}})$ can be bounded as follows. We make use of the law of cosines and obtain

$$\tilde{r}_\Lambda^2=a^2+b^2-2ab\,;\cos\left(\frac{\pi}{2}\pm\frac{\delta}{2}\right) \quad (86)$$

$$=a^2+b^2\left(\frac{1\mp a}{\|h_i(\hat{\mathbf{S}})\|}\right). \quad (87)$$

The derivation is straightforward and by using $\|h_i(\hat{\mathbf{S}})\|\le 1$ and $a\le r_\Lambda<1$ we can state

$$b\le\tilde{r}_\Lambda. \quad (88)$$

Eventually, the angle δ is upper bounded by

$$\delta=2\cdot\arcsin\frac{b/2}{\|h_i(\hat{\mathbf{S}})\|}\le 2\cdot\arcsin\frac{\tilde{r}_\Lambda}{2\cdot\|h_i(\hat{\mathbf{S}})\|} \quad (89)$$

$$\leq 2 \cdot \arcsin \frac{r_\Lambda}{2 \cdot \|h_i(\hat{\mathbf{S}})\|}. \qquad (90)$$

For the second part in (84) we consider the setting shown in Figure 13 and state that the angle β can be calculated as

$$\beta = 2\arcsin\left(\frac{\|\hat{\hat{\mathbf{S}}} - P_i(\hat{\hat{\mathbf{S}}})\|}{2}\right), \qquad (91)$$

and γ_1 is given by

$$\begin{aligned}\gamma_1 = {} & 2\arcsin\left(\frac{\|\hat{\hat{\mathbf{S}}} - P_i(\hat{\hat{\mathbf{S}}})\| + a}{2}\right) \\ & - \beta \leq 2\arcsin\left(\frac{\|\hat{\hat{\mathbf{S}}} - P_i(\hat{\hat{\mathbf{S}}})\| + r_\Lambda}{2}\right) \\ & - \beta \leq 2\arcsin\left(\frac{\|\hat{\hat{\mathbf{S}}} - P_i(\hat{\hat{\mathbf{S}}})\| + r_\Lambda}{2}\right) \\ & - 2\arcsin\left(\frac{\|\hat{\hat{\mathbf{S}}} - P_i(\hat{\hat{\mathbf{S}}})\|}{2}\right), \end{aligned} \qquad (92)$$

where $\|\hat{\hat{\mathbf{S}}} - P_i(\hat{\hat{\mathbf{S}}})\| = 2\sin\frac{\arcsin \hat{x}_n - \alpha_i}{2}$ and therefore the expression is only a function of $\hat{\mathbf{S}}$ and index *i*.

Depending on the relation of the latitudes of **S** and $\hat{\hat{\mathbf{S}}}$ (i.e. $\mathbf{S}_n \geq \hat{\hat{\mathbf{S}}}_n$ or $\mathbf{S}_n < \hat{\hat{\mathbf{S}}}_n$) a second case has to be taken into account: When **S** is closer to the corresponding annulus point P_i(**S**) than $\hat{\hat{\mathbf{S}}}$ this leads to

$$\begin{aligned}\gamma_2 = \beta - {} & 2\arcsin\left(\frac{\|\hat{\hat{\mathbf{S}}} - P_i(\hat{\hat{\mathbf{S}}})\| - a}{2}\right) \leq \beta \\ & - 2\arcsin\left(\frac{\|\hat{\hat{\mathbf{S}}} - P_i(\hat{\hat{\mathbf{S}}})\| - r_\Lambda}{2}\right) \leq 2\arcsin\left(\frac{\|\hat{\hat{\mathbf{S}}} - P_i(\hat{\hat{\mathbf{S}}})\|}{2}\right) \\ & - 2\arcsin\left(\frac{\|\hat{\hat{\mathbf{S}}} - P_i(\hat{\hat{\mathbf{S}}})\| - r_\Lambda}{2}\right). \end{aligned} \qquad (93)$$

We observe that (93) is always smaller than (92) because of the special structure of the expressions and the fact that $\frac{\mathrm{d}}{\mathrm{d}x}\arcsin(x)=\frac{1}{\sqrt{1-x^2}}$ is increasing with larger values of x. We therefore set $\gamma = \max(\gamma_1, \gamma_2) = \gamma_1$ and have derived our desired result.

Remark 7

The bound on the covering angle holds also for the case that a point is quantized in the outside of $\mathscr{B}^{n-1}$ and has to be projected back to the sphere (see remark 3). The reason behind that is that projecting a point in the outside of $\mathscr{B}^{n-1}$ back to the sphere makes it closer to the original point. This can easily be shown using the law of cosine.

References

1. Ingber, A., Courtade, TA., Weissman, T. Compression for quadratic similarity queries. Submitted to IEEE Trans. on Information Theory. 2013. [Online]. Available: http://arxiv.org/abs/1307.6609
2. Ingber, A., Weissman, T. The minimal compression rate for similarity identification. Submitted to IEEE Trans. on Information Theory. 2013. [Online]. Available: http://arxiv.org/abs/1312.2063
3. Gersho, A., Gray, R. Vector Quantization and Signal Compression. Kluwer Academic Publishers; 1992. ser. Kluwer international series in engineering and computer science
4. Dumer I. Covering spheres with spheres. Discrete & Computational Geometry. 2007; 38(4):665–679.
5. Hamkins J, Zeger K. Asymptotically dense spherical codes. I. Wrapped spherical codes. IEEE Transactions on Information Theory. 1997; 43(6):1774–1785.
6. Gallager, RG. Information Theory and Reliable Communication. New York, NY, USA: John Wiley & Sons, Inc; 1968.
7. Cover, TM., Thomas, JA. Elements of Information Theory. John Wiley & Sons: 1991.
8. Li S. Concise formulas for the area and volume of a hyperspherical cap. Asian Journal of Mathematics & Statistics. 2011; (4):66–70.
9. Sloane N. Tables of sphere packings and spherical codes. Information Theory, IEEE Transactions on. 1981; 27(3):327–338.
10. Schurmann A, Vallentin F. Computational approaches to lattice packing and covering problems. Discrete & Computational Geometry. 2006; 35(1):73–116. [Online]. Available: http://dx.doi.org/10.1007/s00454-005-1202-2.
11. Conway, JH., Sloane, NJA. Sphere packings, lattices and groups. Vol. 290. Springer; 1993. ser. Grundlehren der math. Wissenschaften
12. Miller, K. Multidimensional Gaussian Distributions. John Wiley and Sons; 1964. ser. The SIAM series in applied mathematics
13. Lloyd S. Least squares quantization in PCM. IEEE Transactions on Information Theory. 1982; 28(2):129–137.
14. Max J. Quantizing for minimum distortion. Information Theory, IRE Transactions on. 1960; 6(1):7–12.
15. Hamkins, J. Design and Analysis of Spherical Codes. Sep. 1996 dissertation, Univ. of Illinois at Urbana-Champaign
16. Hamkins J, Zeger K. Gaussian source coding with spherical codes. IEEE Transactions on Information Theory. 2002; 48(11):2980–2989.

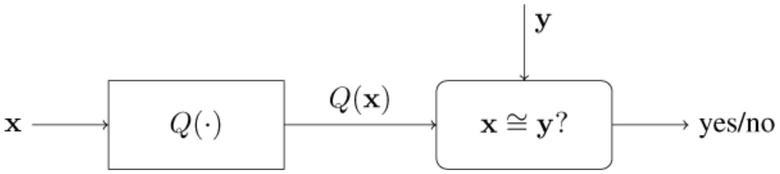

Fig. 1.

Answering a query from compressed data.

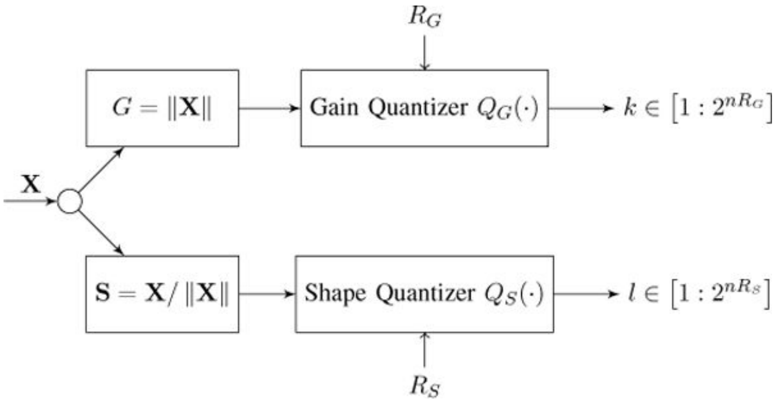
R_G
$G = \|\mathbf{X}\|$
Gain Quantizer $Q_G(\cdot)$
$k \in [1 : 2^{nR_G}]$
$\mathbf{X}$
$\mathbf{S} = \mathbf{X}/\|\mathbf{X}\|$
Shape Quantizer $Q_S(\cdot)$
$l \in [1 : 2^{nR_S}]$
R_S

Fig. 2.
Illustration of a shape-gain quantization process for the proposed achievability part.

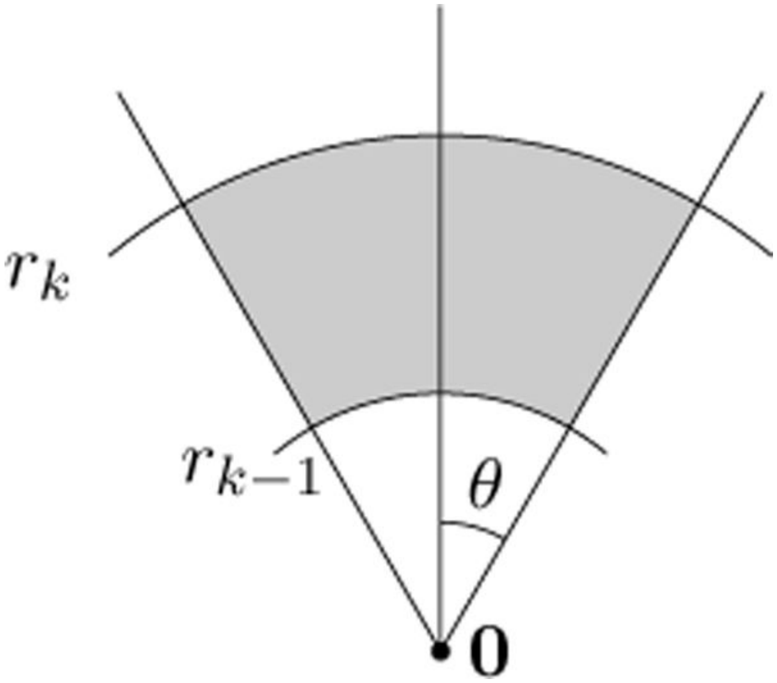

Fig. 3.

Illustration of the spherical cap that contains the set $Q^{-1}(i)$.

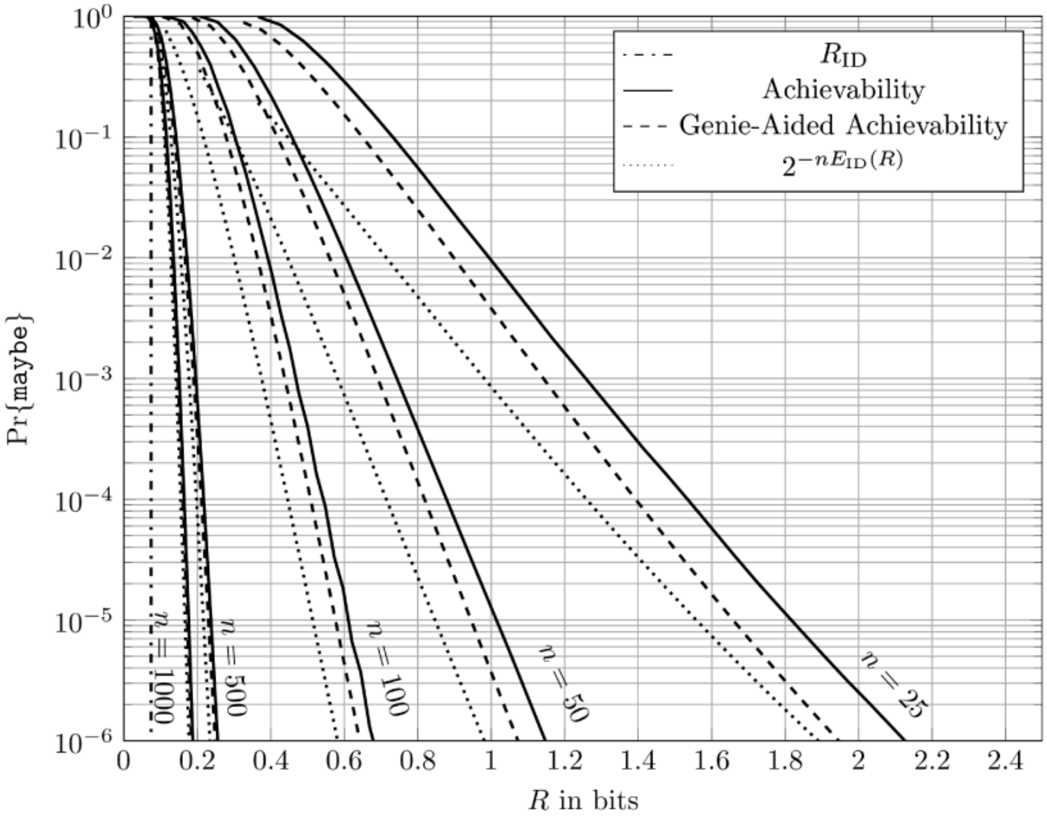

Fig. 4.

Numerical evaluation of the achievability based on Theorem 3 and the guaranteed covering angle of the non-constructive spherical code of [4] with $D = 0.1$.

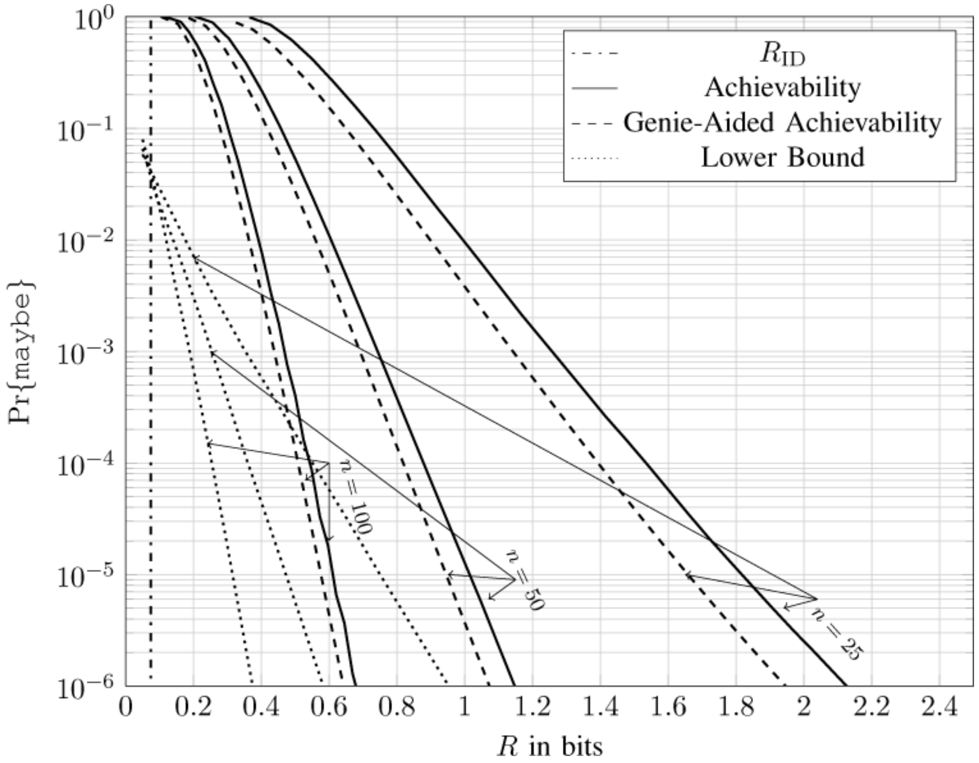

Fig. 5.

Comparison of the achievability and the converse for a desired similarity level of $D = 0.1$.

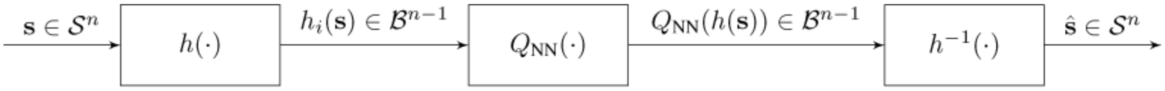
$\mathbf{s} \in \mathcal{S}^n$
$h(\cdot)$
$h_i(\mathbf{s}) \in \mathcal{B}^{n-1}$
$Q_{\mathrm{NN}}(\cdot)$
$Q_{\mathrm{NN}}(h(\mathbf{s})) \in \mathcal{B}^{n-1}$
$h^{-1}(\cdot)$
$\hat{\mathbf{s}} \in \mathcal{S}^n$

Fig. 6.

Block diagram for the shape quantization process of a random vector s using a wrapped spherical code. $h(\cdot)$ denotes the mapping function, $Q_{\mathrm{NN}}(\cdot)$ is the lattice quantizer and $h^{-1}(\cdot)$ is the inverse mapping.

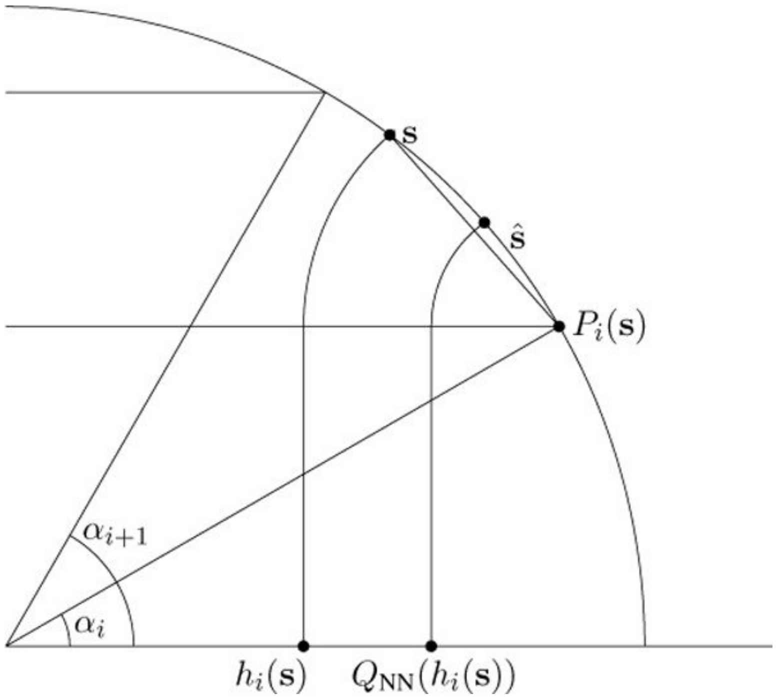

Fig. 7.

Mapping, quantization and inverse mapping of a random vector **S**.

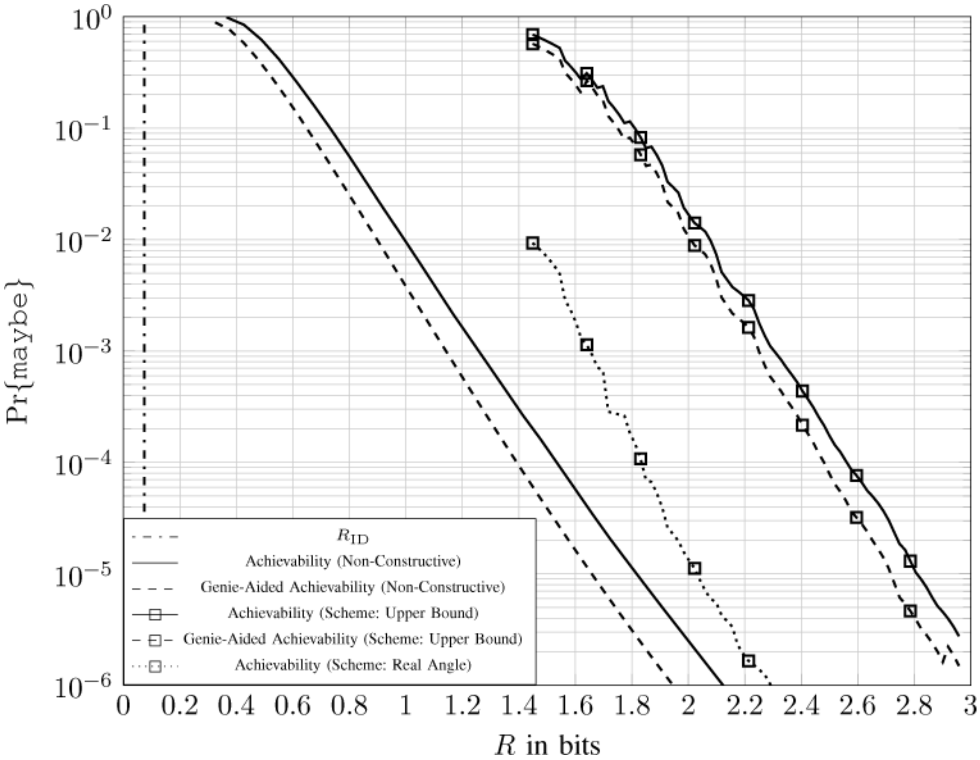

Fig. 8.

Numerical evaluation of the performance of the scheme using wrapped spherical codes (based on the Leech lattice) in $n = 25$ compared to the non-constructive achievability result from Section II

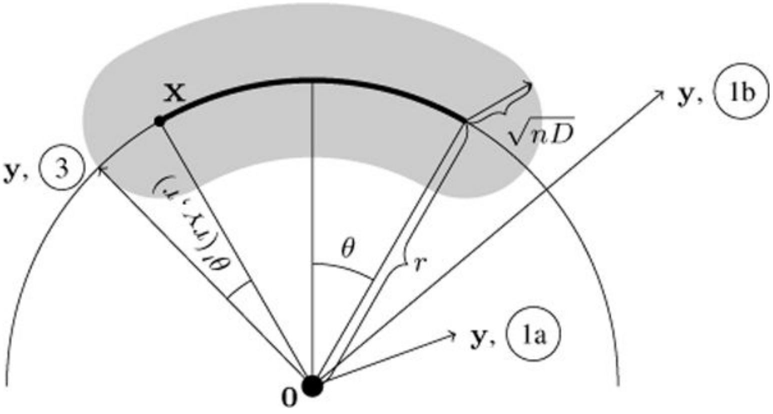

Fig. 9.

Probability of **Y** falling into the Γ^D-expansion of a thin cap: Cases 1a, 1b and 3

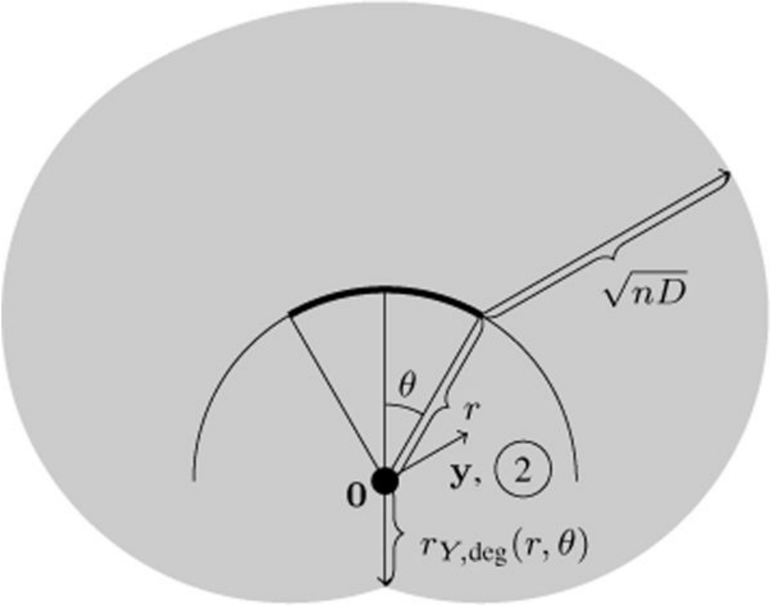

Fig. 10.

Probability of **Y** falling into the Γ^D-expansion of a thin cap: Degenerated case 2

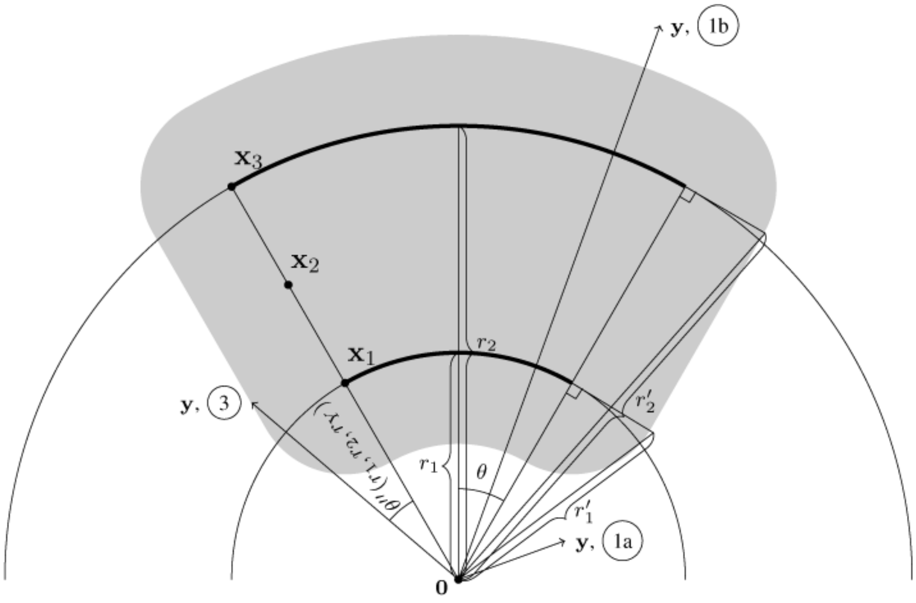

Fig. 11.

Probability of **Y** falling into an expanded thick cap: Cases 1a, 1b and 3

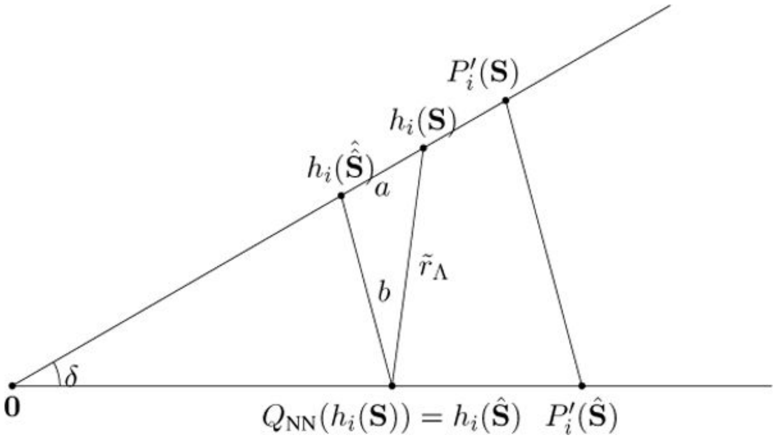

Fig. 12.

Setting in $\mathbb{R}^{n-1}$ to derive a bound on $\angle(\hat{\mathbf{S}}, \hat{\hat{\mathbf{S}}})$ in (84) for the maximum covering angle θ.

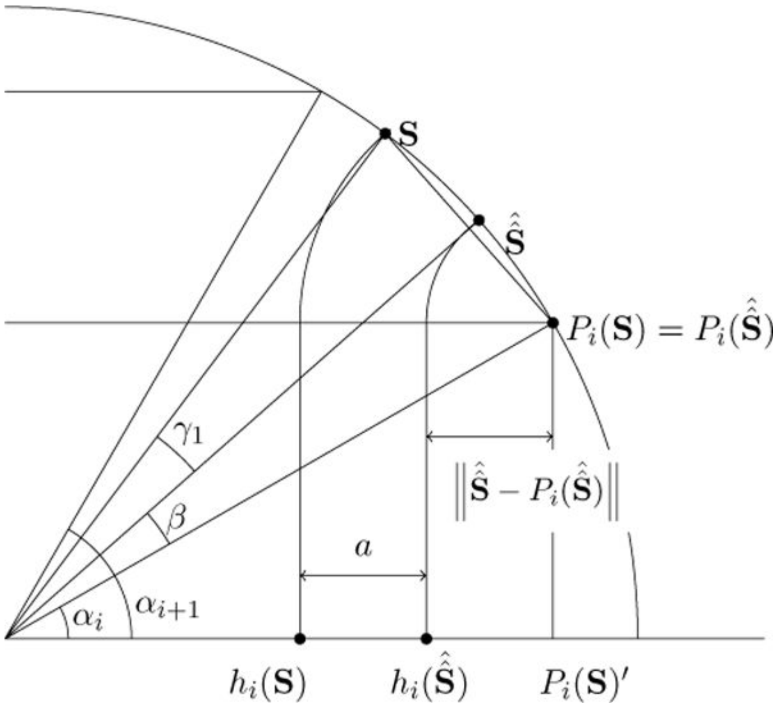
$\mathbf{S}$
$\hat{\hat{\mathbf{S}}}$
$P_i(\mathbf{S}) = P_i(\hat{\hat{\mathbf{S}}})$
γ_1
$\left\| \hat{\hat{\mathbf{S}}} - P_i(\hat{\hat{\mathbf{S}}}) \right\|$
β
a
α_i
α_{i+1}
$h_i(\mathbf{S})$
$h_i(\hat{\hat{\mathbf{S}}})$
$P_i(\mathbf{S})'$

Fig. 13.

Deriving a bound on $\angle(\hat{\hat{\mathbf{S}}}, \mathbf{S})$ in (84) for the maximum covering angle θ.