Tools for Analysis of Shannon-Kotel’nikov Mappings

Pål Anders Floor, Tor A. Ramstad,

P A. Floor is at NTNU Gjovik, Gjovik, Norway
T A. Ramstad is with the Department of Electronics and Telecommunications, Norwegian University of Science and Technology (NTNU), Trondheim, Norway (e-mail: ramstad@iet.ntnu.no).

Abstract—This document introduces tools from differential geometry needed for the analysis of a subset of Joint Source-Channel Codes (JSCC) named Shannon-Kotel’nikov (S-K) mappings. New results based on these concepts are further provided.

I. INTRODUCTION

This document extends the theory of S-K mappings [1], [2], [3] by introducing useful tools from differential geometry for analysis of (hyper) surfaces like curvature, geodesics, first fundamental form, second fundamental form, coordinate mappings, ruled surfaces and developable surfaces. These results will complement the ones provided in [1].

II. USEFUL THEOREMS AND TOOLS FROM DIFFERENTIAL GEOMETRY

This section explains some of the concepts and the terminology used in the extended analysis of S-K mappings. Most results stated in this section are taken from E. Kreyszig’s book Differential Geometry [4].

A. Fundamental concepts of Parametric Curves in \( \mathbb{R}^3 \)

Any continuous (class \( C^1 \)) curve \( C: S \in \mathbb{R}^3 \) can be represented perimetrically as the vector valued function

\[
S(x) = [S_1(x), S_2(x), S_3(x)], \quad x \in \mathbb{R}. \tag{1}
\]

In the following we denote the derivative w.r.t to a general parameter as \( S', S'' \) and \( S''' \) and with respect to (w.r.t.) the special case of arch length as \( S, \dot{S} \) and \( \ddot{S} \). That is, when we parameterize the curve using

\[
\ell(x) = \int_{x_0}^{x} \sqrt{S' \cdot S'} \, dx = \int_{x_0}^{x} ||S'|| \, dx \tag{2}
\]

With this parametrization the norm of the derivative of the curve at any point, corresponding to the length of the tangent vector at that point, is unity. I.e., \( ||t|| = ||S|| = 1, \forall x \) (See Appendix A).

There are three unit vectors connected to any curve \( C: S \in \mathbb{R}^3 \): The unit tangent vector \( t = \dot{S} = S'/||S'|| \), the unit principal normal vector \( p = \ddot{t}/||\ddot{t}|| = \ddot{S}/||\ddot{S}(x_0)|| \) and the unit binormal vector \( b = t \times p \). The vectors \( t, p \) and \( b \) make out a vector space of mutually orthogonal vectors named moving trihedron which is so defined at each point along the curve \( C \). This is nicely illustrated in Kreyszig’s book [4, pp. 36-37]. These three vectors further define three mutually orthogonal planes: i) Osculating Plane spanned by \( t \) and \( p \), ii) normal plane spanned by \( p \) and \( b \), and iii) rectifying plane spanned by \( t \) and \( b \).

For a parametric curve \( S(x) \) one can define the curvature [4, p. 34] w.r.t. arch length as \( \kappa_0 = ||\ddot{S}(x_0)|| \). We then also have \( p = (1/\kappa)\ddot{S} \).

The torsion [4, p. 37-40] is defined as

| \[
\tau(x) = -p \cdot b = \frac{\dddot{S} \cdot S \times S}{||S'||^2}. \tag{3}
|}

When \( \tau = 0, \forall x \), we have a plane curve. Whenever \( \tau \neq 0 \), the curve is not a plane curve, but will “twist” up into space \( \mathbb{R}^3 \). For a general parametrization the curvature and torsion are given by [4, pp.35,39]

| \[
\kappa = \frac{\sqrt{||S'||^2||S''|^2 - (S' \cdot S'')^2}}{||S'||^2}, \tag{4}
|}

| \[
\tau = \frac{|S' \cdot S'' \cdot S'''|}{||S'||^2||S''|^2 - (S' \cdot S'')^2}. \tag{5}
|}

The curvature can locally be interpreted as a circle of radius \( \rho = 1/\kappa \), named radius of curvature, lying in the osculating plane of \( S \). The corresponding circle is named Osculating Circle and its center named centre of curvature. I.e., the curvature in a \( \epsilon \)-neighborhood of a given point is equivalent to that of a circle with radius \( \rho \). This concept of curvature is also valid for curves in \( \mathbb{R}^M, M \geq 3 \).

The Formula of Frenet [4, p. 41] relates the derivatives \( \dot{t}, \dot{p} \) and \( \dot{b} \) to linear combinations of \( t, p, \) and \( b \) as follows:

| \[
\dot{t} = \kappa p, \tag{6}
|}

| \[
\dot{p} = -\kappa t + \tau b, \tag{6}
|}

| \[
\dot{b} = -\tau p. \tag{6}
|}

One can now obtain a canonical representation of a curve. That is, its shape in the neighborhood of any of its points is related to the first three derivatives of the curve: With a curve of class \( S \in C^3 \), and \( \ell \) the curve length, one has the Taylor expansion:

| \[
S = S(0) + \sum_{v=1}^{3} \frac{\ell^v}{v!} \frac{d^v S(0)}{d\ell^v} + O(\ell^3). \tag{7}
|}

From the formula of Frenet one can show that

| \[
\dot{S} = \kappa p, \tag{8}
|}

| \[
\ddot{S} = \dot{t} + \dot{p} = \dot{\kappa} p + \kappa \dot{p} = \kappa p - \kappa^2 t + \kappa \tau b. \tag{8}
|}
By aligning \( t(0), p(0) \) and \( b(0) \) with the positive rays of the \( x_1, x_2 \) and \( x_3 \) axes of \( \mathbb{R}^3 \) respectively, one obtains (\( S_i \) is the \( i \)th component of \( S \)) [4, p.48]

\[
\begin{align*}
S_1(\ell) &= \ell - \frac{\kappa_2}{3} \ell^3 + O(\ell^3) \\
S_2(\ell) &= \frac{\kappa_2}{2} \ell^2 + \frac{\kappa_0}{3!} \ell^3 + O(\ell^3) \\
S_3(\ell) &= \frac{\kappa_0}{3!} \ell^3 + O(\ell^3).
\end{align*}
\]

By discarding all terms in each component except the dominating one we get the canonical representation

\[
S(\ell) \approx \left[ \ell, \frac{\kappa_0}{2} \ell^2, \frac{\kappa_0 \tau_0}{6} \ell^3 \right]
\]

Another useful concept is that of the Osculating Sphere [4, pp. 54-55]. A curve \( S \in \mathbb{R}^3 \) of class \( C^3 \) (functions of continuous 3rd order derivatives) has 3rd order contact (i.e. touches the curve up to 3rd order derivative) in the point \( P \) with the osculating sphere, which is a sphere with center \( a = s + p\dot{p} + r\tau b \) and radius \( R_s = \sqrt{\rho^2 + (\dot{\rho}\tau)^2} \). For the special case where \( \dot{\kappa} = 0 \) in \( P \), the center of the osculating sphere will lie in the osculating plane of \( S \). That is, with constant curvature \( \dot{\rho} = 0 \) and \( R_s = \rho \) (like circles).

The above results imply that any curve of class \( C^3 \) (or above), \( s(x) \in \mathbb{R}^3 \) has a spherical geometry locally, where the sphere is uniquely determined by \( \rho = 1/\kappa, \dot{\rho} = 1/\kappa \) and \( \tau \). Locally, we then have the approximation in (10).

B. Fundamental concepts of Parametric Surfaces in \( \mathbb{R}^3 \)

Many concepts from curves can be extended to surfaces, as any surface can be mathematically described as a set of (interrelated) coordinate curves. However, many new concepts are necessary. We repeat key results from Chapter III in [4, pp.72-117] here.

1) Fundamentals: A surface (or manifold) \( M \in \mathbb{R}^3 \) can be described parametrically as

\[
S(u^1, u^2) = [s_1(u^1, u^2), s_2(u^1, u^2), s_3(u^1, u^2)],
\]

where \( u_1 \) and \( u_2 \) are variables defined on a simply-connected bounded domain \( B \subseteq \mathbb{R}^2 \) (the \( u^1 u^2 \)-plane). For (11) to represent a surface then: 1) \( S(u^1, u^2) \in C^r, \ r \geq 1 \) on \( B \) and each point in the set \( M \) represented by \( S(u^1, u^2) \) corresponds to just one ordered pair \( (u^1, u^2) \in B \). 2) The Jacobian Matrix (see (2.14) in Appendix B), \( J \), is of rank 2 on \( B \).

In the following, partial derivatives will be denoted

\[
S_\alpha = \frac{\partial S}{\partial u^\alpha}, \quad S_{\alpha\beta} = \frac{\partial^2 S}{\partial u^\alpha \partial u^\beta}.
\]

By imposing a coordinate transformation

\[
u^\alpha = u^\alpha(\bar{u}^1, \bar{u}^2),
\]

a new parametric representation \( S(\bar{u}^1, \bar{u}^2) \) of \( M \) is obtained.

**Definition 1:** An Allowable Coordinate Transformation (ACT) is one where the functions in (13):

i) are defined in a domain \( \bar{B} \) such that (s.t.) the corresponding range of values includes the domain \( B \).

ii) are one-to-one of class \( r \geq 1 \) (\( \in C^r \)) everywhere in \( \bar{B} \).

iii) have Jacobian

\[
D = \frac{\partial(u^1, u^2)}{\partial(\bar{u}^1, \bar{u}^2)} = \begin{vmatrix}
\frac{\partial u^1}{\partial \bar{u}^1} & \frac{\partial u^1}{\partial \bar{u}^2} \\
\frac{\partial u^2}{\partial \bar{u}^1} & \frac{\partial u^2}{\partial \bar{u}^2}
\end{vmatrix} \neq 0,
\]

everywhere in \( \bar{B} \).

\[
A \text{ curve on a surface } S : S(u^1, u^2) \text{ can be determined by the parametric representation}
\]

\[
u^1 = u^1(t), \quad u^2 = u^2(t),
\]

doing class \( r \geq 1 \), where the parameter \( t \in \mathbb{R} \). Of special importance are the coordinate curves on \( S \), \( u^1 = \text{constant} \) and \( u^2 = \text{constant} \), corresponding to parallels to the coordinate axes in \( u^1 u^2 \)-plane. A set of curves on \( S \) that depend continuously on a parameter is said to be a one-parameter family of curves. Two one-parameter family of curves are called a net of curves on \( S \) if through every point \( P \) of \( S \) there passes one and only one curve of each of these families, and if the two curves have a distinct direction at \( P \).

The direction of the tangent to a curve \( C : u^1(t), u^2(t) \) on \( S : S(u^1, u^2) \) is determined by:

\[
S' = \frac{dS}{dt} = \frac{\partial S}{\partial u^1} \frac{d\nu^1}{dt} + \frac{\partial S}{\partial u^2} \frac{d\nu^2}{dt} = S_1(\nu^1)' + S_2(\nu^2)',
\]

which depends on \( t \). The tangent vector, \( S' \), is a linear combination of the coordinate curves’ tangent vectors \( S_\alpha \) at \( P \). Whenever \( S_1 \) and \( S_2 \) are linearly independent, they span the tangent plane \( E(P) \) of \( S \) at \( P \). \( E(P) \) therefore contains the tangent of any curve on \( S \) that passes through \( P \). With \( q^1, q^2 \) coordinates of the points of \( E(P) \), the tangent plane can be expressed as

\[
y(q^1, q^2) = S + q^1 S_1 + q^2 S_2, \quad S_1 \times S_2 \neq 0.
\]

Since \( y - S \in E(P), y - S, S_1 \) and \( S_2 \) are linearly dependent, giving the alternative expression for the tangent plane

\[
|y - S| S_1 S_2| = 0.
\]

To cope with all the sum-operations that result when analyzing surfaces and coordinate transformations in general, it is convenient to express all sums through Einstein summation convention to significantly shorten expressions [4, p.84]:

**Definition 2:** Einstein summation convention: If in a product a letter figures twice, once a superscript and once a subscript, summation should be carried out from 1 to \( N \) w.r.t. this letter.

For example, for simple sums

\[
\sum_{\alpha=1}^{N} a^\alpha b_\alpha = a^\alpha b_\alpha,
\]

and for double sums

\[
\sum_{\alpha=1}^{N} \sum_{\beta=1}^{N} a_{\alpha\beta} u^\alpha u^\beta = a_{\alpha\beta} u^\alpha u^\beta,
\]

and so on.

**Contravariant and Covariant vectors and tensors:**

**Definition 3:** Contravariant and Covariant vectors:
A contravariant vector (or 1st order contravariant tensor) at $P$ of $S$ is a vector that under the coordinate transformation in (13), and its inverse, transform as

$$
\bar{a}^\beta = a^\alpha \frac{\partial u^\beta}{\partial u^\alpha} \text{ and } \bar{a}^\gamma = a^\alpha \frac{\partial u^\gamma}{\partial u^\alpha}, \quad \beta, \gamma = 1, \cdots, N, \quad (21)
$$

where $\bar{a}^\beta$ and $a^\alpha$ are vector components in the two coordinate systems. The components of a vector in one system is related to the components in the other system by parallel projection onto the coordinate axis in the original system. See Fig. 1 to the left.

A covariant vector (or 1st order covariant tensor) at $P$ of $S$ is a vector that under the coordinate transformation in (13), and its inverse, transform as

$$
\bar{b}^\beta = b^\alpha \frac{\partial u^\alpha}{\partial u^\beta}, \text{ and } \bar{b}^\gamma = b^\alpha \frac{\partial u^\gamma}{\partial u^\alpha}, \quad \beta, \gamma = 1, \cdots, N, \quad (22)
$$

where $a^\beta$ and $b^\alpha$ are vector components in the two coordinate systems. The components of a vector in one system is related to the components in the other system by orthogonal projection onto the coordinate axis in the original system. See Fig. 1 to the left.

One example of a covariant vector field is the Gradient, $\partial \phi / \partial u^\alpha$, of a scalar function, since

$$
\frac{\partial \phi}{\partial u^\beta} = \frac{\partial \phi}{\partial u^\alpha} \frac{\partial u^\alpha}{\partial u^\beta}. \quad (23)
$$

Note that in orthogonal coordinate systems, contravariant and covariant transformations are the same.

**Definition 4: 2nd order Tensors:**

A 2nd order contravariant tensor at $P$ of $S$ is an entity $a^{\alpha\beta}$ that under the coordinate transformation in (13), and its inverse, transform as

$$
\bar{a}^{\gamma\kappa} = a^{\alpha\beta} \frac{\partial u^\kappa}{\partial u^\gamma} \frac{\partial u^\beta}{\partial u^\alpha}, \quad \gamma, \kappa = 1, \cdots, N, \quad (24)
$$

and

$$
a^{\sigma\tau} = \bar{a}^{\gamma\kappa} \frac{\partial u^\gamma}{\partial u^\sigma} \frac{\partial u^\kappa}{\partial u^\tau}, \quad \sigma, \tau = 1, \cdots, N, \quad (25)
$$

where $\bar{a}^{\gamma\kappa}$ and $a^{\alpha\beta}$ are tensor components in the two coordinate systems.

A 2nd order covariant tensor at $P$ of $S$ is an entity $a_{\alpha\beta}$ that under the coordinate transformation in (13), and its inverse, transform as

$$
a_{\gamma\kappa} = a_{\alpha\beta} \frac{\partial u^\alpha}{\partial u^\gamma} \frac{\partial u^\beta}{\partial u^\kappa}, \quad \gamma, \kappa = 1, \cdots, N, \quad (26)
$$

and

$$
a_{\sigma\tau} = \bar{a}_{\gamma\kappa} \frac{\partial u^\gamma}{\partial u^\sigma} \frac{\partial u^\kappa}{\partial u^\tau}, \quad \sigma, \tau = 1, \cdots, N, \quad (27)
$$

where $\bar{a}_{\gamma\kappa}$ and $a_{\alpha\beta}$ are tensor components in the two coordinate systems.

A 2nd order mixed tensor at $P$ of $S$ is an entity $a^\alpha_\beta$ that under the coordinate transformation in (13), and its inverse, transform as

$$
a^\alpha_{\gamma\kappa} = a^\alpha_\beta \frac{\partial u^\gamma}{\partial u^\alpha} \frac{\partial u^\kappa}{\partial u^\beta}, \quad \gamma, \kappa = 1, \cdots, N, \quad (28)
$$

and

$$
a^\alpha_{\sigma\tau} = \bar{a}^\alpha_{\gamma\kappa} \frac{\partial u^\gamma}{\partial u^\sigma} \frac{\partial u^\kappa}{\partial u^\tau}, \quad \sigma, \tau = 1, \cdots, N, \quad (29)
$$

where $\bar{a}^\alpha_{\gamma\kappa}$ and $a^\alpha_\beta$ are tensor components in the two coordinate systems.\]

Note that the $N^2$ components of any 2nd order tensor can be arranged in a matrix.

The contravariant metric tensor (which will be introduced in Section II-B2) is an example of a contravariant tensor and the metric tensor (or metric of the surface) as described in Appendix B is an example of a covariant tensor.

**2) Fundamental Forms and Curvature:** The fundamental forms are crucial in order to analyze surfaces based on coordinate representations.

**First Fundamental Form (FFF):** In order to measure lengths and areas on a surface, introducing a metric on the surface is necessary. To determine an element of arc (length differential) of a curve $C$ in (15) on $S$ one can use (16):

$$
d^2 = (S_1 du^1 + S_2 du^2) \cdot (S_1 du^1 + S_2 du^2) = S_1 \cdot S_1 (du^1)^2 + 2S_1 \cdot S_2 du^1 du^2 + S_2 \cdot S_2 (du^2)^2. \quad (30)
$$

We recognize the quantities $g_{\alpha\beta} = S_\alpha \cdot S_\beta$ as the components of the metric tensor in Appendix B. By the summation convention we have

$$
d^2 = g_{\alpha\beta} du^\alpha du^\beta. \quad (31)
$$

This quadratic form is named First Fundamental Form (FFF), and defines a metric on $S$ which allows us to measure arc lengths, angles and areas on $S$. This is why the (covariant) tensor $g_{\alpha\beta}$ is named metric tensor. The metric tensor transforms covariantly under coordinate transformations (see Definition 3).

Any vector $v \in E(P)$ of $S$ has a contravariant and covariant representation: A vector $v$ on $S$ at $P$ in $E(P)$ can therefore be decomposed as $v = a^\alpha S_\alpha$. The components $a^\alpha$ are the lengths of the respective parallel projections of $v$ on the $S_\alpha$ axis of the coordinate system in $E(P)$ measured in units of $\sqrt{g_{\alpha\alpha}}$ as shown to the left in Fig. 1 (see section II-B1). One may represent $v$ by its orthogonal projection onto $S_\alpha$.

![Fig. 1. Left: Decomposition into contravariant components $a^\alpha$. Right: Decomposition into covariant components $a_\beta$.](image)

I.e., $v = a_\alpha S_\alpha$, where $a_\alpha = S_\alpha \cdot v$. With angle $\gamma$ between $v$, and $S_\alpha$, the length of the orthogonal projection of $v$ upon $S_\alpha$ will be $||v|| \cos \gamma = S_\alpha \cdot v / \sqrt{S_\alpha \cdot S_\alpha} = a_\alpha / \sqrt{g_{\alpha\alpha}}$. I.e., $a_\alpha$ is the length of the orthogonal projection of $v$ upon the $S_\alpha$ axis of $E(P)$ measured in units of $1 / \sqrt{g_{\alpha\alpha}}$ depicted to the right in Fig. 1.\]

1 Also named fundamental tensor or Riemannian metric.
Covariant and contravariant components of $v$ are related:
\[ a_\alpha = S_\alpha \cdot v = S_\alpha \cdot a^\beta S_\beta = g_{\alpha \beta} a^\beta. \] (32)

Since $g = \det(g_{\alpha \beta}) \neq 0$ one can solve w.r.t. $a^\beta$ to obtain
\[ a_\beta = g^{\alpha \beta} a^\alpha, \] (33)
where
\[ g^{11} = \frac{g_{22}}{g}, \quad g^{12} = \frac{g_{21}}{g}, \quad g^{22} = \frac{g_{11}}{g} \] (34)
and $g_{\alpha \beta} g^{\gamma \beta} = \delta_\alpha^\gamma$, $\alpha^\beta$ are the components of a 2nd order contravariant tensor named contravariant metric tensor.

In a Cartesian coordinate system $g_{\alpha \beta} = \delta_\alpha^\beta$ and $g^{\alpha \beta} = \delta^\beta_\alpha$ and so $a^\alpha = a_\alpha$.

**Second Fundamental Form (SFF):** At any point of a curve $\mathcal{C}$ on a surface $\mathcal{S}$, the corresponding unit normal vector to $\mathcal{S}$
\[ n = \frac{S_1 \times S_2}{\|S_1 \times S_2\|} = \frac{S_1 \times S_2}{-\sqrt{g}}, \] (35)
lies in the normal plane to $\mathcal{C}$ which also contains the principal normal $p$ to $\mathcal{C}$ [4 p.118] (see Section III-A). The angle $\gamma$ between these two vectors depend on the geometric shape of both $\mathcal{C}$ and $\mathcal{S}$ in a neighborhood of the point $P$ under consideration. We have two extremes: i) $\gamma = \pi/2$ for all points along $\mathcal{C}$, implying that $p \perp n$, and so $\mathcal{C}$ is a plane curve with $\mathcal{S}$ a plane (flat space with zero curvature). ii) $\gamma = 0$ for all points along $\mathcal{C}$, implying that $p\|n$, and then $\mathcal{C}$ is a geodesic on $\mathcal{S}$ (see discussion on geodesics in section III-B1). That is the arc with the shortest possible length between its endpoints on $\mathcal{S}$. This is the equivalent of a straight line in a plane (like the great circles on a sphere), and is generally the path that minimizes energy.

We now assume that $C$ is represented by arc length parametrization $u^1(t), u^2(t)$. Since $n$ and $p$ are unit vectors, $\cos \gamma = p \cdot n$ ($\kappa > 0$). This scalar product will generally vary along $\mathcal{C}$. From the formula of Frenet (6) we get
\[ \kappa \cos \gamma = \hat{S} \cdot n. \] (36)

From the product rule we have that
\[ \hat{S} = \frac{\partial S}{\partial u^1} du^1 + \frac{\partial S}{\partial u^2} du^2 = S_\alpha \dot{u}^\alpha. \] (37)

Differentiating w.r.t. $\ell$ again, we get
\[ \ddot{S} = S_{\alpha \beta} \dot{u}^\alpha \dot{u}^\beta + S_\alpha \ddot{u}^\alpha. \] (38)

Since $S_\alpha \cdot n = 0$, Eq. (38) becomes
\[ \kappa \cos \gamma = (S_{\alpha \beta} \cdot n) \dot{u}^\alpha \dot{u}^\beta. \] (39)

The term in the parentheses is an important quantity,
\[ b_{\alpha \beta} = S_{\alpha \beta} \cdot n, \quad \alpha, \beta = 1, \cdots, N. \] (40)

The scalar products $b_{\alpha \beta}$ depend on $\mathcal{S}$ only (i.e. independent of choice of curve $\mathcal{C}$) and are symmetric due to the symmetry of $S_{\alpha \beta}$. The quadratic form
\[ b_{\alpha \beta} du^\alpha du^\beta, \] (41)
is called the second fundamental form (SFF). The SFF is invariant w.r.t. allowable coordinate transformations which preserves the sign (or sense) of $n$. $b_{\alpha \beta}$ are therefore the components of a 2nd order covariant tensor (like $g_{\alpha \beta}$).

An alternative way to obtain $b_{\alpha \beta}$ is by differentiating $S_\alpha \cdot n = 0$. That is, $S_{\alpha \beta} \cdot n + S_\alpha \cdot n_\beta = 0$, where $n_\beta = \partial n/\partial u^\beta$. Therefore
\[ b_{\alpha \beta} = -S_\alpha \cdot n_\beta, \quad \alpha, \beta = 1, \cdots, N. \] (42)

To compute $b_{\alpha \beta}$, the following relation is convenient to use:
\[ b_{\alpha \beta} = S_{\alpha \beta} \cdot n = S_{\alpha \beta} \left( \frac{S_1 \times S_2}{\sqrt{g}} \right) = \frac{1}{\sqrt{g}} |S_1 S_2 S_{\alpha \beta}|. \] (43)

Let now $t$ be any allowable parameter for the curve $\mathcal{C}$. Then
\[ \dot{u}^\alpha = \frac{du^\alpha}{dt} = u^\alpha', \] (44)

Then, in the consequence of (41), Eq. (36) becomes
\[ \kappa \cos \gamma = b_{\alpha \beta} u^\alpha u^\beta = b_{\alpha \beta} u^\alpha du^\beta = g_{\alpha \beta} du^\alpha du^\beta. \] (45)

**Normal curvature, Principal curvature, Lines of curvature:**

We first seek a geometric interpretation of (45). For any fixed point $P \in \mathcal{S}$, $g_{\alpha \beta}$ and $b_{\alpha \beta}$ are fixed and independent of any curve $\mathcal{C}$ on $\mathcal{S}$ passing trough $P$. Therefore, the curvature $\kappa$ of $\mathcal{C}$ at $P$ will only depend on the directions of the tangent $t$ and the principal normal $p$. Since the osculating plane is spanned by $t$ and $p$, then any curve that has the same osculating plane at $P$, will have the same curvature $\kappa$. This includes the curve of intersection between this common osculating plane and $\mathcal{S}$, and therefore one can restrict the investigation to plane curves on $\mathcal{S}$.

Consider now all curves that has the same tangent direction $t_\ell$ at $P$. For these curves, the right hand side of (45) is constant implying that their curvature depend only on the angle $\gamma$. Therefore, for these curves, $\kappa \cos \gamma = \kappa_n$, with $\kappa_n$ a constant. If $\gamma = 0$ then $\kappa = \kappa_n$, and if $\gamma = \pi$ then $\kappa = -\kappa_n$, implying that $|\kappa_n|$ is the curvature of the curve of intersection between $\mathcal{S}$ and a plane passing through both the tangent to $\mathcal{C}$ at $P$ and the normal to $\mathcal{S}$ at $P$. These curves of intersection are called normal sections of $\mathcal{S}$. $\kappa_n$ is called normal curvature of $\mathcal{S}$ at $P$ corresponding to the direction of the tangent to the normal section at $P$. We then have the normal curvature vector $\kappa_n = \kappa_n n$. From this discussion and (45) we see that
\[ \kappa_n = \frac{b_{\alpha \beta} du^\alpha du^\beta}{g_{\alpha \beta} du^\alpha du^\beta}. \] (46)

One can now set $\kappa_n = 1/\rho$ where $|\rho|$ is the radius of curvature of the corresponding normal section at $P$. Since $\kappa = 1/\rho$ we can write
\[ \rho = R \cos \gamma. \] (47)

We have the Theorem of Meusnier [4 p.122]

*Theorem 1:* The centre of curvature of all curves on $\mathcal{S}$ passing through an arbitrary point $P$ and whose tangent at $P$ have the same direction (different from an asymptotic direction, where $b_{\alpha \beta} du^\alpha du^\beta = 0$), lie on a circle $K$ of radius $|\rho|/2$ which lies in the normal plane and has (at least) first
order contact with $S$ at $P$.

This implies that we can restrict the consideration to normal sections of $S$ at $P$ without loss of generality.

A special point is when $b_{\alpha\beta} \sim g_{\alpha\beta}$ named navel point or umbilic. For points that are not umbilics, we seek directions where $\kappa_n$ has extreme values: The formula for normal curvature \(46\) can be rewritten as

$$ (b_{\alpha\beta} - \kappa_n g_{\alpha\beta}) du^\alpha du^\beta = 0. \quad (48) $$

By differentiating w.r.t. $du^\gamma$, treating $\kappa_n$ as a constant, one obtains \[4\] pp.128-129]

$$ (b_{\alpha\gamma} - \kappa_n g_{\alpha\gamma}) du^\alpha = 0, \quad \gamma = 1, 2. \quad (49) $$

By eliminating $\kappa_n$, then

$$ \begin{vmatrix} g_{1\alpha} du^\alpha & b_{1\beta} du^\beta \\ g_{2\alpha} du^\alpha & b_{2\beta} du^\beta \end{vmatrix} = 0 \quad (50) $$

or

$$ \begin{vmatrix} (du^\gamma)^2 & -du^1 du^2 (du^1)^2 \\ g_{11} & g_{12} & g_{22} \\ b_{11} & b_{12} & b_{22} \end{vmatrix} = 0 \quad (51) $$

The roots of \(50\) determine the directions for which $\kappa_n$ is extreme, named principal directions of normal curvature at $P$. The corresponding curvature values, $\kappa_1$ and $\kappa_2$, are the principal normal curvatures of $S$, corresponding the maximal and minimal curvature of $S$ at $P$ respectively. We have the following theorem \[4\] p.129):

**Theorem 2:** The roots of \(50\) are real, and at every point not an umbilic, the principal directions are orthogonal.

**Proof:** See \[4\] p.129].

A curve on $S$ whose direction at every point is a principal direction is known as a line of curvature (LoC) of $S$. The LoC is a solution to the differential equation in \(51\). We have two LoC at each point $P$ (not umbilic), where the following theorem is satisfied \[4\] p.130):

**Theorem 3:** The LoC on any surface $S \in C^r, \ r \geq 3$, are real curves. If $S$ has no umbilics, the LoC form an orthogonal net everywhere on $S$.

**Proof:** See \[4\] p.130].

One may always choose coordinates $u^1, u^2$ on $S$ so that the LoC are allowable coordinates at any point of $S$ not umbilic. We have \[4\] p.130):

**Theorem 4:** The coordinate curves of any allowable coordinate system on $S$ coincide with the LoC $\iff$

$$ g_{12} = 0 \quad \text{and} \quad b_{12} = 0, \quad (52) $$

at any point where those coordinates are allowable.

**Proof:** See \[4\] p.130].

To derive an analytical expression for the principal curvatures, multiply \(49\) with $g^{\beta\gamma}$ and sum w.r.t. $\gamma$ to obtain

$$ b_{\alpha\beta} du^\alpha - \kappa_n du^\beta = 0, \quad \beta = 1, 2, \quad (53) $$

which after some algebraic manipulation leads to \[4\] p.130]

$$ \kappa_n^2 - b_{\alpha\beta} g^{\alpha\beta} \kappa_n + \frac{b}{g} = 0. \quad (54) $$

The principal curvatures $\kappa_1$ and $\kappa_2$, are the roots of this equation.

We have two important definitions \[4\] p. 131):

**Definition 5:** The product

$$ K = \kappa_1 \kappa_2 = \frac{b}{g} \quad (55) $$

is the Gaussian Curvature of $S$ at $P$.

**Definition 6:** The arithmetic mean

$$ H = \frac{1}{2}(\kappa_1 + \kappa_2) = \frac{1}{2} b_{\alpha\beta} g^{\alpha\beta} = \frac{1}{2} b_{\alpha\alpha}, \quad (56) $$

is the Mean Curvature of $S$ at $P$.

Both $K$ and $|H|$ are invariant under allowable coordinate transformations.

When the coordinate curves are LoC, \(53\) holds with $\kappa = \kappa_1$, $du^2 = 0$ and again with $\kappa = \kappa_2$, $du^1 = 0$. Therefore

$$ \kappa_1 = b_{11}^2, \quad \kappa_2 = b_{22}^2, \quad \kappa_2 = b_{22}^2, \quad b_{22}^2 = 0. \quad (57) $$

For this special case

$$ \kappa_1 = \frac{b_{11}}{g_{11}}, \quad \kappa_2 = \frac{b_{22}}{g_{22}}, \quad K = \frac{b_{11} b_{22}}{g_{11} g_{22}}, \quad H = \frac{1}{2}(\frac{b_{11}}{g_{11}} + \frac{b_{22}}{g_{22}}). \quad (58) $$

The normal curvature $\kappa_n$ for any (tangent) direction can be represented in terms of the principal curvatures $\kappa_1$ and $\kappa_2$ according to the Theorem of Euler \[4\] p. 132):

**Theorem 5:** Let $\alpha$ be the angle between an arbitrary direction at $P$ and the direction at $P$ corresponding to $\kappa_1$. Then

$$ \kappa_n = \kappa_1 \cos^2 \alpha + \kappa_2 \sin^2 \alpha. \quad (59) $$

**Proof:** See \[4\] p.132].

We have one last theorem concerning LoC \[4\] p.191):

**Theorem 6:** With the exception of umbilics, the family of LoC for $S$ of class $r \geq 3$ will be orthogonal and conjugate.

**Proof:** See \[4\] p.191].

Note: Two directions $du^1 : du^2$ and $\delta^1 : \delta^2$ are conjugate $\Leftrightarrow b_{\alpha\beta} du^\alpha \delta^\beta = 0$ (see Theorem 60.2 \[4\] p.191))

3) Formulae of Weingarten and Gauss: For curves in $\mathbb{R}^3$ one can at each point $P$ associate three orthogonal unit vectors, $\mathbf{t}$, $\mathbf{p}$ and $\mathbf{b}$. Any vector bound at $P$ will be a linear combination of these vectors, like the derivatives $\mathbf{t}$, $\mathbf{p}$, $\mathbf{b}$ related through the formulae of Frenet. For surfaces in $\mathbb{R}^3$ one have similar relations for the vectors $S_1$, $S_2$ and $n$. Relations for the derivatives of these vectors are provided by the formulae of Weingarten (FoW) and formulae of Gauss (FoG):
The FoW relates the derivative of the normal vector to \( x_\beta \) as shown in [4, p. 139]
\[
  n_\alpha \equiv \frac{\partial n}{\partial u^\alpha} = -g^{\alpha\beta} b_\alpha S_\beta = -b_\alpha S_\beta, \quad \alpha = 1, 2.
\]  
(60)
If LoC coordinate curves are chosen, then FoW reduces to
\[
  n_\alpha = \frac{b_\alpha}{g_{\alpha\alpha}} S_\alpha, \quad \alpha = 1, 2.
\]  
(61)
The FoG relates the 2nd derivatives \( S_{\alpha\beta} \) to \( S_\gamma \) and \( n \) as shown in [4, p.140-142],
\[
  S_{\alpha\beta} \equiv \frac{\partial^2 S}{\partial u^\alpha \partial u^\beta} = \Gamma_{\alpha\beta}^\gamma S_\gamma + b_\alpha n, \quad \alpha, \beta = 1, 2,
\]  
(62)
where
\[
  \Gamma_{\alpha\beta}^\gamma = g^{\lambda\kappa} \Gamma_{\alpha\beta\kappa},
\]  
(63)
are Christoffel symbols of the 2nd kind (CS-2), and where
\[
  \Gamma_{\alpha\beta\lambda} = \frac{1}{2} \left[ \frac{\partial g_{\beta\lambda}}{\partial u^\alpha} + \frac{\partial g_{\alpha\lambda}}{\partial u^\beta} - \frac{\partial g_{\alpha\beta}}{\partial u^\lambda} \right],
\]  
(64)
are Christoffel symbols of the 1st kind (CS-1).

We start off with results concerning curves, that is, the mapping of \( x \) can then be extended to surfaces. The basics of S-K mappings are provided in [3] where distortion analysis is formulated w.r.t. the metric tensor, or FFF. Here we also take curvature into account by considering the SFF (see Section II-B2). We begin with the definition of S-K mappings:

**Definition 7: Shannon-Kotel’nikov mapping**
An S-K mapping \( S \) is a continuous or piecewise continuous nonlinear or linear mapping between \( \mathbb{R}^M \) (source space) and \( \mathbb{R}^N \) (channel space). There are three cases to consider:
1. Equal dimension \( M = N \): \( S \) is a bijective mapping.
2. Dimension expansion \( M < N \): \( S \subseteq \mathbb{R}^N \) is a mapping that can be realized by a hyper surface described by the parametrization
\[
  S(x) = [S_1(x), S_2(x), \ldots, S_N(x)],
\]  
(65)
where each source vector \( x \) should have a unique representation \( S(x) \in S \). \( S \) is then an \( M \)-dimensional (locally Euclidean) manifold embedded in \( \mathbb{R}^N \).
3. Dimension reduction \( M > N \): \( S \subseteq \mathbb{R}^M \) is a mapping that can be realized by a hyper surface described by the parametrization
\[
  S(z) = [S_1(z), S_2(z), \ldots, S_M(z)],
\]  
(66)
\[\text{MMSE decoding is needed at low SNR in order to obtain optimality, effectively weakening this condition.}\]
\[\text{This is not a restriction, i.e. the mapping does not need to be described by a parametrization.}\]

where each channel vector \( z \) should have a unique representation \( S(z) \in S \). \( S \) is then an \( N \)-dimensional (locally Euclidean) manifold embedded in \( \mathbb{R}^M \).

**A. Higher order distortion analysis of S-K mappings using parametric curves**

We treat \( M : 1 \) SK-mappings first. Then use the intuition gained to treat \( 1 : N \) mappings.

1) \( M : 1 \) dimension reducing mappings: In order to account for curvature and higher order behavior of S-K mappings one has to consider the Taylor expansion beyond 1st order. We consider 3rd order Taylor expansion of a curve in \( \mathbb{R}^3 \) here and draw the conclusion that 2nd order behaviour is adequate for the investigation of S-K mappings at high SNR in general.

In the following we denote the derivatives w.r.t. arc length as \( \dot{S} \), \( \ddot{S} \) and \( \dddot{S} \) and w.r.t to a general parameter as \( S' \), \( S'' \) and \( S''' \). With \( x_0 \) transmitted and noise \( n \) we have from the 3rd order Taylor expansion:
\[
  S(x_0+n) \approx S(x_0)+nS'(x_0)+\frac{n^2}{2}S''(x_0)+\frac{n^3}{3!}S'''(x_0).
\]  
(67)
From this we can derive the channel distortion as
\[
  \varepsilon_{ch}^2(x_0) = E\left\{ \left\| nS'(x_0)+\frac{n^2}{2}S''(x_0)+\frac{n^3}{3!}S'''(x_0) \right\|^2 \right\}.
\]  
(68)
To expand this expression and take the expectation, it is advantageous to use arc length parametrization (see Appendix A).

Then the first, second and third derivatives will at each point of \( S \) make out a vector space of mutually orthogonal vectors as explained in Section II-A1 (and with more detail in [4, pp. 36-37]). Since \( \dot{S} \cdot \ddot{S} = 0 \), \( \dddot{S} = 0 \), and \( \dddot{S} \cdot \dddot{S} = 0 \), and using the fact that [6, p.148]
\[
  E\{n^a\} = \begin{cases} 1 \cdot 3 \cdot \cdots (a-1)\sigma_n^a, & a \text{ even} \\ 0, & a \text{ odd} \end{cases}
\]  
(69)
for \( n \) Gaussian, the expectation of the norm in (68) can be found from straightforward calculations
\[
  \varepsilon_{ch}^2(x_0) = \sigma_n^2 \| \dot{S}(x_0) \|^2 + \frac{3\sigma_n^4}{4} \| \dddot{S}(x_0) \|^2 + \frac{5\sigma_n^6}{12} \| \dddot{S}(x_0) \|^2,
\]  
(70)
where the first norm \( \| \dot{S}(x_0) \| = 1 \) according to Theorem 19 in Appendix A. Using the canonical representation defined in (10) in Section II one can express the channel distortion up to 3rd order [4, p.48], valid for any curve \( s \in C^3 \) (of class \( C^r, r \geq 3 \)):
\[
  \varepsilon_{ch}^2(x_0) = E\left\{ \left\| n \frac{\kappa_0}{2} n^2 \frac{\kappa_0 \tau_0}{6} n^3 \right\|^2 \right\}
\]  
(71)
We can conclude from (70) and (71) that the 1st order approximation dominates when \( \sigma_n \) is small, that is at high SNR, as long as the curvature \( \kappa \) and torsion \( \tau \) are not large. That is, \( \sigma_n^4 \) and \( \sigma_n^6 \) goes to zero much faster than \( \sigma_n^2 \), making the 1st order channel distortion approximation accurate at high SNR. As we will see in Section III-A3 one should not
choose a curve with large curvature or torsion to obtain a good mapping (large relative to transmission power and the curvature of the "noise sphere"). With this choice one can safely discard the 3rd order term as the 2nd order term will dominate.

The ML-estimate error also increases with $\kappa$ and $\tau$ as seen from (71) as $\kappa_0 = \tau_0 = 0$ minimizes the error. However, we cannot avoid nonzero $\kappa$ and $\tau$ if we are to fill the space in question in order to keep the approximation distortion low. One should therefore choose a curve that fills the space but at the same time has the smallest possible $\kappa$ and $\tau$.

In general, the curvature is given by (4), leading to a more complicated relations than the ones above. Since we most often will consider arc length parametrization with an amplification $\alpha$, which we name scaled arc length parametrization, we have $\|S'(x_0)\| = \alpha\|S(x_0)\| = \alpha$, $\forall x_0$. Then (4) reduces to $\kappa(x_0) = \|S''_0(x_0)\|/\|S'_0\|^2$ (since $S'_0 \perp S''_0$, still). Therefore we can express the channel error up to second order as

$$
\varepsilon_{cha}^2(x_0) = \sigma_n^2 \|S'_0\|^2 + 3\sigma_n^4 \|S''_0\|^2/4 \|S'_0\|^2 = \alpha^2 \sigma_n^2 + \frac{3\sigma_n^4 \kappa^2(x_0)}{4}. \tag{72}
$$

2) $1:N$ dimension expanding mappings: It is more complicated to compute higher order terms for the error in the expansion case as the ML decoder is a projection onto $S$. We illustrate this for 2nd order approximation: From the Taylor series we have up to 2nd order

$$
S(x) \approx S(x_0) + S'_0 \cdot (x - x_0) + \frac{S''_0}{2}(x - x_0)^2, \tag{73}
$$

where

$$
S'_0 = \frac{dS}{dx}_{|x=x_0}. \tag{74}
$$

With $z = S(x_0) + n$, the ML estimate seek to minimize $\|z - S(x)\|$, which up to 2nd order is the same as

$$
\hat{x}_{ML} = \arg\min_x \|n - S'_0 \cdot (x - x_0) - \frac{S''_0}{2}(x - x_0)^2\|^2. \tag{75}
$$

It is straight forward to show that

$$
\frac{d}{dx} \|n - S'_0 \cdot (x - x_0) - \frac{S''_0}{2}(x - x_0)^2\|^2 = \frac{1}{2} \|S''_0\|^2(x - x_0)^2 + \frac{1}{2} \|S'_0\|^2(x - x_0) - \cdots \tag{76}
$$

$$
n \cdot S'_0 - n \cdot S'_0(x - x_0). \tag{77}
$$

Let now $a = 2(\|S'_0\|^2 - n \cdot S'_0), b = 2n \cdot S'_0/\|S'_0\|^2$, and $w = x - x_0$. Then the minimization problem in (73) is the solution to $w^3 + aw + b = 0$, which is a 3rd order polynomial equation that can be solved using Cardano’s Method. However, this does not result in a simple expression that one can analyze further. It is therefore necessary to use another approach.

Using the concept of the osculating circle in Section HI-A we can apply the higher order analysis of distortion for pulse position modulation (PPM) in [7] pp. 703-704 to obtain a result valid for any $1:N$ mapping. PPM is geometrically described by a curve on a hyper sphere where the arc between any two coordinate axes is like a circle segment as depicted in Fig. 2. The fact that any curve has a local spherical geometry implies that the analysis done for PPM is also valid locally for any other $1:N$ mapping under (scaled) arc length parametrization. In the following we name the signal curve segment in Fig. 2 circle approximation.

In polar coordinates the signal curve can be written

$$
S(x) \approx [R(x), \theta(x)]. \tag{78}
$$

Let $e_r$ be a unit basis vector in the radial direction and $e_\theta$ a unit basis vector in the angular direction in Fig. 2. We further divide the noise into two components $\text{Proj}(n) = n_\parallel$, that is the projection onto the closest point on the circle in Fig. 2 and its normal, $n_\perp$.

Further, for the circle approximation

$$
R(x) = \rho(x) = \frac{1}{\kappa(x)} = \frac{1}{\|S'_0\|}. \tag{79}
$$

iff $t = \|S'_0\| = 1$. That is, under arc length parametrization. For PPM we have a constant radius $R(x) = R$ and the approximation in (78) becomes exact, that is, $S = [R\cos(\theta(x)), R\sin(\theta(x))]$, and so

$$
\frac{dS}{dx} = \left[R\sin(\theta(x))\frac{d\theta(x)}{dx}, -R\cos(\theta(x))\frac{d\theta(x)}{dx}\right]. \tag{80}
$$

By taking the norm we find that $\|dS/dx\|^2 = R^2d\theta(x)/dx$. From this we get

$$
\frac{d\theta(x)}{dx} = \frac{1}{R} \left\|\frac{dS}{dx}\right\| = \frac{\|S'_0\|}{\rho} = \alpha \kappa = \alpha\|S''_0\|. \tag{81}
$$

That is, the rate of change of $S$ is proportional to $\kappa$. Note that the two last equalities in (81) are valid only for scaled arc length parametrization.

To determine the ML estimate for the circle approximation we rewrite (81) as $dx = d\theta(x)/(\|S'_0\|\|S''_0\|) = \|S'_0\|\|S''_0\|$. Therefore

$$
x_0 - \hat{x}_{ML} = \frac{\|S''_0\|}{\kappa(x_0)}\theta = \frac{\theta}{\kappa(x_0)}\|S'_0\| \tag{82}
$$
We need to express $\theta$ in terms of the noise and $\rho$ (or $\kappa$). We use Fig. 2 for this purpose. We have a right-legged triangle, i.e. $\theta + 2\varphi = \varpi + \phi = (\varpi - \theta)/2$. Further, $b = \rho \sin(\theta)$ is the right normal of the triangle. Therefore

$$\sin(\phi) = \frac{b}{\|n\|} \Rightarrow \|n\| = \frac{b}{\sin(\phi)} = \frac{\rho \sin(\theta)}{\sin(\phi)}. \quad (83)$$

Further,

$$\sin(\phi) = \sin\left(\frac{\pi}{2} - \frac{\theta}{2}\right) = \cos\left(\frac{\theta}{2}\right). \quad (84)$$

Since $\sin 2y = 2 \sin y \cos y$, with $y = \theta/2$, then $\sin(\theta) = 2 \sin(\theta/2) \cos(\theta/2)$, implying that

$$\|n\| = \frac{2\rho \sin(\theta/2) \cos(\theta/2)}{\cos(\theta/2)} = 2\rho \sin\left(\frac{\theta}{2}\right). \quad (85)$$

Therefore

$$\theta = 2 \sin^{-1}\left(\frac{\|n\|}{2\rho(x_0)}\right). \quad (86)$$

By the Taylor expansion [8 p. 117]

$$\sin^{-1}(x) = x + \frac{1}{2} \cdot 3^3 \cdot \frac{1}{2} \cdot 4 \cdot 5 \cdot x^5 + \cdots \quad (87)$$

we get $\theta$ up to 2nd order as

$$\theta \approx \frac{\|n\|}{\rho(x_0)} \left(1 + \frac{\|n\|^2}{24\rho^2(x_0)}\right). \quad (88)$$

One can then compute the error (in the absence of anomalies) up to second order from (82) and (88)

$$\varepsilon_{wn}^2 = E\left\{ (x - \hat{x}_{M.L})^2 | x = x_0 \right\} = E\left\{ \frac{\rho(x_0)}{\|S\|} \theta \right\} \quad (89)$$

$$\begin{align*}
\rho^2(x_0) & \frac{\|n\|}{\rho(x_0)} \left(1 + \frac{\|n\|^2}{24\rho^2(x_0)}\right)^2 \quad (90) \\
\frac{\|n\|^2}{\|S\|^2} & \frac{\|n\|^2}{24\rho^2(x_0)} + \frac{\|n\|^6}{242\rho^4(x_0)} \quad (91)
\end{align*}$$

Note that the relation $\kappa(x_0) = \|\tilde{s}(x_0)\|$ is valid under arc length parametrization, that is whenever $\|s'(x_0)\| = \|\tilde{s}(x_0)\| = 1, \forall x_0$. In general, the curvature is given by (4), leading to a more complicated expression. Since we most often consider arc length parametrization with an amplification $\alpha$, that is $\|S'(x_0)\| = \alpha \|\tilde{s}(x_0)\| = \alpha, \forall x_0$, then (4) reduces to $\kappa(x_0) = \|S''(x_0)/\|S\|^2\|$ (since $S'_0 \perp S''_0$ still). In that case we can express the error in terms of the signal curve’s derivatives as

$$\varepsilon_{wn}^2 = \frac{\sigma_n^2}{\|S\|^2} \left(1 + \frac{\sigma_n^2}{4\rho^2(x_0)} + \frac{5\sigma_n^4}{48\rho^4(x_0)}\right)$$

Note that high SNR we can stretch the curve significantly, that is $\|S\|| = \alpha >> 1$, and we see from (92) that the first term clearly dominates, making the accuracy of the 1st order term increasingly dominant with SNR. As was seen in Section III-A1 higher order terms will contribute even less, and therefore the circle approximation above will do.

Further note the similarity to the $M:1$ case given in (72), the distortion is scaled by the curvature in nearly the same manner. Also, distortion in the expansion case is proportional to $1/\|S\|^2$, whereas distortion in the reduction case is proportional to $\|S\|^2$, as expected for the weak noise regime.

It is straight forward to see that the circle approximation above will be valid locally for any signal curve $S$ using the concept of osculating circle. We illustrate with Fig. 3. For a general curve in polar coordinates we have by the product rule

$$\frac{dS}{dx} = \left[ R'(x) \cos(\theta(x)) + \cdots \quad (93) \right.$$ 

$$+ R(x) \sin(\theta(x)), R'(x) \sin(\theta(x)) - R(x) \cos(\theta(x)) \right].$$

When $\rho(x) = \rho = \text{constant}$ we get the result derived above for all $x$. Looking at the curve locally, $R'(x)$ is small if the curvature of $S$ changes slowly with $x$. The more slowly $\kappa(x)$ changes with $x$, the more accurate the circle approximation becomes. As we shall see in Section III-A3 it is not convenient to choose a curve with a rapidly changing curvature $\kappa$. This implies that Shannon’s suggestion for a $1:2$ mapping in (9) may not be the most convenient due to the "half circles" connecting each line segment.

To be a bit more rigorous: For PPM we had $S(x_0) = S_r e_r$. In general we have $S(x_0) = S_r e_r + S_0 e_0$. For the case $\|S\|^2 = \alpha, \forall x$, we have seen that $S_0' \cdot S_0'' = 0$. Then locally around $S(x_0)$ we again have $S_0' = \alpha e_0$, and since $S_0' \perp S_0''$ we must have that $S_0'' = \alpha e_r$. From the formula of Frenet (FoF) we then have

$$\tilde{S}_0 = \kappa(x_0) p.$$
Further, \( S_e(z_i, \ell) = 0, \frac{\partial S_e(z_i, \ell)}{\partial \ell} = 0, \frac{\partial^2 S_e(z_i, \ell)}{\partial \ell^2} = 0. \) (95)

where \( S_e = 0 \) defines a surface in \( \mathbb{R}^3 \) (or hypersurface in \( \mathbb{R}^N \)). The characteristic is therefore a curve in \( \mathbb{R}^3 \) (or a hypersurface of dimension \( N - 2 \) in \( \mathbb{R}^N \)). The characteristic points of the canal surface is the intersection of the characteristic (which here will be points of intersection of circles) given by [4, p. 266] \( S_e(z_i, \ell) = 0, \frac{\partial S_e(z_i, \ell)}{\partial \ell} = 0, \frac{\partial^2 S_e(z_i, \ell)}{\partial \ell^2} = 0. \) (96)

The family \( F \) of spheres with constant radius \( r \) with center on a curve \( C : y(\ell) \), can be represented as

\[
S_e(z_i, \ell) = (z - y(\ell)) \cdot (z - y(\ell)) - r^2 = 0.
\] (97)

To recapitulate in the terminology of S-K mappings one can set \( y(\ell) = S(\ell) \) and \( S_e(z_i, \ell) \), where \( z \) is the channel coordinates and \( \ell \) is the arc length of \( S \). Further, let the curvature of \( S \) be \( \kappa_s \) with corresponding radius of curvature \( \rho_s = 1/\kappa_s \). For 1:N S-K mappings we seek a canal surface that does not intersect itself in order to avoid anomalous distortion. This is the case if and only if the following conditions are satisfied

i) The minimum distance between two folds of \( S \) is \( \geq 2r \)

ii) \( \rho_s > r \), as shown in Proposition [1] below.

**Proposition 1:** For \( \rho_s = 1/\kappa_s \), radius of curvature for \( S(x) \), and \( r \), the radius of the hypersphere \( S^{N-2} \). Then the corresponding canal surface will not intersect itself, that is, the canal surface will have no characteristic points \( \iff \rho_s > r \).

**Proof 1:** (See [4] pp.266-267) The spheres \( S^{N-2} \) in \( F \) can be represented as

\[
S_e(z_i, \ell) = (z - S(\ell)) \cdot (z - S(\ell)) - r^2 = 0, \ i = 1, \ldots, N.
\] (98)

Further,

\[
\frac{\partial S_e}{\partial \ell} = -2S \cdot (z - S) = 0, \ S = \frac{\partial S}{\partial \ell},
\] (99)

this implies that \( (z - S) \perp S \) and so,

\[
\frac{\partial^2 S_e}{\partial \ell^2} = -2\ddot{S} \cdot (z - S) + 2 = -2\kappa_s \dot{p} \cdot (z - S) + 2 = 0,
\] (100)

where the last equality is due to the formula of Frenet (6) with \( \rho_s = 1/\kappa_s \) we get

\[
p \cdot (S - z) - \rho_s = 0.
\] (101)

\( \Rightarrow \): The condition follows directly from (100). \( \iff \): Since \( \| p \cdot (S - z) \| = \| S - z \| = r \), one can see that if \( \rho_s > r \ \forall x \), then the last equation in (100) does not have any real solution. I.e., no real characteristic points exists.

Note that the above concept is more or less the same as a 1:N signal curve in Gaussian noise, where the noise vectors will approximately lie within congruent spheres along \( S(x) \). We will elaborate a bit further for a 1:3 mapping: Fig. 4 shows a canal surface surrounding a signal curve in \( \mathbb{R}^3 \). The radius of the tube is linked to Gaussian noise. A correspondence to the noise in the above discussion results by setting \( n_w = n \) and \( n_{wn} = n_\perp \). We can see from this than any bending of the tube increases the probability for anomalous errors, implying that a straight line has the lowest possible probability. However, it is well known that at high SNR this will produce a sub-optimal mapping due to increased weak noise error. By linking to Proposition [1] and the above discussion, it is easy to see that bending poses no problem if the radius of curvature of \( S \) is small enough: By setting

\[
\rho_s > r = \rho_n \geq \sqrt{\frac{N - 1}{N} b_n \sigma_n^2},
\] (102)

then no characteristic points will exist, and the tube will not intersect itself (if condition i) above is also satisfied).

Of course with noise, \( \rho_n \) usually has no limitation, and the radius of the noise sphere is always linked to a certain probability. However, by setting \( b_n = 4 \), then the probability that \( \rho_n > \rho_s \) will be very small. However, when proving optimality in the limit of infinite dimensionality, then \( \rho_n \to \sigma_n \) with unit probability, meaning that the canal surface concept above is crucial in proving optimality in the limiting case. The concept of canal surfaces extended to general signal hyper surfaces will be crucial when setting down criteria for the construction of well performing mappings later. To summarize the discussion on curves we see that as the SNR gets high, the 1st order approximation (what we later name weak noise regime) takes off with \( \sigma_n^2 \), whereas the largest higher order term takes off with \( \sigma_n^4 \), and therefore becomes negligibly small as \( \sigma_n \to 0 \) (at high SNR) if the curvature is not large, and the 1st order approximation dominates. From the discussion of maximum likelihood estimation and canal surfaces above, one should keep the curvature as small as possible, never larger than \( 1/\rho_n \) (in fact \( \rho_s >> \rho_n \) in a practical situation). This will keep the curvature \( \kappa_s \) in
eqns (72) and (92) at a small enough level for the 2nd order
term to vanish much faster than the 1st order term.

B. S-K mappings using parametric surfaces

Many new concepts concerning differential geometry of surfaces
will be applied here. All relevant concepts are described in 
Section II-B2 A more thorough treatment is provided in [4].

1) M:N dimension expanding mappings: We start off with 
surfaces in $\mathbb{R}^3$, that is, 2:3 mappings. According to theorem of
Meusnier and theorem of Euler the curvature can in all cases be
described by a combination of the principal curvatures $\kappa_1$ and $\kappa_2$
(see Theorems 1 and 5 in Section II-B2). These correspond to the
minimal and maximal curvatures at any point $P \in S$ respectively.
Since the principal curvatures are extremal curvatures, they can be used to analyze higher order
behavior of S-K mappings beyond curves: As it is the maximal
curvature, $\kappa_1$, that bounds the probability of anomalous errors,
it has to be limited to a certain maximum value. To analyze
curvature of S-K mappings one has to consider both FFF and
SFF introduced in Section II-B2.

To make the math simpler one can choose lines of curvature
(LoC) as coordinates on $S$. That is, the coordinates are
aligned with the direction of $\kappa_1$ and $\kappa_2$ at any point (i.e.,
normal sections along principal directions). We first develop a
expression for $\varepsilon_{\text{un}}^2$ when the coordinate curves are chosen as
LoC, since this is the most direct and simple generalization of
the 1:N case. We will then discuss the impact of performance
in choosing LoC as coordinates at the end of this section.

Assume that $x_1$ and $x_2$ are along the LoC. Then, according
to Theorem 4 in Section II-B2 $g_{12} = b_{12} = 0$ has to be
satisfied. Then the curvature is given solely in terms of $g_{\alpha\alpha}$
and $b_{\alpha\alpha}$ as

$$\kappa_i(x_0) = \frac{b_{ii}(x_0)}{g_{ii}(x_0)}, \quad \text{where } g_{ii} = \left\| \frac{\partial S}{\partial x_i} \right\|^2, \quad (103)$$

and where (see Eq. (93))

$$b_{ii} = \frac{1}{\sqrt{g}} \left| \frac{\partial S}{\partial x_1} \frac{\partial S}{\partial x_2} \frac{\partial^2 S}{\partial x_1 \partial x_2} \right|, \quad (104)$$

with $g = \det(G) = g_{11}g_{22} - g_{12}^2$. These expressions are a
direct consequence of (93) and (94).

For 2:N mappings, as $x_1$, $x_2$, $n_1$ and $n_2$ are independent and
i.i.d., and we have two orthogonal coordinate curves, then

$$\varepsilon_{\text{un}}^2(x_0) \approx \frac{\sigma_n^2}{2} \sum_{i=1}^2 \left\{ \frac{1}{g_{ii}(x_0)} \left( 1 + \frac{\sigma_n^2}{4} \kappa_i^2(x_0) \right) \right\}, \quad (105)$$

which is a straight forward generalization of (92).

For $M:N$ mappings, the generalization follows directly as we have $M$ (orthogonal) LoC as coordinate curves:

$$\varepsilon_{\text{un}}^2(x_0) \approx \frac{\sigma_n^2}{M} \sum_{i=1}^M \left\{ \frac{1}{g_{ii}(x_0)} \left( 1 + \frac{\sigma_n^2}{4} \kappa_i^2(x_0) \right) \right\}, \quad (106)$$

Again with $g = \det(G) = g_{11}g_{22} - g_{12}^2$. For $M:2$ mappings,
since $z_1$, $z_2$, $n_1$ and $n_2$ are independent and we have two
orthogonal coordinate curves we get

$$\varepsilon_{\text{ch}}^2(z_0) \approx \frac{\sigma_n^2}{M} \sum_{i=1}^2 \left( g_{ii}(z_0) + \frac{3\sigma_n^2}{4} \kappa_i^2(z_0) \right), \quad (113)$$

For $M:N$ mappings, the generalization follows directly

$$\varepsilon_{\text{ch}}^2(z_0) \approx \frac{\sigma_n^2}{M} \sum_{i=1}^N \left( g_{ii}(z_0) + \frac{3\sigma_n^2}{4} \kappa_i^2(z_0) \right), \quad (114)$$

with $g_{ii}$ and $b_{ii}$ the components of FFF and SFF respectively
(as previously defined).
3) Extended Canal Surfaces and maximal curvature: With the definition of principal curvatures it is straight forward to extend the concept of a non-intersecting canal surfaces to hyper surfaces, that is canal hyper surfaces. We provide conditions for $N > M$. The $M < N$ case is similar.

Proposition 2: Consider $M:N$, $N > M$ S-K mappings. For $\rho_{\text{min}} = 1/\kappa_{\text{max}}$, with $\kappa_{\text{max}}$ the maximal principal curvature of $S$, and $r$, the radius of the hyper-sphere $S^{N-M-1}$, the corresponding canal hypersurface (that is, the envelope of $S^{N-M-1}$), will not intersect itself at any point. That is, the canal surface will have no characteristic points $\iff \rho_{\text{min}} > r$ for all points of $S$.

Proof 2: This result follows directly from the proof of Proposition 1 by letting a curve $C$ be the LoC with maximal principal curvature for all points of $S$. That is $C$ is always in the direction of the maximal curvature on $C$. □

4) Choice of coordinates: The problem with LoC coordinates is that one source is always getting a less exact ML-estimate than the other(s) due to the maximal curvature it has at every point. This poses no significant problem (especially at high SNR) as long as a surface with relatively low curvature is chosen. The alternative is to look at general normal sections with approximately the same curvature (on average) for each source over $S$. Generally, we have from [46]

\[ \kappa_n = \frac{b_{\alpha\beta} du^\alpha du^\beta}{g_{\alpha\beta} du^\alpha du^\beta}, \]  

with $b_{\alpha\beta}$ and $g_{\alpha\beta}$ as defined before in Section II-B, $\kappa_1$ and $\kappa_2$ are then found as the solution to the differential equation

\[ \kappa_n^2 - b_{\alpha\beta} g^{\alpha\beta} \kappa_n + b/g = 0, \]  

or by the differential equations (50) or (51) in Section II-B.

Then, from the Theorem of Euler (Theorem 5 in Section II-B),

\[ \kappa_n = \kappa_1 \cos^2 \alpha + \kappa_2 \sin^2 \alpha. \]  

An important special case of coordinates is Geodesic Coordinates (GeCo). A geodesic is the curve with shortest possible length between any two points on a surface $S$. They are therefore convenient as coordinate curves for dimension reducing S-K mappings as they lead to the smallest possible $g_{iii}$ among all curves on $S$ and therefore minimize weak channel distortion, $\varepsilon_{ch}^2$, as defined in (120) below.

The geodesic is found by the minimization of an energy functional, which is a solution to the Euler-Lagrange equations (ELE) [4] p. 161]. With coordinate curves denoted $u^\gamma$, from [4] p. 162], assuming arc length parametrization

\[ S(u^1(\ell), u^2(\ell)), \quad \ell = \ell(x_1), \]  

a Geodesic curve on $S$ is given by the differential equation

\[ \ddot{u}^\gamma + \Gamma_{\mu\sigma}^\gamma \dot{u}^\mu \dot{u}^\sigma = 0, \]  

where $\Gamma_{\mu\sigma}^\gamma$ is Christoffel symbol of second kind (CS-2) as defined in (63) in Section II-B3. Further, one can show that (119) is equivalent to [4] pp. 160-162

\[ \kappa_g = |\dot{S} \ddot{S} \dot{n}| = 0, \]  

where $n = S_1 \times S_2$ is the normal vector to $S$ at $P$. $\kappa_g$ is the geodesic curvature of the relevant curve, which corresponds to the curvature of the projection of the relevant curve on $S$ onto the tangent plane of $S$ at $P$ (see [4] pp. 154-155 for further explanation). That is, another equivalent definition of a geodesic is a curve with zero geodesic curvature. It is possible to show that this equation can be formulated solely w.r.t. the metric tensor of the surface as the following theorem states [4] p. 156):

Theorem 7: The geodesic curvature, $\kappa_g$, of a curve $C$ on a surface $S$ depends on the FFF of $S$ only (as well as the curve itself).

Proof: See [4] p.156]. □

Considering LoC coordinates, we have the following two important Theorems:

Theorem 8: The coordinate curves $u^1 = \text{constant}$ and $u^2 = \text{constant}$ on a portion of a surface $S : S(u^1, u^2) \subset C'$, $r \geq 2$ are geodesics $\iff \Gamma_{12}^1 = 0$ and $\Gamma_{11}^2 = 0$, respectively.

Proof: See [4] pp.156-158]. □

Theorem 9: If a geodesic, $C_G$, on $S$ is a LoC on $S$, then $C_G$ is a plane curve. Further, any plane geodesic on $S$ is a LoC on $S$.

Proof: See [4] p.158-159]. □

Ideally, with geodesic coordinates and arc length parametrization, one could obtain a diagonal metric tensor with independent (and constant) $g_{ii}$, that is

\[ G = \begin{bmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{bmatrix} = \begin{bmatrix} c_1 & 0 \\ 0 & c_2 \end{bmatrix}, \]  

with $c_1, c_2$ positive constants. However, this is not necessarily possible to obtain over the whole of $S$, only over portions of it: Two common ways of obtaining GeCo is through geodesic parallel coordinates (GPC) or geodesic polar coordinates (GPoC) [4] pp. 162-168].

As we shall see in Section IV there are subsets of surfaces named developable surfaces that can be mapped isometrically to the Euclidean plane, as well as subsets of surfaces that can be mapped conformally to the Euclidean plane, where (122) is obtained all over $S$.

C. Definition of weak noise regime for S-K mappings

We are now able to make a more rigourous definition of the weak noise regime connected to the distortion analysis of S-K mapping:

1) Weak noise distortion: We begin with the definition of weak noise distortion for dimension expanding S-K mappings: Let $S_{lin}(x)$ denote the linear approximation of $S(x)$ at $x_0$

\[ S_{lin}(x) = S(x_0) + J(x_0)(x - x_0), \]
where \( J(x_0) \) is the Jacobian matrix (see Appendix B) of \( S \) evaluated at \( x_0 \). We have the following proposition providing the exact distortion under linear approximation:

**Proposition 3: Minimum weak noise distortion**

For any continuous i.i.d. source \( x \in \mathbb{R}^M \) with unimodal pdf \( f_x(x) \) communicated on an i.i.d. Gaussian channel of dimension \( N \) using a continuous expanding S-K mapping \( S \) where \( S_i \in C^r(\mathbb{R}^M), \ r \geq 1, \ i = 1, \ldots, M, \) the minimum distortion under the linear approximation in (125) is given by

\[
\varepsilon_{chw}^2 = \frac{\sigma^2}{M} \int_0^1 \cdots \int_0^1 \sum_{i=1}^N g_{ii}(z) f_z(z) \mathrm{d}z, \tag{126}
\]

where \( f_z(z) \) is the channel signal pdf, and \( g_{ii}(z) \) are the diagonal components of the metric tensor \( G \) of \( S \).

**Proof:** See [5]. □

The name weak noise distortion is due to Definition 9 given below.

Consider now the weak noise error given by the 2nd order approximation to \( S \) in (106).

**Definition 8: Weak noise regime (dimension expansion)**

Let \( x_0 \) denote the transmitted vector and \( S(x_0) \) its representation in the channel space. We say that we are in the weak noise regime whenever the second order term (as well as higher order terms) in (106), i.e., the terms containing \( \kappa_i \), are negligible compared to the 1st order term. That is, (123) is a close approximation to \( S \) and the weak noise distortion in (124) provides an accurate approximation to the actual distortion in the absence of anomalies. □

From (106) it is then clear why a linear system is convenient at low SNR, as \( \kappa_i = 0 \) and the higher order terms vanishes. For higher SNR one would seek nonlinear mappings with low curvature \( \forall x \). And therefore the weak noise regime is a good approximation for any reasonably chosen mappings at high SNR.

2) Weak channel distortion: The received vector \( \hat{z} = z + n \) must be passed through \( S \) to reconstruct \( x \). When the noise is weak enough, distortion analysis can be done by considering the tangent space of \( S \). That is, one can consider the linear approximation \( S_{lin}(z_0) \) of \( S(z) \) at \( z_0 \)

\[
S_{lin}(z_0 + n) = S(z_0) + J(z_0)n, \tag{125}
\]

We have the following proposition providing the exact distortion under linear approximation:

**Proposition 4: Minimum Weak Channel Distortion**

For any continuous i.i.d. Gaussian channel of dimension \( N \) and any dimension reducing S-K mapping \( S \) where \( S_i \in C^r(\mathbb{R}^M), \ r \geq 1, \ i = 1, \ldots, M, \) the distortion under the linear approximation in (125) is given by

\[
\varepsilon_{chw}^2 = \frac{\sigma^2}{M} \int_0^1 \cdots \int_0^1 \sum_{i=1}^N g_{ii}(z) f_z(z) \mathrm{d}z, \tag{126}
\]

where \( f_z(z) \) is the channel signal pdf, and \( g_{ii}(z) \) are the diagonal components of the metric tensor \( G \) of \( S \).

**Proof:** See [5]. □

The name weak channel distortion is due to Def. 9 given below.

Consider now the channel error given by the 2nd order approximation to \( S \) in (114).

**Definition 9: Weak noise regime (dimension reduction)**

Let \( z_0 \) denote the transmitted vector and \( S(z_0) \) its representation in the source space. We say that we are in the weak noise regime whenever the second order term (and higher order terms) in (114), i.e., the terms containing \( \kappa_i \), are negligible compared to the 1st order term. That is, (125) is a close approximation to \( S \) and the weak channel distortion in (126) provides an accurate approximation to the actual distortion due to channel noise. □

IV. NEW RESULTS ON MAPPING CONSTRUCTION USING SURFACES

There are a myriad of possible surfaces, as exemplified in the Encyclopedia of Analytical Surfaces. Many of these will correspond with Definition 7. Trying out many of these as potential S-K mappings is not a constructive approach. However, the criteria laid out above as well as in this section will rule out most surfaces as good candidates for well-performing S-K mappings. Having many specific criteria that has to be satisfied, constraining the set of relevant surfaces, can bring an advantage to numerical solutions to variational calculus approaches, which tend to be stuck in local minima if not already well performing mappings are provided as input. In fact, with these criteria “baked” into the solution process of the variational calculus problem, one can better constrain the solution, and this will likely produce mappings closer to the global optimum.

We provide results here for surfaces in \( \mathbb{R}^3 \), some of which can easily be extended to higher dimensional spaces. Our earlier investigations showed that a constant metric tensor with independent components like in (123) would be advantageous, as it would lead to a shape preserving mapping. However, to incorporate a general pdf, it may be better to let

\[
G(x_1, x_2) = \begin{bmatrix} g_{11}(x_1) & 0 \\ 0 & g_{22}(x_2) \end{bmatrix}, \tag{127}
\]

where \( g_{ii}(x_i) \) could be "adjusted" to the sources pdf. Both of these should ideally lead to the correct scaling with channel SNR. That is, the performance of the relevant mapping has the same slope as the corresponding OPTA curve as the SNR increases when its free variables are chosen optimally. However, things are not that straightforward.
For more general surfaces we need further results from [4]. We first consider isometric surfaces, that is, surfaces that can be mapped onto each other in a length preserving manner.

**Theorem 10:** An allowable mapping of a PoS S onto a PoS S* is isometric ⇐⇒ $g_{\alpha\beta} = g_{*\alpha*\beta}$ at corresponding points when referred to the same coordinate system on S and S*.

**Proof:** See [4, pp.176-177]. □

**Theorem 11:** Isometric surfaces have the same Gaussian curvature at corresponding points. Corresponding points on those surfaces have the same geodesic curvature, $\kappa_g$, at corresponding points.

**Proof:** See [4, p.177]. □

A subset of surfaces named developable surfaces (DS) can be mapped isometrically to the Euclidean plane [4, p.189]. This implies a metric like (122) for any such surface. The DS is a special case of a ruled surface (RS). A RS is obtained by a set of straight lines, named generators $z(\ell,t)$, which are related through a space curve $y(\ell)$, named indicatrix, as follows

$$S(\ell,t) = y(\ell) + t\mathbf{z}(\ell), \quad (128)$$

where $\mathbf{z}$ is a unit vector linearly independent of the tangent $\dot{y}$, that is $\dot{y} \times \mathbf{z} \neq 0$. The indicatrix acts like the trajectory for a straight line through space, $\mathbf{z}$ and $\dot{y}$ also act as coordinate curves. An example on a RS is the Helicoid (or Archimedes Screw) which is depicted in Fig. 7 below. Here one can see a straight line $\mathbf{z}$ moved along a helix $y$.

If an RS should also be a DS, the tangent planes need to be the same at all points along the same generator $\ell$, the curvature $\kappa_\ell$ is zero. Also, when $K = 0$ we always have a DS. This implies that only DS can be mapped onto a plane without distorting/changing distances. A DS will therefore guarantee a constant metric tensor $G$.

We have three examples of RS: Let $t = \dot{y}$ (tangent vector), $\mathbf{p} = \dot{y}$ (principal normal vector) and $\mathbf{b} = \mathbf{t} \times \mathbf{p}$ (binormal vector):

i) **Tangent Surface (TS):**

$$S(\ell,t) = y(\ell) + t\mathbf{t}(\ell), \quad (130)$$

ii) **Principal Normal Surface (PNS):**

$$S(\ell,t) = y(\ell) + t\mathbf{p}(\ell), \quad (131)$$

iii) **Binormal Surface (BNS):**

$$S(\ell,t) = y(\ell) + t\mathbf{b}(\ell), \quad (132)$$

Under certain conditions these surfaces are also DS [4, p.182].

**Theorem 15:** The TS is always a DS. PNS and BNS are DS ⇔ $y(\ell)$ is a plane curve, i.e., the torsion $\tau = 0$.

**Proof:** See [4, p.182]. □

**Proof:** See [4, p.185]. □

From the above results it seems like DS is the correct choice of surface for 3:2 and 2:3 mappings. However, as we show in the following, this is not true in general.

A DS is one of many mappings that can be decomposed into sub-mappings. Consider the 3:2 system in Fig. 5 consisting of a 1:1 and 2:1 system. For a DS, in particular, this sub-system consists of a linear 1:1 mapping ($z$) and a nonlinear 2:1 mapping ($y$). However, many decomposable systems may also have a nonlinear 1:1 mapping (or general nonlinear subsystems of $M:N$ mappings). In the following we will show that any mapping that can be split into sub-mappings will not obtain the

A PNS in ii) was applied in [14] for a bivariate Gaussian to utilize correlation among two variables. This provided a significant gain over linear and piecewise linear approaches.

---

\[\text{Fig. 5. Splitting of a 3:2 mapping into 1:1 and 2:1 sub-mappings.}\]

---

\[\text{\footnotemark[4]A PNS in ii) was applied in [14] for a bivariate Gaussian to utilize correlation among two variables. This provided a significant gain over linear and piecewise linear approaches.}\]
optimal slope in SDR as SNR increases. That is, they diverge away from OPTA at high SNR.

Consider the general form of OPTA for \( M:N \) mapping:

\[
\text{SDR}_{M:N} = \frac{\sigma_x^2}{D_{M:N}} = \left(1 + \frac{P_{M:N}}{\sigma_n^2}\right)^{-\frac{1}{m}}
\]

(133)

We have that the minimal distortion for the 1:1 system is (solve (133) w.r.t. \( D_t \) for \( N/M = 1 \))

\[
D_{1:1} = \frac{\sigma_n^2}{1 + \frac{P_t}{\sigma_n^2}}
\]

(134)

and for the 2:1 system is (solve (133) w.r.t. \( D_t \) for \( N/M = 1/2 \))

\[
D_{2:1} = \left(\frac{\sigma_n^2}{1 + \frac{P_t}{\sigma_n^2}}\right)^{\frac{1}{2}}
\]

(135)

With \( P_t \) the total power of the 3:2 system one can allocate power to the two sub-systems through a factor \( \kappa \in [0, 1] \). Let SNR= \( P_t/\sigma_n^2 \) and use the fact that \( 1 + x \approx x \) as \( x \) becomes large, then the total distortion in the limit of large SNR becomes,

\[
\lim_{\text{SNR} \to \infty} D_{t(3:2)} = \frac{\sigma_n^2}{3} \left[ \left(1 + \kappa \text{SNR}\right)^{-\frac{1}{m}} + \left(1 + (1 - \kappa)\text{SNR}\right)^{\frac{1}{m}} \right]
\]

\[
\frac{\sigma_n^2}{3} \lim_{\text{SNR} \to \infty} \left[ \frac{1}{\kappa \text{SNR}} + \frac{1}{\sqrt{(1 - \kappa)\text{SNR}}} \right].
\]

(136)

According to the laws of limits \( \lim_{x \to \infty} \kappa x = \kappa \lim_{x \to \infty} x \), and so the power allocation factor can be moved outside the limit. Then, since \( \sqrt{\text{SNR}} \) grows more slowly than SNR, \( D_{t(3:2)} \) will be dominated by \( 1/\sqrt{\text{SNR}} \) as \( \text{SNR} \to \infty \). This implies that a 3:2 system realized by a DS will have the slope of 2:1 system at high SNR. The generalization to a general dimension reducing mapping \( (M > N) \) follows along the same lines.

As DS can be seen as a straight line (1:1 system) moved along a curve \( y \) (a \( M:1 \) or \( 1:N \) system) a DS has the wrong slope and will therefore diverge from the OPTA bound as SNR grows large (this will include the suggestion for higher dimensional mappings in [15]).

We show that this is also the case for expanding mappings: Consider the 2:3 system in Fig. 6 consisting of a 1:1 and 1:2 system. The minimal distortion for the 1:1 system is as in [139]. For the 1:2 system we have

\[
D_{1:2} = \left(\frac{\sigma_n^2}{1 + \frac{P_t}{\sigma_n^2}}\right)^{\frac{1}{2}}
\]

(137)

With \( P_t \) the total power of the 2:3 system one can allocate to the two sub-systems according to a allocation factor \( \kappa \in [0, 1] \subseteq \mathbb{R} \). The total distortion in the limit of large SNR becomes

\[
\lim_{\text{SNR} \to \infty} D_{t(m+n:2)} = \frac{\sigma_n^2}{m+n} \left[ \left(1 + \kappa \text{SNR}\right)^{\frac{1}{m}} + \left(1 + (1 - \kappa)\text{SNR}\right)^{\frac{1}{m}} \right]
\]

\[
= \frac{\sigma_n^2}{m+n} \lim_{\text{SNR} \to \infty} \left[ \frac{1}{\kappa^{1/n} \text{SNR}^{1/n}} + \frac{1}{(1 - \kappa)^{1/m} \text{SNR}^{1/m}} \right].
\]

(140)
According to the laws of limits \( \lim_{x \to \infty} \frac{\kappa x}{x} = \kappa \lim_{x \to \infty} x \), and so the power allocation factor(s) can be moved outside the limit. Then, for \( m > n \) since \( \text{SNR}^{1/m} \) grows more slowly than \( \text{SNR}^{1/n} \), \( D_{l(m+n;2)} \) will be dominated by the second term in (140) as \( \text{SNR} \to \infty \).

In the expansion case, a similar derivation leads to

\[
\lim_{\text{SNR} \to \infty} D_{l(2;m+n)} = \frac{\sigma_x^2}{m+n} \lim_{\text{SNR} \to \infty} \left[ \frac{1}{\kappa \text{SNR}^m} + \frac{1}{(1-\kappa) \text{SNR}^n} \right] \quad (141)
\]

If \( m > n \), the first term in (141) dominates as \( \text{SNR} \to \infty \). □

Despite these results, DS provide a simple mapping which is based on combinations of lower dimensional mappings which may provide quite good performance in the small to medium SNR range as illustrated in Sections IV-A as well as in [15].

To avoid the problem of wrong slope, one has to widen the set of relevant mappings, but still keep a similar type metric tensor. A more general and larger set of surfaces having a uniform dimension reducing mapping exists a portion of a minimal surface. Another subset of surfaces that is of importance is the helicoid and \( \frac{\alpha}{\beta} \) is conformal to the plane [4, pp. 193]:

**Theorem 16**: An allowable mapping of a PoS of \( \mathcal{S} \) onto a PoS of \( \mathcal{S}^* \) is conformal \( \Leftrightarrow \) when on \( \mathcal{S} \) and \( \mathcal{S}^* \) the same coordinate systems have been introduced, the coefficients of the FFF \( g_{\alpha\beta} \) and \( g'_{\alpha\beta} \) of \( \mathcal{S} \) and \( \mathcal{S}^* \), respectively, are proportional:

\[
g'_{\alpha\beta} = \gamma(u^1, u^2)g_{\alpha\beta}, \quad \gamma > 0, \quad \alpha\beta = 1,2. \quad (142)
\]

**Proof**: See [4] p.193-194]. □

**Theorem 17**: Any simply-connected portion \( \mathcal{S} \) of a surface which has a representation of class \( r \geq 3 \) can be conformally mapped into a plane.

**Proof**: See [4] p.196-198]. □

Another subset of surfaces that is of importance is minimal surfaces. A minimal surface is a surface \( \mathcal{S} \) of class \( r \geq 2 \) whose mean curvature

\[
H = \frac{1}{2} k^\alpha = \frac{1}{2} h_{\alpha\beta} g^{\alpha\beta} = 0 \quad (143)
\]

at every point of \( \mathcal{S} \). We have the following theorem [4] p. 244]

**Theorem 18**: Let \( \mathcal{C} \) be a simple closed curve. If among all portions of surfaces of class \( r \geq 2 \) bounded by \( \mathcal{C} \) there exists a portion \( \mathbf{x}(u^1, u^2) \) of minimum area then \( \mathbf{x}(u^1, u^2) \) is necessarily a portion of a minimal surface.

**Proof**: See [4] p.244-245]. □

That is, minimal surfaces are natural solution candidates for the Euler-Lagrange equations for surfaces. However, one cannot conclude that a surface needs to be minimal in order for S-K mappings to be optimal as we will see in the following sections.

Finally, we state a condition that has to be satisfied if a uniform \( M:N \) dimension reducing mapping \( ^5\) (\( M > N \)) is to follow the slope of OPTA: From the proof of Proposition 6, i.e., Eq. (58) in [5], we have that

\[
D_l = \varepsilon_a^2 + \varepsilon_{ch}^2 = \frac{M-N}{4M(M-N+2)} \Delta^2 + M \frac{\Delta^2}{2} \bar{B}^2 \quad (144)
\]

Then, for a 3:2 mapping, we have that

\[
\varepsilon_{ch}^2 \sim \frac{1}{\Delta}, \quad \varepsilon_a^2 \sim \Delta^2, \quad (145)
\]

in order to obtain \( \text{SDR} \sim \text{SNR}^{2/3} \).

**A. Examples on 3:2 mappings**

We analyze and simulate several 3:2 mappings selected based on the characteristics of surfaces presented in the previous sections. For comparison we consider OPTA, Block Pulse Amplitude Modulation (BPAM) [16] and Power Constrained Channel Optimized Vector Quantizer (PCCOVQ) [17], [18]. BPAM is the optimal linear mapping (i.e., a plane in the source- or channel space) for any \( r = N/M \in \mathbb{Q} \). The PCCOVQ is a numerically optimized discrete mapping (for any \( r = N/M \in \mathbb{Q} \)) that often replicate continuous or piece-wise continuous curves or surfaces when the number of points in its constellations is large. Obviously, one would like any nonlinear mapping to rise well above BPAM as the SNR increases for any \( r \neq 1 \). The PCCOVQ is often hard to beat with many points in the constellation since one is more free to move points in space than what is the case when one is stuck to a given curve or surface. Approaching or beating a PCCOVQ system with many points is therefore a good indication of a well performing mapping.

1) Helicoid: The first structure chosen for \( \mathcal{S} \) is the Helicoid (also known as Archimedes’ screw) depicted in Fig. 7. The

![Fig. 7. 3:2 Helicoid mapping shown in the source space. The structure corresponds to equation (146) where \( R = \Delta = 1, \alpha_1 = \alpha_2 = 1 \), and \( z_1, z_2 \in [-\pi, \pi] \).](image-url)
Helicoid has a parametric equation \( S(z_1, z_2) \) with components
\[
S_1(z_1, z_2) = \frac{R\alpha_1}{\pi} z_1 \cos(\alpha_2 z_2),
S_2(z_1, z_2) = \frac{R\alpha_1}{\pi} z_1 \sin(\alpha_2 z_2),
S_3(z_2) = \frac{\Delta\alpha_2}{\pi} z_2,
\]
where \( \Delta \) corresponds to the distance between two helical discs of the surface (at some distance from the \( x_3 \) axis). \( R \) is determining the size of the “radius” of the surface in the \( x_1 x_2 \)-plane (see Fig. 7 from the "top"). To obtain a symmetric channel signal both positive and negative parameter values are used and \((z_1, z_2) = (0, 0) \) corresponds to the origin in \( \mathbb{R}^3 \). The source vectors \( x \) are projected onto the Helicoid. The point \((y_1, y_2)\), corresponding to the point on the Helicoid, represents the two channel signals. The channel signals are further scaled with a factor \( \alpha_i \) to satisfy a given power constraint, that is \( z_i = y_i/\alpha_i, i = 1, 2 \).

Channel power: The channel power can be found by considering equation (146) and Fig. 7. The derivation is done leaving out the scaling factors \( \alpha_1 \) and \( \alpha_2 \), so the statistics of the channel signals \( y_1 \) and \( y_2 \) are considered.

Looking at the Helicoid from above (\( z_3 \) direction), one can see a disc in the \( x_1 x_2 \) plane. The radius of this disc is given by \( r = (R/\pi)y_1 \) (i.e., directrix lines, following a helical indicatrix), while \( y_2 \) traces out all angles (the “parallel helices”, i.e., the indicatrix). Thus \( y_1 \) is given by the function
\[
y_1 = \pm \frac{\pi}{R} \sqrt{y_2^2 + x_2^2}.
\]

According to [6, p. 190], the mapping of two Gaussian random variables by this function gives a Rayleigh distribution. But due to the \( \pm \) sign a zero mean “double Rayleigh” distribution is obtained
\[
f_{y_1}(y_1) = \frac{1}{2} \frac{R^{2}|y_1|}{\pi^2 \sigma_x^2} e^{-\frac{R^2 y_2^2}{\pi^2 \sigma_x^2}}.
\]

To get the constants right, one has to use the fact that the random variable \( h(y) = c_2 \) has distribution \( f_h(h) = (1/|c|) f_y(h/c) \) [6, p. 131]. Further in [6, p 148] it is shown that the second moment of the Rayleigh distribution \((y/\nu^2)e^{-y^2/(2\nu^2)} = E(y^2) = 2\nu^2\), and so the variance on channel one is given by
\[
\sigma^2_{y_1} = 2\left(\frac{\sigma_x \pi}{R}\right)^2
\]

The other parameter, \( y_2 \), will trace out parallel helices with “radius” given by (147). Consider first \( y_2 = 0 \). Then \( y_2 \) will lie on the \( x_3 \) axis and have a Gaussian distribution since then \( x_3 = (\Delta/\pi)y_2 \), and so the variance of \( y_2 \) will be
\[
\sigma^2_{y_2|y_1=0} = \left(\frac{\pi \sigma_x}{\Delta}\right)^2.
\]

Considering other cases \((y_1 \neq 0)\) it seems like the variance (and the pdf) stays more or less the same, independent of the value of \( y_1 \). Although \( y_2 \) will be on a helix, it seems to trace out all of \( z_3 \) in a very similar way. The discrepancy from (150) seems to be more or less constant, equal to 1/2, except when \( \Delta \) is very close to zero, so that
\[
\sigma^2_{y_2} = \left(\frac{\pi \sigma_x}{\Delta}\right)^2 + \frac{1}{2}.
\]

Taking the scaling, \( \alpha_1 \) and \( \alpha_2 \), into account, the total channel power (per channel dimension) becomes
\[
P = \frac{1}{2} \left(\frac{\sigma^2_{y_1}}{\alpha_1^2} + \frac{\sigma^2_{y_2}}{\alpha_2^2}\right) = \frac{1}{2} \left(\sigma^2_{y_1} + \sigma^2_{y_2}\right).
\]

Channel distortion: To determine the channel distortion, the diagonal components of the metric tensor of \( S \) is needed
\[
g_{11} = \left\| \frac{\partial S}{\partial z_1} \right\|^2 = \left\| \frac{R\alpha_1}{\pi} \cos(\alpha_2 z_2), \frac{R\alpha_1}{\pi} \sin(\alpha_2 z_2), 0 \right\|^2
= \left(\frac{R\alpha_1}{\pi}\right)^2,
\]
and
\[
g_{22} = \left\| \frac{\partial S}{\partial z_2} \right\|^2 = \left\| -\frac{R\alpha_1}{\pi} z_1 \sin(\alpha_2 z_2) / \frac{R\alpha_1}{\pi} z_1 \cos(\alpha_2 z_2) / \frac{\Delta\alpha_2}{\pi} \right\|^2
= \left(\frac{R\alpha_1\alpha_2}{\pi}\right)^2 \frac{\Delta\alpha_2}{\pi}^2 + \left(\frac{\Delta\alpha_2}{\pi}\right)^2.
\]

It is straightforward, using the derivatives in (153) and (154), to show that \( g_{12} = 0 \). One can now calculate the weak channel distortion as
\[
\varepsilon^2_{ch} = \frac{\sigma^2}{3\pi^2} \int \left( \sum_{i=1}^{2} g_{ii} f_i(z) \right) dz
= \frac{\sigma^2}{3} \int \int \int \int \int \left( \left(\frac{R\alpha_1}{\pi}\right)^2 + \left(\frac{R\alpha_1\alpha_2}{\pi}\right)^2 \right) \frac{\Delta\alpha_2}{\pi}^2 f_{z_1, z_2}(z_1, z_2) dz_1 dz_2.
\]
This leads to
\[
\varepsilon^2_{ch} = \frac{\sigma^2}{3\pi^2} \left(\left(\frac{R\alpha_2}{\pi}\right)^2 + \left(R\alpha_1\right)^2(1 + \alpha_2^2 \sigma_x^2)\right).
\]

From (156) one can observe that the channel distortion increases linearly with the power on channel one (\( \sigma_x^2 \)). This can also be seen from the metric component \( g_{22} \). I.e., the length of the velocity vector along the \( z_2 \) direction is dependent on \( z_1 \). Trying to compensate for the dependence in one direction, i.e., using arc-length along that coordinate curve, will affect the other coordinate curve and vice versa, as will be demonstrated at the end of this section.

Approximation distortion: The approximation distortion is not feasible to find analytically, particularly at low SNR. The distance between two folds of \( S \) is determined by \( \Delta \), which determines the size of the approximation distortion. The approximation distortion can then be approximated by nonlinear curvefitting: a model that coincide well when \( \sigma_x = 1 \)
\[
\varepsilon^2_a \approx \beta \Delta^3 + \gamma \Delta^2 + \delta \Delta
= -0.0036 \Delta^3 + 0.024 \Delta^2 + 0.0056 \Delta, \quad \Delta \in [0, 3], \quad \sigma_x = 1.
\]

Optimization of the Helicoid system: Assume \( \sigma_x = 1 \). To find the optimal performance, one need to optimize over
The performance of the 3R PCCOVQ and BP AM. Lagrangian of this problem is 

\[ L(R, \Delta, \alpha_1, \alpha_2) = \epsilon_a^2(\Delta \alpha_1) \epsilon_{ch}(R, \Delta, \alpha_1, \alpha_2) - \lambda c_4(R, \Delta, \alpha_1, \alpha_2), \]  

(158)

where the constraint is given by

\[ c_4(R, \Delta, \alpha_1, \alpha_2) = P_{max} - P(R, \Delta, \alpha_1, \alpha_2) \geq 0. \]  

(159)

\[ f(R, \Delta, \alpha_1, \alpha_2) \]  

is the maximum allowed power per channel and \( P(R, \Delta, \alpha_1, \alpha_2) \) is given by (152). The first order necessary conditions for a minimum is given by the the Karush-Kuhn-Tucker (KKT) equations [10, p.328]. One of the criterion that has to be satisfied to be in a KKT point is

\[ \nabla_{R\Delta \alpha_1 \alpha_2} L(R, \Delta, \alpha_1, \alpha_2, \lambda) = 0 \]  

(160)

Solving (160) analytically is impossible since it contains an equation of high order with no algebraic solution. Therefore numerical optimization has to be used.

Fig. 8 shows the comparison between the theoretical model given above and its simulation, as well as OPTA, PCCOVQ and BPAM. One can observe that the performance of the theoretical model and the simulation coincide quite well at high SNR (as expected, since higher order terms related to curvature becomes negligible). The PCCOVQ system clearly outperforms the Helicoid.

The Helicoid is a Ruled Surface as well as a minimal surface, but is not a developable surface since it is a PNS with a helix indicatrix \( \gamma(\ell) \) (which is not a plane indicatrix). As such, its a potentially good 3:2 mapping, as seen at low to medium SNR. However, due to the fact that the Helicoid is a ruled surface it is not isometric to the plane, and \( g_{22} \) is both dependent and increases with \( z_2 \) as seen from (154). Therefore the Helicoid diverges away from OPTA as SNR gets high. Can one compensate for this effect using arc length parametrization?

**Arc length parametrization:** Assume now that \( z_1 \) is mapped through a function \( \varphi(z_1) \) before mapping onto the Helicoid, leading to arc length parametrization along the \( z_2 \) coordinate, effectively removing the dependence of \( g_{22} \) w.r.t. \( z_1 \). Then \( g_{22} \) becomes (with \( \alpha_1 = \alpha_2 = 1 \))

\[ g_{22} = \frac{\Delta^2 + R^2 z_1^2}{\pi^2 \varphi^2(z_1)}. \]  

(161)

Choosing \( \varphi(z_1) = \sqrt{\Delta^2 + R^2 z_1^2}/\pi \) then \( g_{22} = 1 \). But now \( g_{11} \) will become (with \( R = \Delta = 1 \))

\[ g_{11} = \frac{z_1^4 + (z_1 z_2)^2 + 2z_2^2 + 1}{(1 + z_1^2)^2/\pi^2}. \]  

(162)

Fig. 9 depict (162). This is even worse than \( g_{22} \sim z_1^2 \), since now, if the amplitudes on channel two is relatively large, the noise on channel one will be scaled the most for the most probable channel symbols on channel one, giving a large average distortion after decoding.

This implies that the Helicoid is not a good choice of mapping. However, this interdependence is a problem for many surfaces. One exception is DS.

2) **Right Cylinder with Archimedes Spiral Directrix (RCASD):** The parametric equation for RCASD is [11, p.51]

\[ \mathbf{S}(x) = [a \varphi(z_1) \cos(\varphi(z_1)), a \varphi(z_1) \sin(\varphi(z_1)), \alpha_2 z_2], \]  

(163)

with \( \alpha_2 \) some amplification factor and \( a = \Delta/\pi \). \( \Delta \) is the distance between the spiral arms. Fig. 10 depicts a surface plot of the RCASD with the (LoC) coordinate grid included.

The tensor components of the first two fundamental forms are

\[ g_{11} = \left( \frac{\Delta}{\pi} \right)^2 (\varphi'(z_1))^2 (1 + \varphi^2(z_1)), \]  

(164)

\[ g_{22} = \alpha^2_2, \]  

\[ g_{12} = g_{21} = 0, \]

Fig. 9. The first diagonal component of the metric tensor \( g_{11}(z_1, z_2) \) for the Helicoid after a arc length transformation of \( z_2 \).
with noise added to the channel signal $z$. Then the channel distortion becomes

$$
\varepsilon_{\text{ch}}^2 = \frac{\sigma_n^2}{3} \iint \sum_{i=1}^{2} g_i(z)f_z(z)dz
$$

$$
= \frac{\sigma_n^2}{3} (\alpha_1^2 + \alpha_2^2) f_z(z)dz
$$

$$
= \frac{\sigma_n^2 (\alpha_1^2 + \alpha_2^2)}{3}.
$$

(167)

From this we see that the metric tensor is as in (122), which was one of the criteria we sought.

To compute the approximation distortion one can observe from Fig. 10 that we have an approximation operation similar to archimedes spiral, i.e., a uniform S-K mapping (see 5, Definition 6), but now over 3 sources in $\mathbb{R}^3$ implying that Eq. (18) in 5 will apply as a lower bound with channel dimension $N = 2$ and source dimension $M = 3$

$$
\bar{\varepsilon}_q^2 \geq \frac{M - N}{4M(M - N + 2)} \Delta^2 = \frac{\Delta^2}{36}.
$$

(168)

**Power:** In the $x_1, x_2$-plane we have the same configuration as for the Archimedes spiral applied for 2:1 compression in 11. That is, we map via an arc length parametrization from a two dimensional circular region (a disc) to a one dimensional representation. As shown in 11, for $y_1$ this results in a Laplace distribution with variance

$$
\sigma_{y_1}^2 = 2 \left( \frac{2\pi}{\Delta} \sigma_2^2 \right)^2, \quad \eta = 0.16,
$$

at high SNR. $y_2$ is simply a linear mapping of $x_3$ so that we have a Gaussian distribution with variance $\sigma_{y_2}^2 = \sigma_3^2$.

The channel signals have to be scaled by some factor $\alpha_i$ in order to satisfy the power constraint as the SNR changes. That is, $z_i = \eta \alpha_i$ and so $\sigma_z^2 = \sigma_y^2/\alpha_i^2$. The total channel power then becomes

$$
P_t = \frac{1}{2} \left( \frac{\sigma_1^2}{\alpha_1^2} + \frac{\sigma_2^2}{\alpha_2^2} \right)^2.
$$

(170)

**Optimization:** With constraint $C_t = P_{\text{max}} - P_t(\Delta, \alpha_1, \alpha_2) \geq 0$, we get the following objective function

$$
\mathcal{L}(\Delta, \alpha_1, \alpha_2) = \bar{\varepsilon}_q^2(\Delta) + \bar{\varepsilon}_\text{ch}(\alpha_1, \alpha_2) - \lambda C_t(\Delta, \alpha_1, \alpha_2).
$$

(171)

We use the same procedure to determine the optimal parameters as for the Helicoid in Section IV-A1.

The performance of the optimized RCASD is shown in Fig. 11. The RCASD clearly improves with SNR, beating both the linear BPAM as well as PCCOVQ (see footnote 6) at high SNR. The RCASD is also noise robust. However, the slope at high SNR is lower than that of OPTA. We will show that this slope corresponds to that of a 2:1 system:

**High SNR analysis.** The Karush-Kuhn-Tucker (KKT) conditions 10 for the optimization problem in 186 are

$$
\frac{\partial \mathcal{L}}{\partial \Delta} = 0, \quad \frac{\partial \mathcal{L}}{\partial \alpha_i} = 0, \quad i = 1, 2, \quad C_t \geq 0, \quad \lambda \geq 0.
$$

(172)

From these conditions it is straightforward to show that no non-zero solution exists for inner points, and therefore the only

$$
g_{11} \approx \alpha_1^2, \quad \forall z_1, z_2.
$$
solution of interest lies on the boundary $C_1 = 0$. Solving this equation w.r.t. $\alpha_1^2$ we get
\[\alpha_1^2 = \frac{\alpha_2^2}{\Delta^2(2\alpha_2^2 P_{max} - \sigma_x^2)}, \quad \kappa = 2(2\pi^2 \sigma_x^2)^2. \tag{173}\]

By inserting this in the total distortion we get the following unconstrained objective function
\[\tilde{D}_t = \frac{\Delta^2}{36} + \frac{\sigma_x^2}{3} \left( \frac{\alpha_2^2}{\Delta^2(2\alpha_2^2 P_{max} - \sigma_x^2)} - \alpha_2^2 \right). \tag{174}\]

We have the system of equation $\partial \tilde{D}_t/\partial \Delta = 0, \partial \tilde{D}_t/\partial \alpha_2 = 0$. By solving $\partial \tilde{D}_t/\partial \alpha_2 = 0$ w.r.t. $\alpha_2$ one get the four possible solutions
\[\alpha_2 = \pm \sqrt{\frac{\sigma_x}{2P_{max}} (\sigma_x \pm \sqrt{\kappa/\Delta})}. \tag{175}\]

By inspecting this expression one can see that the first sign must be positive as $\alpha_1 \geq 0$. By some further inspection one can conclude that also the sign inside the square root should be positive. By inserting \[\frac{\Delta^4}{18} - \frac{\kappa \sigma_x^2}{3SNR} \Delta - \frac{\kappa}{3SNR} = 0, \quad SNR = P_{max}/\sigma_n^2. \tag{176}\]
This equation does not have an analytical solution. However, as $\Delta$ gets small (SNR gets large) the first and last term dominates, and so an approximate solution is
\[\Delta \approx \frac{4\sqrt{6\kappa}}{\sqrt{SNR}}. \tag{177}\]

Since at high SNR, $D_t = \bar{e}_g^2 = \bar{e}_ch^2$, we see that SDR~SNR$^{-1/2}$ at high SNR, corresponding to the slope of a 2:1 system. This is also exactly what one can expect from Prop. \[\text{[5]} \] and Eq. \[\text{[156]}, \] since the RCASD is a developable surface which always can be split into a 1:1 and a 2:1 system (here, a line and a spiral).

Curvature Analysis. As we consider LoC, the principal curvatures are easily found from the fundamental forms in \[\text{[164]} \] and \[\text{[165]}. \]
\[\kappa_1 = \frac{b_{11}}{g_{11}} = -\frac{2 + \varphi^2(z_1)}{a(\sqrt{1 + \varphi^2(z_1)})^{3/2}}, \quad \kappa_2 = 0. \tag{178}\]

With $\varphi(z_1)$ as in \[\text{[166]} \] one can evaluate how curvature is affected by the free parameters $\Delta$ and $\alpha_1$. The maximal curvature as function of both $\alpha_1$ and $\Delta$ is shown in Fig. \[\text{[18(a)]}. \]

As a particular example, we consider 30dB SNR. By inserting optimized parameters for 30dB SNR found by the optimization procedure \[\text{below (}$\Delta^* = 0.068$, $\alpha_1^* = 3.33$\[\text{)} one obtains $|\kappa_1| < 1$ averaged over the relevant range of $z_1$. By considering the distortion terms in \[\text{[106]}, \] one can see that the 1st order term is in the order of about 0.001/0.001$^2 = 1000$ over the 2nd order term. Therefore, RCASD is a mapping following Definition \[\text{[8]} \] at high SNR.

### 3) Menge Surface with Cylindrical Directrix Surface (MS-CDS):
This surface has many configuration with the RCASD as a special case. The surface is described in \[\text{[11, pp.189-197], \} \] and its parametric equation has components
\[S_1(z_1, z_2) = \frac{\Delta}{\pi} \cos(\varphi(z_1)) - \cdots \]
\[S_2(z_1, z_2) = \frac{\Delta}{\pi} \sin(\varphi(z_1)) - \cdots \]
\[S_3(z_1, z_2) = B\alpha_2 z_2 \sin(\theta_0) \tag{179}\]
Note that we display the equation for $\theta_0 = \pm \pi/2$ which will leave out some terms multiplied with $\cos(\theta_0)$. The reason is that any other angle of $\theta_0$ will lead to a non-symmetric surface across the $x_1, x_2$ plane, and therefore a sub-optimal structure w.r.t. approximation operation. We also map $z_1$ through a function $\varphi(z_1)$ for the same reason as for the RCASD (enable arc length parametrization in one of the coordinate directions). A variant of the RCASD results when $\theta_0 = \pi/2, a = \alpha = 0$ and $B = 1$. Fig. \[\text{[12]} \] show the MS-CDS for $\theta_0 = \pm \pi/2$. The special case of MS-CDS shown in Fig. \[\text{[12(a)]} \] fills space in a “spherical” manner that should fit a 3D i.i.d Gaussian distribution quite well. We will consider the performance of this special case later.

The nonzero components of the metric tensor under LoC coordinates are \[\{g_{12} = 0 \}\]
\[g_{11}(z_1, z_2) = (\varphi(z_1))^2 \left( \frac{\Delta}{\pi} \right)^2 \left( \alpha_0 - \varphi(z_1) \right)^2 + \cdots + a(\alpha_2 z_2)^2 \sin(\theta_0) \left( a(\alpha_2 z_2)^2 \sin(\theta_0) - \frac{\Delta}{\pi} (\alpha_0 - \varphi(z_1)) \right). \tag{180}\]
and, since the MS-CDS is convenient to use only for $\theta_0 = \pm \pi/2$,
\[ g_{22}(z_2) = B^2\alpha_2^2 + 4a^2\alpha_2 z_2^2. \quad (181) \]

**Channel pdf’s and Power:** Since the MS-CDS is convenient to use only for $\theta_0 = \pm \pi/2$, we have from (179) that $x_3 = \pm B\alpha_2 z_2$. Therefore $z_2 \sim N(0, \sigma_{z_2}^2)$ implying that the power on channel 2 becomes:
\[ P_2 = \text{Var}(z_2) = \sigma_{z_2}^2 = \frac{\sigma_z^2}{\alpha_2^2 B^2}. \quad (182) \]

For Channel 1 we have a similar situation as for the RCASD (a spiral covering the $x_1, x_2$ plane), except that the spiral radius shrinks somewhat for larger $x_3$ values. However, this effect is generally negligible, and so the same approximation as for RCASD works at high SNR. By choosing $\varphi$ as in (166) we get the same power for channel 1:
\[ P_1 = \text{Var}(z_1) = \frac{2}{\alpha_1^2} \left( 2\pi \varphi \sigma_z^2 \right)^2, \quad \eta = 0.16. \quad (183) \]

**Approximation distortion:** As for the RCASD, the MS-CDS leads to a uniform structure and so the approximation distortion is given by (168).

**Channel distortion:** From (126) we have
\[ \tilde{\varepsilon}_{ch}^2 = \frac{\sigma_z^2}{3} \int \sum_{i=1}^{2} g_{ii}(z)f_z(z)dz \]
\[ = \frac{\sigma_z^2}{3} \left[ \int g_{11}(z)f_z(z)dz + \int g_{22}(z)f_z(z)dz \right] \]
\[ = \sigma_z^2 \left( I_1 + I_2 \right) \]
\[ \frac{3}{1}, \quad (184) \]

where
\[ I_2 = \int g_{22}(z_2)f_{z_2}(z_2)dz_2 \]
\[ = \int \left( B^2\alpha_2^2 + 4a^2\alpha_2 z_2^2 \right)f_{z_2}(z_2)dz_2 \]
\[ = B^2\alpha_2^2 + 4a^2\alpha_2^4 \text{Var}(z_2) = \alpha_2^2 \left( B^2 + \frac{4a^2\sigma_z^2}{B^2} \right). \quad (185) \]

Further, we insert (166) into (180) and solve the relevant integral numerically.

**Optimization:** With constraint $C_t = P_{\text{max}} - P_1(\Delta, \alpha_1, \alpha_2) \geq 0$, we get the following objective function
\[ L(\Delta, \alpha_1, \alpha_2) = \tilde{\varepsilon}_{opt}^2(\Delta) + \tilde{\varepsilon}_{ch}^2(\alpha_1, \alpha_2) - \lambda C_t(\Delta, \alpha_1, \alpha_2). \quad (186) \]

We use the same procedure to determine the optimal parameters as for the RCASD for the special case $\theta_0 = -\pi/2$. The performance is shown in Fig. 13. The MS-CDS has similar performance as RCASD, being somewhat inferior at high SNR. The slope at high SNR is again lower than that of OPTA. For the MS-CDS in general, the 1:1 function mapping $x_2$ to $z_1, z_2$ has curvature $\kappa \neq 0$, unlike the RCASD. According to (172), the MS-CDS will then have a higher error in the ML estimate than the RCASD. As the approximation distortion is the same for the two structures, the MS-CDS will at best be on-par with the RCASD (at high SNR).

**Nonoptimality of slope of RCASD and MS-CDS**

There are two main reasons that RCASD and MS-CDS cannot obtain the OPTA 3:2 slope:

i) Both mappings can be seen as a composition of a 1:1 and a 2:1 mapping which is suboptimal according to Proposition 5.

ii) Any uniform S-K mapping has $\tilde{\varepsilon}_{ch}^2 \sim 1/\Delta^2$. According to Eq. (145) any uniform 3:2 mapping then has to have $\tilde{\varepsilon}_{ch}^2 \sim 1/\Delta$ to follow the same slope as 3:2 OPTA. For the RCASD we have that $\tilde{\varepsilon}_{ch}^2 \sim 1/\Delta^2$ according to (174), which is the wrong exponent. This is not as clear for the MS-CDS, but it is possible to verify that also for this case we have $\tilde{\varepsilon}_{ch}^2 \sim 1/\Delta^2$.

Point ii) is easily verified for any mapping that has distortion $D_t = \kappa_1\Delta^2 + \kappa_2\Delta^{-2}$ with $\kappa_1, \kappa_2$ constants. Differentiation w.r.t. $\Delta$ and solving for $\Delta$ one obtains $\Delta_{\text{opt}} = \sqrt{2\kappa_1/\kappa_2}$. As the two distortion contributions $\tilde{\varepsilon}_q^2 = \tilde{\varepsilon}_{ch}^2$ at high SNR (19), we can insert $\Delta_{\text{opt}}$ in e.g. $\tilde{\varepsilon}_q^2$ and so $\text{SDR} \sim \sqrt{\text{SNR}}$, which is the exponent of a 2:1 system.

4) **Snail Surface:** The Snail Surface is a simple example of a surface that cannot be decomposed into sub-mappings. The parametric equation of the Snail Surface consists of the components (111 p. 280)
\[ S_1(z_1, z_2) = a\varphi(z_1) \sin(\varphi(z_1)) \cos(\alpha_2 z_2 + \phi) \]
\[ S_2(z_1, z_2) = b\varphi(z_1) \cos(\varphi(z_1)) \cos(\alpha_2 z_2 + \phi) \quad (187) \]
\[ S_3(z_1, z_2) = c\varphi(z_1) \sin(\alpha_2 z_2 + \phi), \]

Fig. 12. The MS-CDS surface for $\Delta = 1, B = 4, a = 1$ and $\alpha_0 = -0.5$ with: (a) $\theta_0 = -\pi/2$, (b) $\theta_0 = \pi/2$.

Fig. 13. Theoretical performance of MS-CDS when $\theta_0 = -\pi/2$ compared to OPTA, BPAM and PCCOVQ.
which is valid for $0 \leq z_1 \leq k\pi, -\pi \leq z_2 \leq \pi$. To include negative values for $z_1$ we combine the above parametric equation with

\begin{align}
S_1(z_1, z_2) &= -a\varphi(z_1) \sin(\varphi(z_1) - \psi) \cos(\alpha_2 z_2 + \phi) \\
S_2(z_1, z_2) &= -b\varphi(z_1) \cos(\varphi(z_1) - \psi) \cos(\alpha_2 z_2 + \phi) \\
S_3(z_1, z_2) &= c \sin(\alpha_2 z_2 + \phi),
\end{align}

valid for $-k\pi \leq z_1 \leq 0, -\pi \leq z_2 \leq \pi$, and so obtain a double Snail Surface.

By choosing $\psi = \pi/2$ and $a = b = c = 2\Delta/\pi$ one obtains a spherical symmetry which further results in a (close to) uniform S-K mapping (see [5], Definition 6), and therefore \[168\] will be a good approximation of a lower bound of here as well, the choice of $\eta$ for $S^n$ scales with $S$, and so obtain a spherical symmetry which further results in a double Snail Surface. Note the spherical symmetry, well suited for a 3D i.i.d Gaussian pdf.

Further (as will be seen later), by choosing $\phi = \pi/2$ or 0, the pdf of $z_2$ will be symmetric around the origin. The first choice leads to lower probability for the highest channel values. The choice of mapping function $\varphi(z_1)$ becomes apparent when considering the metric tensor.

For a general mapping function on $z_1$, the two non-zero metric components become

\begin{align}
g_{11} &= (a\varphi'(z_1))^2(1 + \varphi(z_1) \cos^2(\alpha_2 z_2 + \phi)), \\
g_{22} &= a^2 \alpha_2^2 \varphi^2(z_1).
\end{align}

By choosing $\varphi(z_1) = \alpha_1 z_1$ we get the metric tensor

\begin{align}
g_{11} &= a^2 \alpha_1^2 (1 + \alpha_1^2 z_1^2 \cos^2(\alpha_2 z_2 + \phi)), \\
g_{22} &= a^2 \alpha_1^2 \alpha_2^2 z_1^2, \\
g_{12} &= g_{21} = 0.
\end{align}

One can observe that $g_{ii} \sim z_1^2$, $i = 1, 2$, implying that $\varepsilon_{ch}^2$ increases with $z_1^2$. One can compensate this for both $g_{ii}$ components simultaneously by choosing $\varphi \sim \sqrt{z_1}$. As the RCASD also has $g_{11} \sim z_1^2$ when $\varphi(z_1) = \alpha_1 z_1$, and that SnaSu scales with $\Delta$ like RCASD, it makes sense to use $\alpha_1$ for here as well, the choice of $\eta$ being arbitrary.

**Optimization of Snail Surface as 3:2 mapping.**

**Channel Power and Density Function:** To derive the pdf of $z_1$ consider Fig. [15(a)] By letting $z_2$ vary slightly ($\pm \epsilon$) around some constant value (here $\pi/2$) with $z_1$ free we get a cork-screw like structure. In the limit of constant $z_1$ we get a spiral with torsion $\tau \neq 0$. The rate at which the spiral rises from the $x_1$-$x_2$-plane depending on the value of $z_2$. Whenever $z_2 = \pm(2n + 1)\pi/2$, $n \in \mathbb{N}$ the torsion is maximal, whereas when $z_2 = \pm n\pi$, $n \in \mathbb{N}$ the spiral is a plane curve ($\tau = 0$). This implies that the mapping from the snail surface to $z_1$ can be seen as a “radius” of a sphere that traces out all points inside this sphere depending on the value of $z_1$ and $z_2$, depending on the values of the sources $x_1, x_2, x_3$. This radius will then be $\rho = \sqrt{x_1^2 + x_2^2 + x_3^2}$. By assuming that the snail surface is dense enough, i.e., with very small values of $\Delta$, one can approximate $\mathcal{S}$ as a mapping $h : \mathbb{R}^3 \to \mathbb{R}$ (this assumption becomes more and more accurate as the SNR gets larger since $\Delta$ becomes smaller). This mapping depends on $\varphi$. We first assume a general power function $\varphi = (\gamma z_1)^n, n \in \mathbb{Q}^+$. Then we get a mapping $\mathcal{Z} : h(x_1, x_2, x_3) = \pm\gamma a^n(x_1^2 + x_2^2 + x_3^2)^{n/2}$, and so $z_1 = \ell(\rho) = \pm\gamma a^{-n}\rho^n = \pm\gamma a^{-n}(x_1^2 + x_2^2 + x_3^2)^{n/2}$.

The cumulative distribution is then given by (a straight forward generalization of the $h : \mathbb{R}^2 \to \mathbb{R}$ case in [6], pp. 180-181)

\begin{equation}
F_{z_1}(z_1) = p_r\{z_1 \leq z_1\} = p_r\{(x_1, x_2, x_3) \in D^+_z \cup D^-_z\} = \int_{D^+_z \cup D^-_z} f_{X_1X_2X_3}(x_1, x_2, x_3)dx_1dx_2dx_3,
\end{equation}

where $f_{X_1X_2X_3}(x_1, x_2, x_3)$ is the joint Gaussian distribution and

\begin{align}
D^+_z &= \{(x_1, x_2, x_3)\mid (x_1^2 + x_2^2 + x_3^2)^{n/2} \leq \rho^n, z_1 \geq 0\}, \\
D^-_z &= \{(x_1, x_2, x_3)\mid (x_1^2 + x_2^2 + x_3^2)^{n/2} \geq -\rho^n, z_1 < 0\}.
\end{align}

The pdf is found by differentiating (191) with respect to $z_1$. To simplify, one can consider only the positive domain in (192) then use the fact that the channel pdf will be symmetric about the origin with the chosen $\mathcal{S}$. Further, since the domain in (192) is spherical, it is convenient to do the integration in spherical coordinates [20]

\begin{equation}
f_{z_1}(z_1) = \frac{1}{2z_1} \int_0^{2\pi} \int_0^\pi \int_{a^\varphi(z_1)} \rho^2 \sin(\theta) d\rho d\theta d\phi,
\end{equation}

where $f_r = \exp(-\rho^2/(2\sigma^2_r))/((2\pi)^{3/2}\sigma^2_r)$. The integrals
over \(\theta, \phi\) become \(I(\theta, \phi) = \pi\). The integral over \(\rho\) remains,

\[
\frac{d}{dz_1} \int_{0}^{\alpha z_1} f_{\rho}(\rho) \rho^2 d\rho = \frac{1}{(2\pi)^{3/2}\sigma_\varphi^2} \int_{0}^{\alpha z_1} e^{-\frac{\rho^2}{2\sigma_\varphi^2}} d\rho
\]

\[
= \frac{na^3 \gamma^3 |z_1|^{3n-1}}{(2\pi)^{3/2}\sigma_\varphi^2} e^{-\frac{a^2 \gamma |z_1|}{2\sigma_\varphi^2}}.
\]

(194)

By multiplying in \(\pi\) and further using absolute value (to make the pdf symmetric around the origin) we get

\[
f_{z_1}(z_1) = \frac{na^3 \gamma^3 |z_1|^{3n-1}}{2\sqrt{2\pi\sigma_\varphi^2}} e^{-\frac{a^2 \gamma |z_1|}{2\sigma_\varphi^2}}.
\]

(195)

If we now assume that \(\varphi(z_1)\) is as in (160), then \(\gamma = \alpha_1/(\eta\Delta)\) and so

\[
f_{z_1}(z_1) = \frac{a^3 \alpha_1^{3/2} \sqrt{|z_1|}}{2\sqrt{2\pi\sigma_\varphi^2}(\eta\Delta)^{3/2}} e^{-\frac{a^2 \alpha_1 |z_1|}{2\sigma_\varphi^2}}.
\]

(196)

According to [6, p.87] the gamma distribution has the general form \(f_\gamma(x) = u(x) e^{-x} e^{-\frac{x}{(\Gamma(b))b}}\). From this it is easy to realize that (196) is a double gamma distribution with \(c = 3/2\) and \(b = (2\eta\Delta\sigma_\varphi^2)/(a^2 \alpha_1)\). From (6, p.154) we have that the second moment of gamma distribution is \(E\{x^2\} = c(c+1)b^2\). Then, since (196) with zero mean, its variance, and thereby the power of channel 1 becomes

\[
P_1 = \sigma_{z_1}^2 = \frac{15(\eta\Delta\sigma_\varphi^2)^2}{a^2 \alpha_1^4} = \frac{15(\eta\pi^2\sigma_\varphi^2)^2}{16\alpha_1^4\Delta^2}.
\]

(197)

To derive the pdf of \(z_2\) consider Fig. [5(b)] By perturbing \(z_2\) with \(\varepsilon_2\) around some constant value (here 2\pi) with \(z_2\) free, we get two Möbius strips. In the limit \(\varepsilon_2 \to 0\) we get a circle "rotating" about an axis whose radius increases as \(2\Delta z_2/\pi\). Consider \(\phi = 0\). Then the rotation axis is at \(\pi/2\). The radius of the rotating circle is insignificant as \(z_2 \in [-\pi, \pi]\), independent of \(\pi\). From the perspective of \(z_2\), as the joint pdf of \(x\) is spherically symmetric, we have a uniform mass distribution over a virtual spherical shell of arbitrary radius, \(r_0\), as depicted in Fig. [10]. To find the probability mass associated with different values of \(z_2\), one considers the sum of all points along circles (green circle in Fig. [10]) resulting from intersections of this virtual sphere by planes perpendicular to the rotation axis. The radius, \(r_1\), of such a circle is \(r_1 = r_0 \cos(\nu)\), where \(\nu = z_2\) is the angle from the equatorial plane, \(\nu = \pi/2 - \phi\), and \(\phi\) the polar angle. The circumference as a function of \(z_2\) is \(O(2\pi) = 2\pi r_0 \cos(z_2)\).

Since \(r_0\) is arbitrary, one can set \(r_0 = 1/(2\pi)\) implying that \(f_{z_2}(z_2) \sim |\cos(z_2)|\), \(z_2 \in [-\pi, \pi]\). To avoid high probability for the largest channel amplitude values, one can set \(\phi = \pi/2\) to obtain zero probability there. I.e., a sine distribution results. To normalize, as \(\int_{-\pi}^{\pi} \sin(z_2) dz_2 = 2\pi\) then

\[
f_{z_2}(z_2) = \frac{\alpha_2}{4} |\sin(\alpha_2 z_2)|.
\]

(198)

Since \(f_{z_2}(z_2)\) is proportional to Gilberts sine distribution \(f_s(x) = \sin(2x)\), which according to [21] has variance \(E\{x^2\} = \pi^2/4 - 1/2\), the power for channel 2 becomes

\[
P_2 = \text{Var}(z_2) = \frac{2\alpha_2^2}{\pi^2/4 - 1}.
\]

(199)

\[
\begin{align*}
\text{Fig. 16.} \quad \text{The sphere used to compute the channel pdf of } z_2 \text{ for the snail surface.}
\end{align*}
\]

\[
\begin{align*}
\text{Weak Channel Distortion: From (126) we get}
\end{align*}
\]

\[
\varepsilon_{ch}^2 = \frac{\alpha_2^2}{3} \int \left( g_{11}(z) + g_{22}(z) \right) f_k(z) dz
\]

\[
= \frac{\alpha_1^2}{3} \int \left( I_1 + I_2 \right)
\]

\[
I_2 = \int \left( g_{22}(z) f_k(z) dz = (\alpha_2)^2 \int \varphi(\alpha_1 z_1) f_{z_1}(z_1) dz_1 = (2\alpha_2)^2 \int \varphi(\alpha_1 z_1) f_{z_1}(z_1) dz_1 \right)
\]

\[
= 2(\alpha_2)^2 \frac{\alpha_1}{\eta\Delta} \int_{-\infty}^{\infty} z_1 f_{z_1}(z_1) dz_1 = 6\alpha_2^2 \sigma_{z_2}^2.
\]

(200)

(201)

The last equality comes from the fact that \(z_1\) is gamma distributed and therefore the integral in (201) becomes [6, p.154] \(bc/2 = 3\eta\Delta\sigma_\varphi^2/(a^2 \alpha_1)\).

Further, we have (using (189))

\[
I_1 = \int g_{11}(z) f_k(z) dz = a^2 \left( \int \varphi'(\alpha_1 z_1)^2 f_{z_1}(z_1) dz_1 \right)
\]

\[
+ \int \varphi'(\alpha_1 z_1)^2 f_{z_1}(z_1) dz_1 \int \cos^2(\alpha_2 z_2 + \phi) f_{z_2}(z_2) dz_2
\]

\[
= a^2 \left( \frac{\alpha_1}{4\eta\Delta} E[z_1^{-1}] + \frac{\alpha_2^2}{6\eta\Delta^2} \right) = \frac{\alpha_1}{\eta\Delta} E[z_1^{-1}] + \frac{2\alpha_1^2}{3\pi^2\eta^2},
\]

(202)

where we used that \(\varphi'(\alpha_1 z_1) = \alpha_1/(4\eta\Delta z_1)\). Through the series expansion \((1 + x)^{-1} = 1 - x + x^2 - x^3 + x^4 + \cdots\), setting \(x = z_1 - 1\), we get \(E[z_1^{-1}] \approx E[4 - 6z_1 + 4z_1^2 - z_1^3] = 4(1 + \sigma_{z_1}^2)\) up to 3rd order, with \(\sigma_{z_1}^2\) as in (197). Then

\[
I_1 \approx \frac{4\alpha_1\Delta}{\eta\pi^2} (1 + \sigma_{z_1}^2) + \frac{2\alpha_1^2}{3\pi^2\eta^2}.
\]

(203)

Therefore the channel distortion becomes

\[
\varepsilon_{ch}^2 \approx \frac{\sigma_{z_2}^2}{3} \left( \frac{4\alpha_1\Delta}{\eta\pi^2} (1 + \sigma_{z_1}^2) + \frac{2\alpha_1^2}{3\pi^2\eta^2} + 6\alpha_2^2 \sigma_{z_2}^2 \right).
\]

(204)
Optimization: With constraint \( C_t = P_{\text{max}} - P_t(\Delta, \alpha_1, \alpha_2) \geq 0 \), we get the following objective function

\[
\mathcal{L}(\Delta, \alpha_1, \alpha_2) = \varepsilon_i^2(\Delta) + \varepsilon_{\text{sh}}^2(\Delta, \alpha_1, \alpha_2) - \lambda C_t(\Delta, \alpha_1, \alpha_2).
\]

(205)

The performance of the optimized Snail Surface is shown in fig. 17.

![Performance of Double Snail Surface](image)

Fig. 17. Performance of Snail Surface (simulated and calculated) compared to OPTA, BPAM and PCCOVQ.

The Snail Surface clearly improves with SNR, beating both the linear BPAM as well as PCCOVQ (see footnote 9), and is noise robust.

Curvature Analysis: By expanding (112) using the rules of determinants [22, p. 227] we get

\[
b_{ij} = b_{ji} = \frac{1}{\sqrt{g_{11}g_{22}}} \left( \frac{\partial S_1}{\partial z_1} \frac{\partial^2 S_1}{\partial z_2^2} \frac{\partial^2 S_1}{\partial z_1 \partial z_2} - \cdots \right)
\]

\[
\frac{\partial S_1}{\partial z_2} \frac{\partial S_1}{\partial z_1} + \frac{\partial^2 S_1}{\partial z_2 \partial z_1} \frac{\partial^2 S_1}{\partial z_1 \partial z_2} \right).
\]

(206)

The calculation of all derivatives and determinants above is a quite tedious process, so we do not show the detail here. With \( \theta = \alpha_2 z_2 + \phi \), one arrives at the following expressions

\[
b_{11} = -\frac{a(\varphi(z_1))}{\sqrt{1 + \varphi^2(z_1)}} \cos^2 \theta, \quad b_{22} = -\frac{a \alpha_2 \varphi^2(z_1) \cos \theta}{\sqrt{1 + \varphi^2(z_1)}}
\]

\[
b_{12} = \frac{a \alpha_1 \alpha_2 \varphi^2(z_1) \sin \theta}{\sqrt{1 + \varphi^2(z_1)}}
\]

(207)

From this we can conclude that we do not have LoC coordinates: According to Theorem \( \ref{locc} \) LoC coordinates require that \( g_{12} = b_{12} = 0 \). From the expression for \( b_{12} \) in (207) we see that \( b_{12} \neq 0 \), except when \( \alpha_2 z_2 + \phi = 0 \), which is not a realistic possibility when both channel variables are used for communication.

As the coordinates are not LoC, the principal curvatures are the roots of \( \xi \):

\[
\kappa_{1/2} = \frac{1}{2} \left( \frac{b_{11} + b_{22}}{g_{11}} + \sqrt{\left( \frac{b_{11} + b_{22}}{g_{11}} \right)^2 - 4 \left( \frac{b_{11} b_{22} - b_{12}^2}{g_{11} g_{22}} \right)} \right).
\]

(208)

Evaluation of \( \kappa_2 \) as function of \( \Delta \) and \( \alpha_1 \) is shown in Fig. 18(a). Note that we have set all sine and cosine terms in (189) and (207) to one, implying that the 3D plot above shows a somewhat higher curvature than the actual value. The curvature is getting smaller as \( \Delta \) decreases and \( \alpha_1 \) increases, corresponding to the high SNR case. However, the curvature is quite large when \( \Delta \) is large and \( \alpha_1 \) is small, corresponding to the low SNR case. The increase in curvature is much more dramatic than what was the case for RCASD, which may explain why the RCASD is performing better than the Snail Surface at lower SNR (see comparison in Fig. 19). By studying the geometric configurations of RCASD in Fig. 14 and Snail Surface in Fig. 13 one can expect the Snail Surface to have somewhat higher curvature.

By inserting optimized parameters for 30dB SNR found by the optimization procedure below (\( \Delta^* = 0.539, \alpha_1^* = 4.76, \alpha_2^* = 2.57 \)) one obtains maximal curvature \( |\kappa_2| < 1 \) averaged over the relevant range of \( z_1 \). By considering the distortion terms in (106), with total transmission power 1, then \( \sigma_n^2 = 0.001 \), and one can see that the 1st order term is in the order of about 0.001/0.001^2 = 1000 over the 2nd order term. Therefore, the Snail Surface is also a mapping following Definition \( \ref{mapping} \) at high SNR.

Further investigation of coordinate curves on Snail Surface 3:2 mapping: We already know that we do not have LoC coordinates. To figure out whether or not we have geodesic coordinates (GeCo) we compute the Christoffel symbols \( \Gamma^2_{12} \) and \( \Gamma^1_{11} \). Using the definitions in (64) and (63) we get

\[
\Gamma_{\alpha \beta \gamma} = \frac{1}{2} \left[ \frac{\partial g_{\beta \lambda}}{\partial z^\alpha} + \frac{\partial g_{\lambda \alpha}}{\partial z^\beta} - \frac{\partial g_{\alpha \beta}}{\partial z^\gamma} \right].
\]

(209)

With \( g^{11} = g^{22} = g = 1/g_{11} \) (since \( g_{12} = 0 \)), \( g_{22} = g^{11}/g = 1/g^{22} \) and \( g_{12} = -g^{12}/g = 0 \), then from (64), we get

\[
\Gamma^2_{11} = -\frac{1}{2 g_{22}} \frac{\partial g_{11}}{\partial z_2} = \frac{(\varphi(z_1))^2}{\alpha_2} \cos(\alpha_2 z_2 + \phi) \sin(\alpha_2 z_2 + \phi),
\]

(210)

and

\[
\Gamma^2_{22} = -\frac{1}{2 g_{11}} \frac{\partial g_{22}}{\partial z_1} = \frac{\alpha_2^2 \varphi(z_1)}{\varphi(z_1)(1 + \varphi(z_1) \cos^2(\alpha_2 z_2 + \phi))}.
\]

(211)
By inserting \( \varphi(z_1) \) from equation (166) one will see that both of these are \( \neq 0 \) in general (\( \Gamma^2_{11} = 0 \) for \( \alpha_2 z_2 + \phi = 0, \pi/2, \pi \)). Therefore, according to Theorem 8 we do not have GeCo in the SnSu system above.

What one would like to obtain, if possible, is to find coordinates where \( g_{ij} \) is a function of \( z_i \) only and where \( \varepsilon_{ch}^2 = \Delta^{-1} \). If this can be obtained with geodesic coordinates is unknown. This is quite a demanding problem and further research is needed to conclude.

5) Comparison of 3:2 Mappings: In Fig. 19 we compare all 3:2 mappings proposed in this section. We see that Helicoid and RCASD is quite well performing in the range 10-22 dB channel SNR. However, the Snail Surface is clearly superior at higher channel SNR, following the same slope as OPTA to a much higher SNR. However, it will eventually also diverge from OPTA.

**APPENDIX A**

**UNIT SPEED PARAMETRIZATION.**

Assume a parametrization of a curve \( s(\varphi(x)) \). For every continuous parameter curve there exists a special \( \varphi \) that makes all tangent vectors along the curve unit vectors.

Let \( s: x \in [a, b] \subseteq \mathbb{R} \rightarrow s(x) \in \mathbb{R}^N \) be a parametrization for the curve \( C \). Let \( \ell(x) \) denote the length function of \( s \) given by

\[
\ell(x) = \int_a^x \| s'(q) \| \, dq
\]

and \( \varphi \) denote its inverse.

**Theorem 19:** Let \( y(\ell) \) be a curve length parametrization of \( C \). Then \( y(\ell) \) and \( s(\varphi(x)) \) will have the same image, and \( \|y'(\ell)\| = \|s'(\varphi(\ell))\| = 1, \quad \forall \ell \).

**Proof 3:** See [24] pp. 115-116.

**APPENDIX B**

**THE METRIC TENSOR**

Consider a hyper surface \( S \) realized by the parametric equation (65) (or 66). The **metric tensor** (also named Riemannian **Metric**) [24] pp.301-343 for a smooth embedding \( S \) in \( \mathbb{R}^N \) is given by:

\[
G = J^T J = \begin{bmatrix}
g_{11} & g_{12} & \cdots & g_{1M} 
g_{21} & g_{22} & \cdots & g_{2M} 
\vdots & \vdots & \ddots & \vdots 
g_{M1} & g_{M2} & \cdots & g_{MM}
\end{bmatrix}
\]

(213)

where \( J \) is the Jacobian [25] p.47 of \( S \), given by

\[
J = \begin{bmatrix}
\frac{\partial s_1}{\partial x_1} & \frac{\partial s_1}{\partial x_2} & \cdots & \frac{\partial s_1}{\partial x_M} \\
\frac{\partial s_2}{\partial x_1} & \frac{\partial s_2}{\partial x_2} & \cdots & \frac{\partial s_2}{\partial x_M} \\
\vdots & \vdots & \ddots & \vdots \\
\frac{\partial s_N}{\partial x_1} & \frac{\partial s_N}{\partial x_2} & \cdots & \frac{\partial s_N}{\partial x_M}
\end{bmatrix}^T
\]

(214)

The metric tensor \( G \) is symmetric and positive definite [24], \( g_{ij} \) can be interpreted as the squared length of the tangent vector in the direction of parameter \( x_i \), where \( x_i \) is the \( i \)'th parameter in a parametric description of \( S \). All “cross terms” \( g_{ij} \), are the inner product of the tangent vectors in the direction of \( x_i \) and \( x_j \). See [24] pp.301-347 for details.

**REFERENCES**

[1] F. Hekland, P. A. Floor, and T. A. Ramstad, “Shannon-Kotel’nikov mappings in joint source-channel coding,” IEEE Trans. Commun., vol. 57, no. 1, pp. 94–105, Jan. 2009.

[2] F. Hekland, “On the design and analysis of Shannon-Kotel’nikov mappings for joint source-channel coding,” Ph.D. dissertation, Norwegian University of Science and Engineering (NTNU), 2007.

[3] P. A. Floor, “On the theory of Shannon-Kotel’nikov mappings in joint source-channel coding,” Ph.D. dissertation, Norwegian University of Science and Engineering (NTNU), 2008.

[4] E. Kreyszig, *Differential Geometry*. Dover Publications, Inc., 1991.

[5] P. A. Floor and T. Ramstad, “Shannon-Kotel’nikov mappings for analog point-to-point communications,” arXiv:1101.5716v2 [cs.IT], 2012.

[6] A. Papoulis and S. U. Pillai, *Probability, Random Variables and Stochastic Processes*, 4th ed. New York: McGraw-Hill higher education, Inc., 2002.

[7] J. M. Wozencraft and I. M. Jacobs, *Principles of Communication Engineering*. New York: John Wiley & Sons, Inc., 1965.

[8] K. Rottmann, *Mathematische Formelsammlung*. Bibliographisches Institut & F. A. Brockhaus, 1991.

[9] C. E. Shannon, “Communication in the presence of noise,” Proc. IRE, vol. 37, pp. 10–21, Jan. 1949.

[10] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer-Verlag, 1999.

[11] S. N. Krvoshapko and V. N. Ivanov, *Encyclopedia of Analytical Surfaces*. Springer International Publishing Switzerland, 2015.

[12] E. Akyol, K. B. Viswanatha, K. Rose, and T. A. Ramstad, “On zero-delay source-channel coding,” IEEE Trans. Information Theory, vol. 60, no. 12, pp. 7473–7489, Dec. 2014.

[13] M. S. Mehmetoglu, E. Akyol, and K. Rose, “Deterministic annealing-based optimization for zero-delay source-channel coding in networks,” IEEE Transactions on Communications, vol. 63, no. 12, pp. 5089–5100, Dec. 2015.

[14] P. A. Floor, A. N. Kim, T. Ramstad, and I. Balasingham, “Zero delay joint source-channel coding for multivariate Gaussian sources over orthogonal Gaussian channels,” Entropy, vol. 15, no. 6, pp. 2129–2161, Jun. 2013.

[15] Y. Hu, J. Garcia-Frias, and M. Lamarca, “Analog joint source-channel coding using non-linear curves and mmse decoding,” IEEE Trans. Commun., vol. 59, no. 11, pp. 3016–3026, Nov. 2011.

[16] K.-H. Lee and D. P. Petersen, “Optimal linear coding for vector channels,” IEEE Trans. Commun., vol. COM-24, no. 12, pp. 1283–1290, Dec. 1976.

[17] A. Fuldt and T. A. Ramstad, “Bandwidth compression for continuous amplitude channels based on vector approximation to a continuous subset of the source signal space,” in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Proc. (ICASSP)*, 1997.
[18] A. Fuldseth, “Robust subband video compression for noisy channels with multilevel signaling,” Ph.D. dissertation, Norwegian University of Science and Engineering (NTNU), 1997.

[19] F. Hekland, G. E. Øien, and T. A. Ramstad, “Using 2:1 Shannon mapping for joint source-channel coding,” in Proc. Data Compression Conference, IEEE. Snowbird, Utah: IEEE Computer Society Press, Mar. 2005, pp. 223–232.

[20] W. D. Richter, “Generalized spherical and simplicial coordinates,” Journal of Mathematical Analysis and Applications, vol. 336, pp. 1187–1202, 2007.

[21] A. Edwards, “Gilberts sine distribution,” Teaching Statistics, vol. 22, no. 3, pp. 70–71, 2000.

[22] G. Strang, Linear Algebra and its Applications, 3rd ed. Harcourt Brace Jovanovich, 1988.

[23] J. J. Callahan, The Geometry of Spacetime: An Introduction to Special and General Relativity, S. Axler, F. W. Gehring, and K. A. Ribet, Eds. New York: Springer-Verlag, Inc, 2000.

[24] M. Spivak, A Comprehensive Introduction to Differential Geometry, 3rd ed. Publish or Perish, Houston Texas, Inc, 1999.

[25] J. R. Munkres, Analysis on Manifolds. Westview Press, 1991.