Phase Shift Keying on the Hypersphere: Peak Power-Efficient MIMO Communications

Christoph Rachinger, Ralf R. Müller and Johannes B. Huber

***Abstract*—Phase Shift Keying on the Hypersphere (PSKH), a generalization of conventional Phase Shift Keying (PSK) for Multiple-Input Multiple-Output (MIMO) systems, is introduced. In PSKH, constellation points are distributed on a multidimensional hypersphere. The use of such constellations with a Peak-To-Average-Sum-Power-Ratio (PASPR) of 1 allows to use load-modulated transmitters which can cope with a small backoff, which in turn results in a high power efficiency. In this paper, we discuss several methods how to generate PSKH constellations and compare their performance. After applying conventional Pulse-Amplitude Modulation (PAM), the PASPR of the continuous time PSKH signal depends on the choice of the pulse shaping method. This choice also influences bandwidth and power efficiency of a PSKH system. In order to reduce the PASPR of the continuous transmission signal, we use spherical interpolation to generate a smooth signal over the hypersphere and present corresponding receiver techniques. Additionally, complexity reduction techniques are proposed and compared. Finally, we discuss the methods presented in this paper regarding their trade-offs with respect to PASPR, bandwidth, power efficiency and receiver complexity.**

***Index Terms*—Multiple-input multiple-output systems, wireless communications, peak-to-average-power, load-modulation.**

I. Introduction

POWER efficiency is one of the driving forces behind the development of current communication technologies. Unfortunately, one of the main sources of power consumption are amplifiers operating at low efficiency. This holds even for state-of-the-art amplifiers. They cannot be operated at their optimal power level, because signals with suboptimal Peak-To-Average-Power-Ratio (PAPR), i.e., PAPR > 1, require to backoff the amplifier to avoid serious clipping. Constant envelope modulation schemes such as continuous phase modulation (CPM) provide an optimal PAPR and reduce the backoff compared to other modulation schemes, such as the widely used QAM [1]. While this improves the power efficiency of a transmission system, the bandwidth efficiency suffers: Since the radius in the complex plane is fixed for constant envelope transmission, phase remains the only degree of freedom to represent information. This results in a rate loss compared to QAM, which is the reason why constant envelope modulation did not receive much attention since the development of GSM-Mobile Communiation, a system employing Gaussian Minimum Shift Keying (GMSK), a variant of CPM.

C. Rachinger and J. B. Huber are with the Institute for Information Transmission, University of Erlangen-Nueremberg (email: christoph.rachinger@fau.de, johannes.huber@fau.de).

Ralf R. Müller is with the Institute of Digital Communications, University of Erlangen-Nueremberg (email: ralf.r.mueller@fau.de).

This paper has been submitted to IEEE Transactions on Wireless Communications.

The use of MIMO systems allows for another method to increase power efficiency: Not the PAPR, but the Peak-to-Average-Sum-Power-Ratio (PASPR) of a vector-valued signal can determine the required amplifier backoff. For an arbitrary vector-valued function $\mathbf{x}(t) \in \mathbb{C}^n$, this quantity is defined as

$$\mathrm{PASPR}(\mathbf{x}(t)) = \frac{\max_t \|\mathbf{x}(t)\|^2}{\mathcal{E}\{\|\mathbf{x}(t)\|^2\}}. \tag{1}$$

The PASPR is a decisive factor when recently proposed load-modulated MIMO amplifiers are used [2], [3]. Since the degrees of freedom are reduced by only one for all antennas, the relative rate loss is smaller the more antennas are used [4]. For massive MIMO systems, the central limit theorem (CLT) guarantees that the PASPR of the continuous-time signal becomes optimal as long as the data points are distributed on a multidimensional hypersphere. We call these constellations *Phase Shift Keying on the Hypersphere* (PSKH), because it is a natural extension of ordinary PSK. In conventional MIMO with only a handful of antennas, large fluctuations of the continuous transmit signal are still possible and therefore the PASPR is far from being optimal. Thus some more adaptations are necessary in order to reduce the PASPR of the transmission signal.

The rest of this paper is organized as follows: In Sec. II we introduce our system model. Unlike in PSK, there are multiple ways to construct constellation on the hypersphere. Thus we discuss several algorithms to generate PSKH constellations and their advantages and disadvantages in Sec. III. Secs. IV and V presents two approaches to reduce the PASPR in detail. This includes receiver structures for the corresponding signals as well as numerical results for their performance. Sec. VI discusses how much the previously introduced approaches reduce the PASPR of the transmitter output signal and how they affect the spectrum and thus the bandwidth efficiency. The paper ends with conclusions in Sec. VII.

II. System Model

We define a PSKH constellations as a set of $M = 2^{R_m}$ data points $\mathcal{A} = \{\mathbf{a}_0, \ldots, \mathbf{a}_{M-1} \mid \mathbf{a}_i \in \mathbb{C}^{n_T}, \|\mathbf{a}_i\| = \sqrt{E_s}\}$ where E_s is the energy per symbol, n_T is the number of transmit antennas and R_m the rate per modulation interval. Unless otherwise mentioned, these constellations are modulated using conventional PAM with a pulse shaping filter $h(t)$ with normalized symbol period $T = 1$ to generate the continuous-time transmitter output signal in the equivalent complex baseband

(ECB) domain

$$\mathbf{s}(t) = \sum_{k=-\infty}^{\infty} \mathbf{x}[k]h(t-k), \quad \mathbf{x}[k] \in \mathcal{A} \tag{2}$$

for a given data sequence $\langle \mathbf{x}[k] \rangle$.

If $h(t)$ is a $\sqrt{\text{Nyquist}}$ filter, the channel is not frequency selective, and the corresponding matched filter is applied at the receiver, the overall model is the well known discrete-time MIMO ECB channel model

$$\mathbf{y}[k] = \mathbf{H}\mathbf{x}[k] + \mathbf{n}[k] \tag{3}$$

where $\mathbf{x}[k] \in \mathbb{C}^{n_T}, \mathbf{y}[k] \in \mathbb{C}^{n_R}$ are transmit and receive vector at time k, respectively, $\mathbf{H} \in \mathbb{C}^{n_T \times n_R}$ is the channel matrix and $\mathbf{n}[k] \in \mathbb{C}^{n_R}$ is complex i.i.d. additive white Gaussian noise with variance $\sigma^2 = \frac{N_0}{T} = N_0$ per complex component. n_R and n_T denote the number of transmit and receive antennas respectively, but for the remainder of this paper we assume that $n = n_R = n_T$ and omit the time index k unless confusion is possible. If other pulse shaping filters than $\sqrt{\text{Nyquist}}$ are used, it will be discussed in detail.

For this work, both the continuous and the discrete time models (eqs. (2) and (3)) are important, because the first one determines the PASPR and bandwidth of the signal whereas the latter can be used for the detection of the transmitted sequence in the receiver.

Because every point in a PSKH constellation in n dimensions has energy E_s and uses $\sqrt{\text{Nyquist}}$ impulses, the equivalent energy per bit for uncoded transmission is given as E_s/R_m at the transmitter side. In this paper, we use two different channel models: $\mathbf{H}$ can be unitary, which corresponds to a vector AWGN channel after equalization. In this case, transmitter and receiver energy are equal and the received energy per bit over the one-sided noise-spectral density is given as

$$\frac{E_b}{N_{0\,\text{AWGN}}} = \frac{E_s}{R_m \sigma^2}. \tag{4}$$

Our second channel model is the flat-fading Rayleigh model, i.e., the entries of $\mathbf{H}$ are i.i.d. complex Gaussian random variables with unit variance. Thus each antenna receives an average signal energy of E_s per symbol and hence the total received energy over n receive antennas is nE_s. The *average* received energy per bit over the one-sided noise-spectral density is then given as

$$\frac{E_b}{N_{0\,\text{Fading}}} = n\frac{E_b}{N_{0\,\text{AWGN}}} = \frac{nE_s}{R_m \sigma^2}. \tag{5}$$

We omit the subindices AWGN or Fading and instead specify the channel we use in a given scenario.

III. PSKH Constellations

A. Constellation Construction

As explained in Sec. II, a PSKH constellation is a set of $M = 2^{R_m}$ points on the hypersphere with radius $\sqrt{E_s}$ representing R_m bits. The vectors $\mathbf{a}_i \in \mathcal{A} \subset \mathbb{C}^n$ are n-dimensional corresponding to n transmit antennas in a MIMO system. We note that PSKH constellations are also known as *spherical codes* in literature, but to our knowledge they have never been used to improve power efficiency of communication systems by means of PASPR reduction. Without further restrictions than the radius, there are many possible ways to create constellations, which might differ quite vastly in terms of quality. A reasonable measure for quality, as in all PAM schemes, is the minimum distance between signal points. Optimal codes in this sense and their analytic description, however, are known only for some restricted constellation sizes and dimensions [5].[1] Therefore, we compare four different algorithms to generate PSKH constellations:

- *Equal Area Partitioning Algorithm* (EQPA) from [7]: Generates a constellation with equally sized areas, which are usually not the Voronoi regions of a data point.
- *k-means Clustering* (kMC) using the spherical k-means algorithm [8]: Generates a large number of uniformly distributed points on the sphere, clusters them using the spherical k-means algorithm.
- *Potential Minimization* (PM): Generates particles on a sphere and minimizes the potential energy between particles. This can be done via a molecular dynamics simulation [9].
- *Per-Antenna PSK* (PA-PSK): Generates independent PSK constellations on each antenna, then scales them to fit the power constraint.

The algorithms can be distinguished in terms of construction complexity: EQPA and PA-PSK are analytic constructions, kMC and PM rely on numerical methods and are therefore more expensive to construct. Of course, such a construction needs to be done only once and can be computed offline. If construction is nonanalytic, it is further necessary to store the constellation in memory, which we think is reasonable to implement for $R_m \lesssim 16$. Additionally, it is possible to construct a constellation for only half the number of antennas and duplicate it, which results in a small degradation of quality.

In order to compare constellations with respect to their performance, we take a look at three different properties: Constellation-constrained capacities, minimum distance of the constellation (also known as *packing radius* in the context of spherical codes and packings) and error probabilities.

B. Capacity of PSKH

In [4], it is proven that the capacity of PSKH for unitary $\mathbf{H}$ and continuous input is achieved for a uniform distribution on the hypersphere. In this section, we compare the capacities if the input is discrete and constrained to a certain size, but $\mathbf{H}$ is still unitary. Fig. 1 shows the constellation constrained capacities for $n = 3$ antennas and $M = 64$ as well as $M = 512$ points. The results are similar for different constellation and antenna array sizes, which is why we restrict ourselves to one exemplary case. As a baseline, we also plot the AWGN capacity for continuous input. For $n = 3$ antennas, this capacity is $C = 3\log(1 + \text{SNR}_{\text{ch}})$ with $\text{SNR}_{ch} = \frac{E_s}{n\sigma^2}$ being the SNR on each individual AWGN channel. For Fig. 1, we use the standard representation over $\frac{E_b}{N_0}$. For the vector AWGN channel, we have $\frac{E_b}{N_0} = \frac{\text{SNR}_{\text{ch}}}{C/n}$. The general result

[1] Some examples of such optimal packings can be found in [6].

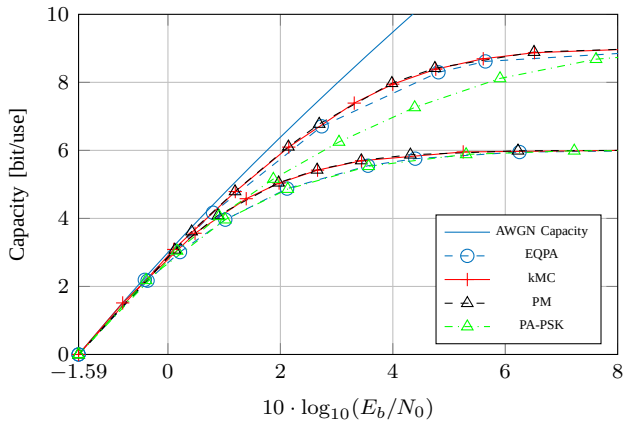


Fig. 1. Capacity of different constellations for a constellation sizes $M = 64$ and $M = 512$ points and $n = 3$ antennas over a vector AWGN channel.

regarding capacities can be summarized as follows: PM and kMC have the best capacities with PM outperforming kMC by only approx. 0.01 dB. For a constellation size of 1 bit per real dimension, PA-PSK and EQPA lose up to 0.5 dB at a capacity of $C = 5.5$ bit/use. If we increase the constellation size, PM and kMC still remain on top, EQPA reduces the loss to about 0.3 dB, whereas PA-PSK loses up to 2 dB compared to PM at $C = 8.5$ bit/use. The reason for this big loss when increasing the constellation size is that PA-PSK is the only constellation where no form of global optimization, i.e., over all 6 dimensions, takes place. While a PSK constellation might be perfect on an individual antenna, increasing the total constellation size requires to use different constellations on each antenna. This can have devastating influence on the overall distance properties of the constellation. On the other hand, EQPA works in such a way that the distribution of points becomes more and more uniform as the constellation size increases. This algorithm profits from packing the hypersphere more densely.

Fig. 1 also shows that for coded transmission with a target rate well below R_m, the larger constellation can get very close to the AWGN capacity: Using $M = 512$ points per constellation and assuming a code of rate $R_c = \frac{1}{2}$, i.e., a total rate of 4.5 bit, the gap to AWGN capacity is only approx. 0.1 dB. The capacities fit nicely to a coded modulation rule of thumb that 0.5 bit redundancy per real dimension for coding [10] and another 0.5 bit redundancy for shaping [11] are sufficient to close most of the gap to capacity.

C. Distance Properties

In order to compare the distance properties of the individual algorithms, Table I lists the *minimum distance* and *average neighbor distance* of a constellation defined as

$$d_{\min}(\mathcal{A}) = \min_{\substack{\mathbf{a}_i, \mathbf{a}_j \in \mathcal{A} \\ i \neq j}} ||\mathbf{a}_i - \mathbf{a}_j|| \tag{6}$$

and

$$d_{\mathrm{nb,avg}}(\mathcal{A}) = \frac{1}{M} \sum_{i=0}^{M-1} \min_{\substack{\mathbf{a}_j \in \mathcal{A} \\ \mathbf{a}_i \neq \mathbf{a}_j}} ||\mathbf{a}_i - \mathbf{a}_j||. \tag{7}$$

TABLE I
AVERAGE AND TOTAL MINIMUM DISTANCE OF CONSTELLATIONS

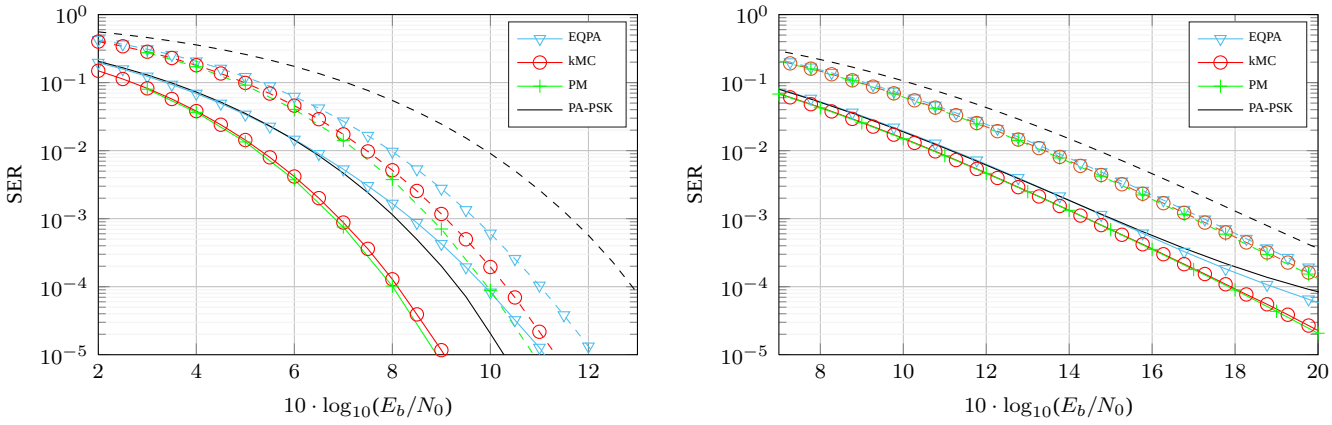
	$d_{\min}$	$d_{\mathrm{nb,avg}}$
EQPA, $M = 64$	0.6611	0.7282
kMC, $M = 64$	0.8674	0.9207
PM, $M = 64$	0.9139	0.9474
PA-PSK, $M = 64$	0.8165	0.8165
EQPA, $M = 512$	0.4654	0.5350
kMC, $M = 512$	0.5235	0.5767
PM, $M = 512$	0.5894	0.6217
PA-PSK, $M = 512$	0.4419	0.4419

PSKH constellations, especially the ones which were generated numerically, often have asymmetric distance profiles: In conventional modulation schemes like PSK or QAM, every point has at least one neighbor which is the minimum distance apart. A lot of points (in PSK even all points) have the same distance profile to neighboring points. In PSKH, there may be only two neighboring points which are the minimum distance apart, whereas all other points in the constellation only have neighbors which are further apart. This is the reason why we also include the average neighbor distance $d_{nb,avg}$. The results in Table I show for example, that EQPA and PA-PSK differ only slightly in terms of minimum distance for $M = 512$. Nevertheless, their capacities are quite different. The qualitative result of this is that the distance profile has an effect on the capacity, but it is not possible to estimate capacities from these values alone. Various PSKH constellations with similar minimum distance might have significantly different overall distance profiles, resulting in varying capacities and performances.

D. Power Efficiency

In order to elaborate how the distance profile effects the power efficiency of PSKH constellations, Fig. 2 shows the symbol error rate (SER) for constellations when 3 antennas are used with constellation sizes $M = 64$ and $M = 512$. Fig. 2a shows transmission over a vector AWGN channel, i.e., $\mathbf{H} = \mathbf{I}$. This corresponds to the discussion of the capacity on a channel with unitary channel matrix in Sec. III-B. For Fig. 2b we use the standard Rayleigh fading model, i.e., every element of $\mathbf{H}$ is a complex i.i.d. Gaussian random variable with unit variance and we average over several thousand realizations. In both cases, one can observe that the error performance of the constellation is ordered according to the minimum distance of the constellation (like the capacity), but due to the size of a constellation and asymmetry of the distance profiles (see last section), a simple quantitative estimation is not developed yet.

Remark: We note that PA-PSK with $M = 64$ for 3 antennas is regular 4-PSK on each antenna. Without further modification, an optimized constellation such as PM can give a substantial gain compared to 4-PSK/antenna of almost 1.5 dB on the vector AWGN channel. Such a channel is equivalent to subsequent transmissions over a regular AWGN channel. Constellations which achieve a coding gain by being spread over subsequent transmissions on a Single-Input Single-Output



(a) Vector AWGN (b) Rayleigh Fading

Fig. 2. Symbol error rate (SER) for transmission of $R_m = 6$ (solid) or $R_m = 9$ (dashed) bits per constellation point over a vector AWGN channel (a), and a Rayleigh fading channel (b). The system has $n = 3$ antennas.

(SISO) channel are also known as *multidimensional constellations* [12]. Multidimensional constellations can be used to combat fading [13], [14], to exploit four available dimensions in optical communications [15], [16] and to introduce a more flexible trade-off between bandwidth the power efficiency of trellis-coded signals [17]. Such constellations are usually based on conventional 2-D modulation schemes or on some lattices, which generally does not result in fixed radius constellations. To our knowledge, no work has dealt with multidimensional constellations with fixed radius in MIMO systems to exploit load-modulation transmitters.

IV. Sinc^2 Pulse Shaping

The simplest method to reduce the PASPR is to use pulse shaping filters which show better PASPR properties. PAM employing a $\mathrm{sinc}^2(t)$-function[2] for pulse shaping shows very good properties even for very few antennas (see Sec. VI). This means that the continuous transmission signal is

$$\mathbf{s}(t) = \sum_{k=-\infty}^{\infty} \mathbf{x}[k] \, \mathrm{sinc}^2\left(\frac{t - kT}{T}\right). \tag{8}$$

Since sinc^2 is not a $\sqrt{\mathrm{Nyquist}}$-function, some ISI has to be equalized at the receiver. This ISI is not generated by the channel, but only by the pulse shaping filter and its corresponding matched filter, i.e., there is no ISI between different receive antennas. Thus there is no need to make use of equalization techniques developed for MIMO ISI channels. Instead, we filter the received signal of each antenna using Forney's *Whitened Matched Filter* (WMF) [18] to get a minimum phase impulse. ISI can then be expressed by a one-dimensional, causal minimum-phase impulse $h_W[i]$ and the resulting discrete time transmission model becomes

$$\mathbf{y}[k] = \mathbf{H} \sum_{i=0}^{L} h_W[i] \mathbf{x}[k - i] + \mathbf{n}[k] \tag{9}$$

with ISI-length L. ISI can be equalized with Maximum Likelihood Sequence Estimation (MLSE) using a vector-valued

[2]We define $\mathrm{sinc}(x) = \sin(\pi x)/(\pi x)$ for $x \neq 0$ and 1 otherwise.

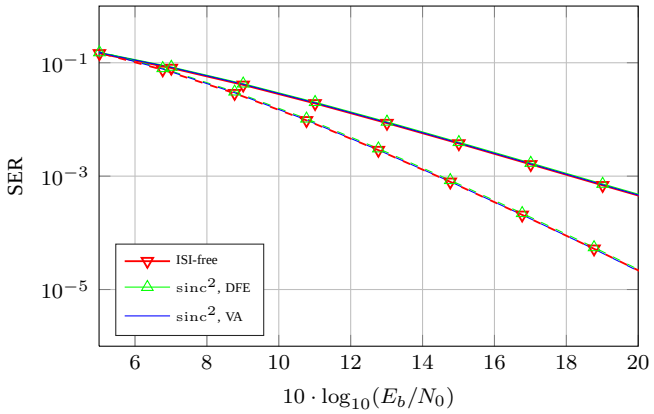


Fig. 3. Comparison of a ISI-free tranmission using a $\sqrt{\mathrm{Nyquist}}$ impulse and sinc^2 pulse shaping. Transmission is over $n = 2$ (solid) and $n = 3$ (dashed) antennas with one bit per real dimension ($M = 16$ and $M = 64$). The VA uses $\nu = 2$ memory elements. Simulations were performed over several thousand Rayleigh fading channels and averaged.

Viterbi Algorithm (VA) [19], Decision Feedback Equalization (DFE) or Delayed Decision Feedback Sequence Estimation (DDFSE) [20], which allows a performance trade-off between DFE and MLSE. In this specific case, almost all energy of h_W is stored in the very first coefficient $h_W[0]$, such that there is only a minimal loss in terms of error probability when using DFE. Results of numerical simulations can be found in Fig. 3. There is virtually no loss between sinc^2 pulse shaping with DFE (the simplest equalization method in this scenario) and ISI-free transmission by means of a $\sqrt{\mathrm{Nyquist}}$ pulse shaping in terms of power efficiency.

V. Spherical Interpolation Signaling

Spherical Interpolation (SI) signaling tries to smoothen the transmission signal by forcing it onto the hypersphere also in-between data samples. This is achieved by inserting interpolation points at a certain oversampling rate. The positive effect is a significantly reduced PASPR compared to conventional PAM because the signal becomes smoother and deviations from the hypersphere are reduced, especially zero-crossings.

The disadvantage is ISI introduced by the interpolation points and thus an increased receiver complexity.

Before presenting two different approaches, we define spherical interpolation also known as SLERP (SphericaL intERPolation) [21]: Given two points $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{R}^N$ with $||\mathbf{x}_1|| = ||\mathbf{x}_2|| = 1$ and $\cos(\theta) = \langle \mathbf{x}_1, \mathbf{x}_2 \rangle$, for any $0 \leq \tau \leq 1$, the spherical interpolation of $\mathbf{x}_1$ and $\mathbf{x}_2$ is given as

$$\mathbf{SI}(\mathbf{x}_1, \mathbf{x}_2, \tau) = \frac{\sin\left((1-\tau)\theta\right)}{\sin\theta}\mathbf{x}_1 + \frac{\sin\left(\tau\theta\right)}{\sin\theta}\mathbf{x}_2. \tag{10}$$

A. $\frac{T}{2}$-Pulse Shaping

Generating a signal using spherical interpolation values in-between T-spaced data symbols introduces ISI. In order to simplify equalization, our first approach is to use a $\sqrt{\text{Nyquist}}$-filter with respect to $T/2$, i.e., a pulse shaping filter $h(2t)$. The corresponding matched filter is $h^\star(-2t)$. In-between data symbols, the interpolation of two adjacent points is transmitted. This way, no two transmitted symbols are opposite of the hypersphere and hence the PASPR is reduced. The resulting output signal of the transmitter is

$$\begin{aligned}\mathbf{s}(t) = \sum_{k=-\infty}^{\infty} \Bigg(& \mathbf{x}[k]\, h\left(2(t-kT)\right) \\ & + \mathbf{SI}\left(\mathbf{x}[k], \mathbf{x}[k+1], \frac{1}{2}\right) h\left(2\left(t - \left(k+\frac{1}{2}\right)T\right)\right)\Bigg)\end{aligned} \tag{11}$$

which is $\sqrt{\text{Nyquist}}$ with respect to half the symbol rate. Filtering with the corresponding matched filter and sampling with a rate of $\frac{T}{2}$ at the receiver gives a sequence consisting of alternating data points and interpolation values. Because we chose a filter which is $\sqrt{\text{Nyquist}}$ with respect to $\frac{T}{2}$, every sample at the receiver is ISI-free.

To keep this system comparable with conventional PAM, both data and interpolation symbols contain only half the original symbol energy. Therefore it is necessary to use all points at the receiver to estimate the data sequence, otherwise half of the energy would be wasted. Data estimation for $\frac{T}{2}$-pulse shaping is done via the Viterbi algorithm. Due to the interpolation, every metric in the receiver depends on current and previous received value, i.e., the VA requires exactly $\nu = 1$ memory element.

It is obvious that this method has a huge disadvantage due to occupying twice the spectrum. The reason why we nevertheless include it in this comparison is that $\frac{T}{2}$-pulse shaping increases the slope of the error curve such that in the medium- to high-SNR regime it might still be a valid alternative given the largely reduced PASPR compared to conventional PAM. Results for this pulse shaping method are plotted in Fig. 4.

The increased slope can be explained by the linear transformation of the hypersphere induced by $\mathbf{H}$: At high SNRs, symbol errors will usually occur because the noise moves the data symbol into the decision region of a neighboring symbol. Symbol errors to the opposite side of the hypersphere occur only rarely, because such points are farthest apart. If the receiver constellation $\mathbf{H}\mathcal{A} = \{\mathbf{Hx} \mid \mathbf{x} \in \mathcal{A}\}$ is distorted enough, such errors may be much more likely because the distance between opposing points might be drastically reduced. In the case of $\frac{T}{2}$-pulse shaping, not only the distance between constellation points, but also the distance between interpolation points affects the performance of the system. The spherical interpolation thus corresponds to a nonlinear convolutional code. For data points on opposing sides of the hypersphere, $\frac{T}{2}$-pulse shaping generates interpolation points which are usually far away from each other. This increases the total minimum distance and thus the performance of the transmission. The magnitude of this effect is dependent on $\mathbf{H}$. A good measure for this effect is the ratio $\frac{\sigma_{\text{SVD,max}}}{\sigma_{\text{SVD,min}}}$ with $\sigma_{\text{SVD,max}}$ and $\sigma_{\text{SVD,min}}$ being the maximum and minimum singular values of the real representation of $\mathbf{H}$, respectively[3]. The two extreme cases would be $\frac{\sigma_{\text{SVD,max}}}{\sigma_{\text{SVD,min}}} = 1$ in which $\mathbf{H}\mathcal{A}$ would still be a hypersphere (possibly with a different radius) and $\frac{\sigma_{\text{SVD,max}}}{\sigma_{\text{SVD,min}}} \to \infty$. In the latter case, the n-dimensional hypersphere would be compressed down to fewer dimensions, effectively reducing the distance between opposing points and possibly making them nearest neighbors.

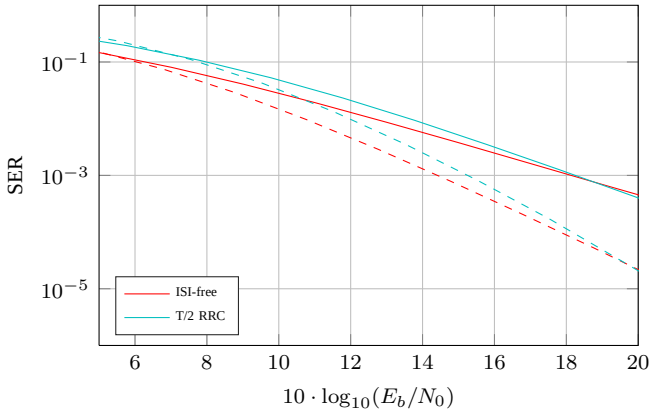


Fig. 4. Comparison of conventional ISI-free PAM transmission and $\frac{T}{2}$ pulse shaping. Transmission is over $n = 2$ (solid) and $n = 3$ (dashed) antennas with one bit per real dimension ($M = 16$ and $M = 64$). Simulations were performed over several thousand Rayleigh fading channels and averaged.

B. Spherical Interpolation

The main problem of $\frac{T}{2}$-pulse shaping is the large bandwidth due to the use of pulse shaping filters at higher frequency. In order to mitigate the problem, we combine SI with a conventional pulse shaping filter being $\sqrt{\text{Nyquist}}$ with respect to T.

This method is characterized by the interpolation frequency $f_{\text{IP}} \in \mathbb{N}$: In each symbol interval, the original data point as well as $f_{\text{IP}} - 1$ interpolation points are transmitted. $f_{\text{IP}} = 1$ corresponds to conventional PAM without SI. The resulting output signal of the transmitter is

[3]Every complex-valued model $\mathbf{y} = \mathbf{Hx} + \mathbf{n}$ of dimension n can be transformed into an equivalent real-valued model of dimension $2n$, see e.g. [22].

$$\mathbf{s}(t) = \sum_{k=-\infty}^{\infty} \left(\mathbf{x}[k]\, h(t-kT) + \sum_{l=1}^{f_{\mathrm{IP}}-1} \mathbf{SI}\left(\mathbf{x}[k], \mathbf{x}[k+1], \frac{l}{f_{\mathrm{IP}}}\right) h\left(t - \left(k + \frac{l}{f_{\mathrm{IP}}}\right) T\right)\right). \tag{12}$$

At the receiver, matched filtering with $h^*(-t)$ is performed followed by T-spaced sampling. This is slightly suboptimal, but simplifies the receiver structure greatly. Introducing the autocorrelation

$$\varphi_{hh}(\tau) = \int_{-\infty}^{\infty} h(t+\tau) h^{\star}(t) \mathrm{d}t, \tag{13}$$

the received discrete-time signal is

$$\mathbf{y}[k] = \mathbf{y}(kT) = \mathbf{H} \left(\sum_{\bar{k}=-\infty}^{\infty} \mathbf{x}[\bar{k}] \varphi_{hh}(kT - \bar{k}T) + \sum_{l=1}^{f_{\mathrm{IP}}-1} \mathbf{SI}\left(\mathbf{x}[\bar{k}], \mathbf{x}[\bar{k}+1], \frac{l}{f_{\mathrm{IP}}}\right) \cdot \varphi_{hh}\left(kT - \left(\bar{k} + \frac{l}{f_{\mathrm{IP}}}\right) T\right)\right) + \mathbf{n}[k]. \tag{14}$$

By using a $\sqrt{\text{Nyquist}}$-pulse and T-spaced sampling, the direct influence of adjacent data symbols may be suppressed and the resulting noise at the receiver is white, but the influence of the interpolation values inserted at rate $T_{\mathrm{IP}} = T/f_{\mathrm{IP}}$ remains. Thus ISI-equalization in form of MLSE via the VA has to be performed at the receiver. Fig. 5 shows the system model used at the receiver to estimate the data sequence. The VA has to consider all f_{IP} vectors, which were transmitted during one symbol interval, to calculate a branch metric. Since T-spaced sampling is used, it is vital to use the contribution from the interpolation values, otherwise serious ISI would be unprocessed and its energy would be wasted.

For SI pulse shaping, the choice of the filter does have an influence on the error probability, because it determines the shape of φ_{hh}. For the following results, we used a root-raised cosine (RRC) pulse shaping filter with roll-off factor $\beta = 0.25$. Additionally, the choice of f_{IP} provides a trade-off between receiver complexity and smoothness of the output signal (which in term improves PASPR and bandwidth, see Sec. VI). The results in Figs. 9 and 8 were generated using $f_{\mathrm{IP}} = 4$. Increasing f_{IP} to 16 showed a 0.15 dB improvement in PASPR, whereas the error probability is unaffected by increasing f_{IP}.

In order to achieve full ML detection, all coefficients of the impulse response have to be considered and MLSE can be performed using the VA. The system model for $\nu = 3$ memory elements is depicted in Fig. 5: Two adjacent symbols in time generate SI data, which is weighted according to φ_{hh} and summed up. Given the size of PSKH constellations, full ML detection is computationally impossible. Thus we need to apply various complexity reduction methods, which are described in the next subsection.

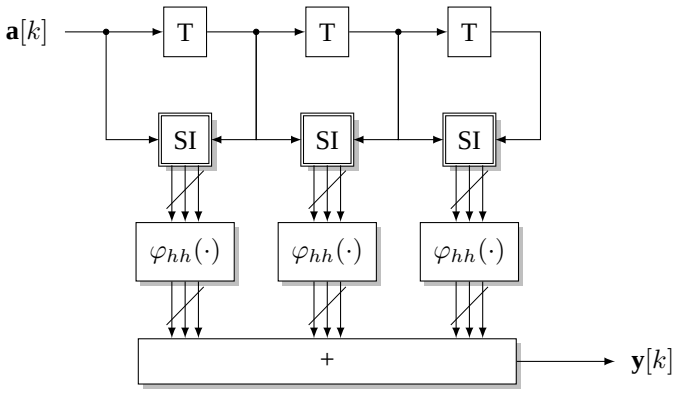


Fig. 5. System used to model the ISI produced by spherical interpolation transmission. An SI block calculates f_{IP} vectors and $\varphi_{hh}(\cdot)$ weighs them with the autocorrelation of the pulse shaping filter. Thus each block processes all interpolation vectors within one symbol period. This model omits the channel matrix $\mathbf{H}$ and noise $\mathbf{n}$. This example employs $\nu = 3$ memory elements.

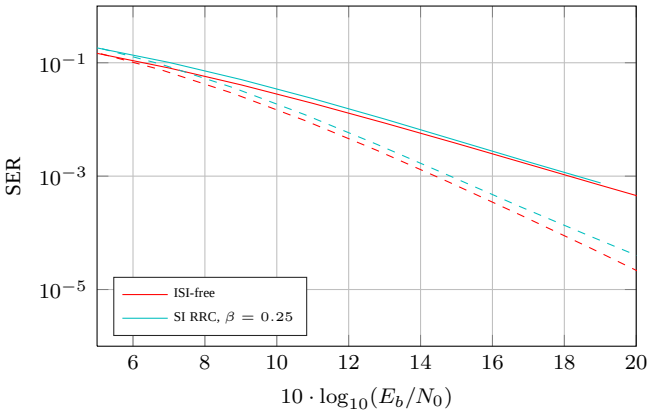


Fig. 6. Comparison of ISI-free transmission and SI pulse shaping with $f_{\mathrm{IP}} = 4$. Transmission is over $n = 2$ (solid) and $n = 3$ (dashed) antennas with one bit per real dimension ($M = 16$ and $M = 64$). Detection was performed using DDFSE employing $\nu = 2$ memory elements for $n = 3$ antennas and $\nu = 3$ memory elements for $n = 2$ antennas. Simulations are averaged over several thousand Rayleigh fading channels.

C. Complexity Reduction Techniques

In order to make a detector computationally feasible, we only use those intervals of φ_{hh} which have the highest energy and use it for sequence estimation in the VA. Prior taps are considered as noise and remaining taps at the end are equalized using DFE, which makes the overall scheme a Delayed Decision-Feedback Sequence Estimation (DDFSE) [20]. Fig. 6 shows numerical results for the spherical interpolation shaping using $f_{\mathrm{IP}} = 4$. Detection was performed using DDFSE employing $\nu = 3$ memory elements for the 16-ary constellation and $\nu = 2$ memory elements for the 64-ary constellation (corresponding to 4096 states in both cases).

For DDFSE, usually a prefilter is applied to make the overall impulse response minimum-phase, see e.g. [23]. In this case, we have to deal with a number of problems: First of all, SI is a nonlinear operation. Secondly, the overall impulse response has a large linear phase portion. Thirdly, for a given interval width, the minimum phase response captures less of the total

energy than the overall impulse response: The original impulse response has a large fraction of its total energy concentrated around the center, whereas the minimum phase part spreads its energy over a wider interval in the beginning of the impulse. Thus, more memory elements in the VA are required to capture the same amount of energy. This is computationally not feasible. Fourthly, given the total length of φ_{hh}, numerical inaccuracies might occur when calculating the prefilter. It is thus advantageous to use the original filter instead of applying any prefilter to make the overall filter minimum phase and simply treat the influence of the first taps as noise.

The power efficiency of the SI pulse compared to conventional PAM employing a RRC filter depends on the roll-off factor and the number of states in the VA: For a fixed number of states in the receiver, increasing the roll-off factor β of the RRC filter improves power efficiency, because more of the energy of $\varphi_{hh}(t)$ is concentrated around the center. This is in contrast to conventional PAM, where power efficiency is unaffected by the roll-off factor. Our analyses show that for $\beta = 0.1$, a loss of approx. 1.1 dB occurs at a target symbol error rate of 10^{-4}. This loss shrinks quickly with an increasing roll-off factor: For $\beta = 0.25$, the gap is closed. Increasing the number of memory elements also reduces the loss, because the amount of energy used for sequence estimation is increased. The feasibility of this is restricted by computational complexity. Therefore, we now discuss how the gap between SI RRC and conventional RRC can be closed with reduced computational complexity.

It is usually sufficient to use only 2 or 3 delay elements to capture almost all energy of the pulse. The remaining energy at the end of the filter can be equalized by means of DDFSE. But since every delay element is M-ary, the number of states can become infeasible even for such a small number of elements. We thus compare system performance and complexity when using two different complexity reduction techniques: The well-known Reduced State Sequence Estimation (RSSE) [24] and a newly proposed iterative application of the VA.

For RSSE, we use a Viterbi algorithm with $\nu = 2$ memory elements and generate hyperstates by using hypersymbols in the second delay element only. Combining different input symbols into a hypersymbol in the first element leads to large performance degradation due to two effects: Hyperstates are calculated in advance based on the original constellation $\mathcal{A}$. We did this by numerically optimizing the minimum distance within each hyperstate [25]. The effective constellation at the receiver, however, is $\mathbf{H}\mathcal{A}$ which might have a drastically different distance profile than the constellation with originally optimal hyperstates. The other negative impact is the fact that both the impulse response $\varphi_{hh}(t)$ as well as its minimum-phase component do not have monotonously decreasing values. The decision in the first delay element is thus based on only a small fraction of the total pulse energy.

Our second approach to reduce the complexity is to apply the Viterbi algorithm iteratively. This works well if the performance gap between the use of ν and $\nu+1$ memory elements is not too large. The idea behind it is that if each error pattern is a neighboring symbol of the correct signal point, it is sufficient to consider these neighboring symbols in future steps. This is a valid assumption if the SNR is sufficiently high. Our algorithm works as follows:

1) Initialize $\nu = 1$.
2) Run the Viterbi algorithm with one memory element.
3) For each estimate $\hat{\mathbf{x}}[k]$, find the n_{NB} nearest neighboring points.
4) Set $\nu = \nu + 1$.
5) Run the Viterbi algorithm with ν memory elements, only allowing $n_{\mathrm{NB}} + 1$ points in each time step.
6) If $\nu = \nu_{\max}$ finish, otherwise go back to 3.

The neighboring points can be calculated in advance and stored in a table. This works best if the neighboring points are taken from $\mathbf{H}\mathcal{A}$, but a reasonable performance can also be achieved if they are taken directly from $\mathcal{A}$, which reduces the overhead to recalculate them every time $\mathbf{H}$ changes.

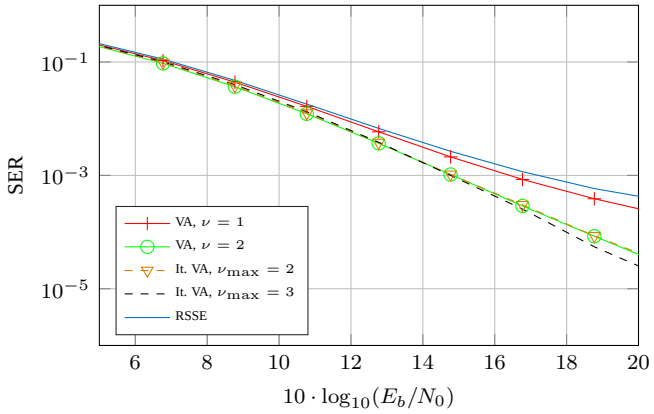


Fig. 7. Comparison of SI pulse shaping for $n = 3$ antennas with a RRC pulse shape with $\beta = 0.25$. RSSE used a quaternary second delay element and $n_{\mathrm{NB}} = 4$ for all iterative variants. Simulations were performed over several thousand Rayleigh fading channels and averaged.

Fig. 7 shows the performance of a 64-ary alphabet transmitted via SI signaling with a RRC pulse with $\beta = 0.25$ employing $n = 3$ antennas. The VA curve using two memory elements ($\nu = 2$) is the same as in Fig. 6 and is our baseline. As a measure of complexity we count the number of trellis branches in each time step

$$\Xi = \begin{cases} M^{\nu+1}, & \text{Standard VA} \\ M \cdot \prod_{i=1}^{\nu} M_i, & \text{VA, RSSE} \\ M^2 + \sum_{i=2}^{\nu_{\max}} (n_{\mathrm{NB,i}} + 1)^{\nu_i + 1}, & \text{Iterative VA.} \end{cases} \quad (15)$$

In this term, M_i is the number of possible values in the i-th delay element if RSSE is used (the number of hypersymbols) and ν_{i} is the number of memory elements in the i-th iteration of the iterative VA. Table II shows the complexity for the algorithms used to create Fig. 7. The computational complexity for a VA with $\nu = 2$ is already impractical. The iterative VA, however, provides the same performance as the VA with $\nu = 2$ with only minor complexity increase compared to the VA with $\nu = 1$. Increasing the number of iterations by one allows to improve the power efficiency (approx. 0.5 dB at $\mathrm{SER} = 10^{-4}$) such that the iterative VA outperforms the VA with two delay elements. The exact results for the iterative VA depend on the shape of the overall impulse response which changes with the

TABLE II
COMPLEXITY COMPARISON OF SI DEMODULATION (M = 64) FOR FIG. 7

Algorithm	Ξ
Viterbi, $\nu = 1$	4096
Viterbi, $\nu = 2$	262144
Viterbi, RSSE	16384
It. Viterbi, $\nu_{\max} = 2$	4221
It. Viterbi, $\nu_{\max} = 3$	4846

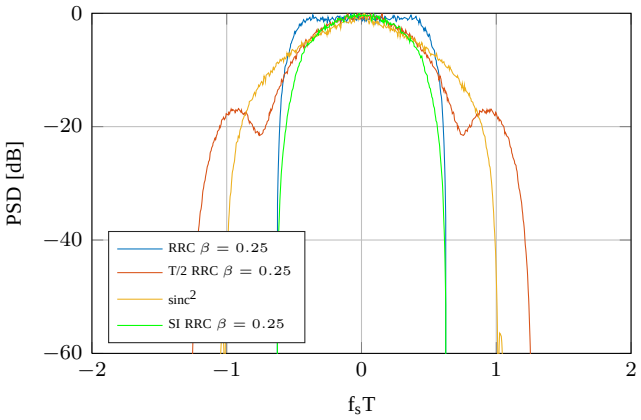


Fig. 8. Occupied spectra of different pulse shaping methods. The spectra are calculated for a $n = 4$ antenna system.

roll-off factor. For practical values $\beta > 0.2$, we found the differences to be only marginal.

VI. PASPR, SPECTRUM AND BANDWIDTH EFFICIENCY

In the previous sections, we introduced several methods to reduce the PASPR of a signal and discussed their power efficiency. Some methods may have a negative impact on the bandwidth, but a wider spectrum may be tolerable, if the gain in PASPR is substantial. In this section we discuss how much PASPR reduction can be achieved and how the corresponding spectrum behaves.

Our baseline is a RRC pulse shaping filter with roll-off factor $\beta = 0.25$. The comparison of bandwidth and PASPR is given in Figs. 8 and 9, respectively. In these plots, RRC and sinc^2 describe conventional pulse shaping (see Sec. IV for sinc^2 pulse shaping), T/2-RRC describes pulse shaping at twice the symbol rate and SI based pulse shaping is named SI RRC (see Sec. V).

A simple conclusion is that PAPSR reduction can be traded for bandwidth, i.e., the widest spectrum produces the lowest PASPR: T/2-RRC has the lowest PASPR, followed by sinc^2 pulse shaping and SI RRC has the least PASPR reduction, but it also is the only method which does not widen the spectrum. One should also take into account the receiver complexity for these techniques: sinc^2 requires almost no additional complexity compared to ISI-free PAM, T/2-RRC and SI RRC require a sequence estimation to achieve a reasonable power efficiency. For all RRC based methods, the well-known trade-off between bandwidth and roll-off still holds. Additionally, increasing β also improves the PASPR slightly.

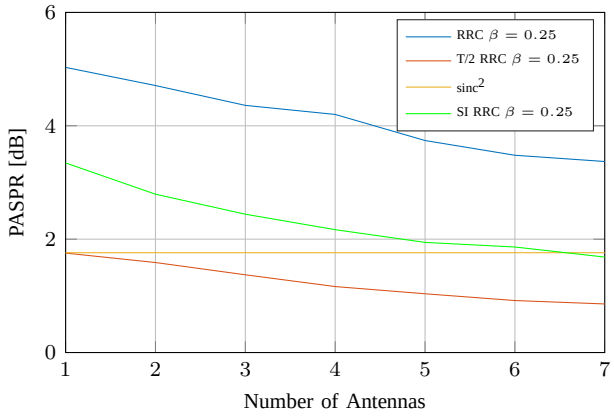


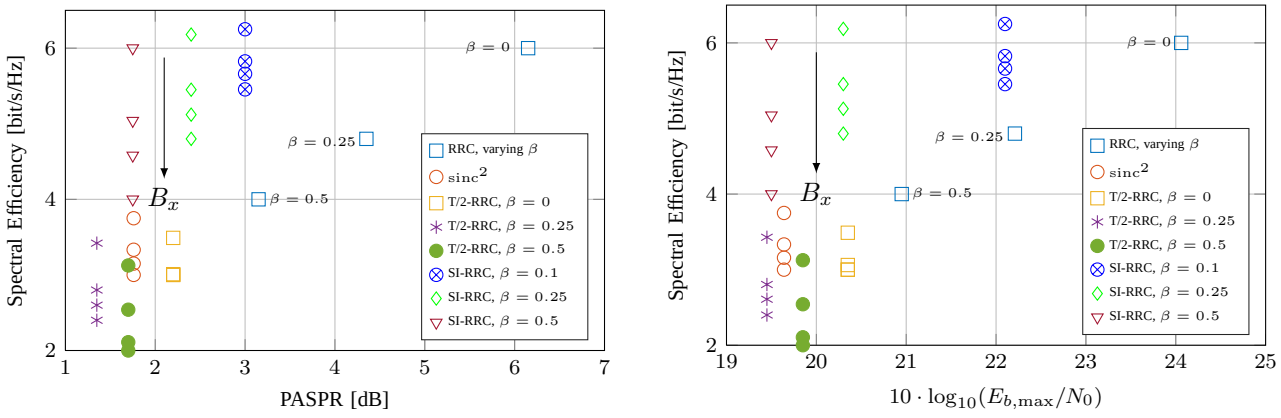
Fig. 9. Resulting PASPRs of different pulse shaping methods. A modulation rate of 1 bit per real dimension and an interpolation frequency $f_{\mathrm{IP}} = 4$ for SI RRC were used.

To summarize the results, we compare PASPR as well as power and spectral efficiencies of the methods presented in this paper. Since some pulse shaping methods have wide spectra, we also consider the bandwidth B_x which includes a fraction x of the total signal energy, e.g., $B_{99\%}$ is the bandwidth which holds 99% of the total energy of a signal.

In Fig. 10, the spectral efficiencies are plotted for a transmission system employing $n = 3$ antennas and a constellation size of $M = 64$ over a Rayleigh fading channel. As a baseline, a RRC with $\beta = \{0, 0.25, 0.5\}$ is used. For all other pulse shaping methods, we plot the spectral efficiencies for B_x with $x \in \{99\%, 99.9\%, 99.99\%, 100\%\}$. Fig. 10a plots the spectral efficiencies over the PASPR, whereas Fig. 10b uses the maximum energy per bit over the noise spectral density, i.e., $E_{b,\max}/N_0 = E_b \cdot \mathrm{PASPR}/N_0$. This takes both the maximum instantaneous power and the power efficiency for a given error rate into account and thus provides a fair comparison. These results are especially applicable for load-modulated transmitters.

The general result can be summarized as follows: All methods presented in this paper provide reasonable reduction of the PASPR. Some methods, however, do so at the cost of reduced spectral efficiency. SI pulse shaping is the only method to reduce the PASPR without sacrificing spectral efficiency. The cost to be paid in this case is a more complex receiver architecture.

In Fig. 10b, the power-bandwidth plane for a target symbol error rate of $\mathrm{SER} = 10^{-4}$ is shown. Because all losses in terms of power efficiency for a given symbol error rate are minor (if existing), especially SI RRC is superior to the conventional PAM. The gain due to the reduced PASPR generally outweighs the loss due to suboptimal detection of ISI if $\beta > 0.1$. Depending on the roll-off factor, the final gain in $E_{b,\max}/N_0$ is in the range of 1 to 2 dB. Substantial gain can also be achieved by T/2-RRC and sinc^2 pulse shaping. Since these variants have almost no loss in power efficiency, they can realize their whole PASPR gain. The downside of them, again, is the reduced spectral efficiency.



(a) Plot over PASPR

(b) Plot over $E_{b,\max}/N_0 = E_b \cdot \mathrm{PASPR}/N_0$

Fig. 10. Spectral efficiencies for different pulse shaping averaged over many Rayleigh fading channels with $n = 3$ antennas and a constellation size of $M = 64$. For all methods, except RRC, we plot the spectral efficiency based on the bandwidth B_x for a fraction x of the total energy with $x \in \{99\%, 99.9\%, 99.99\%, 100\%\}$. A target symbol error rate of SER $= 10^{-4}$ was used for the power bandwidth plane in Fig. 10b. Simulations were performed over several thousand Rayleigh fading channels and averaged.

VII. Conclusion

PSKH is a novel modulation scheme for MIMO that is applicable in various scenarios: Because PSKH constellations are optimized over all antennas, both the constellation-constrained capacity as well as the error rate for a given power are improved compared to conventional PSK. This shows that PSKH is an interesting alternative even for MIMO systems which do not employ load modulation.

If PSKH is combined with load-modulation amplifiers, additional improvements are possible. The distribution on the hypersphere can be exploited to achieve a transmit signal with a low PASPR. By reducing the PASPR, amplifiers can be driven at a higher efficiency and thus the power loss is reduced. To achieve this, there is a trade-off between three degrees of freedom: Power efficiency, bandwidth efficiency and receiver complexity. It is possible to improve power efficiency at the cost of either bandwidth efficiency or receiver complexity. These results underline that load-modulation transmitters are a valid alternative for power-efficient communications of MIMO systems, which only employ a small number of antennas.

References

[1] F. H. Raab, P. Asbeck, S. Cripps, P. B. Kenington, Z. B. Popovic, N. Pothecary, J. F. Sevic, and N. O. Sokal, "RF and Microwave Power Amplifier and Transmitter Technologies - Part 1," *High Frequency Electronics*, vol. 2, no. 3, pp. 22–36, 2003.

[2] M. A. Sedaghat, R. R. Müller, and G. Fischer, "A Novel Single-RF Transmitter for Massive MIMO," in *Proc. ITG Workshop on Smart Antennas*, 2015.

[3] R. R. Müller, M. A. Sedaghat, and G. Fischer, "Load Modulated Massive MIMO," in *Proc. of IEEE Global Conference ond Signal and Information Processing*, 2014.

[4] M. A. Sedaghat, R. R. Müller, and C. Rachinger, "(Continuous) Phase Modulation on the Hypersphere," *IEEE Transactions on Wireless Communications*, vol. PP, no. 16, May 2016.

[5] J. H. Conway and N. J. A. Sloane, *Sphere Packing, Lattices and Groups*. Springer, 1998.

[6] N. J. A. Sloane, R. H. Hardin, W. D. Smith *et al.* Tables of Spherical Codes. [Online]. Available: http://neilsloane.com/packings/

[7] P. Leopardi, "A partition of the unit sphere into regions of equal area and small diameter," *Electronic Transactions on Numerical Analysis*, vol. 25, pp. 309–327, 2006.

[8] I. S. Dhillon and D. S. Modha, "Concept Decompositions for Large Sparse Text Data Using Clustering," *Machine Learning*, vol. 42, no. 1, pp. 143–175, January 2001.

[9] J. M. Haile, *Molecular Dynamics Simulation: Elementary Methods*. Wiley, 1997.

[10] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Transactions on Information Theory*, vol. 28, no. 1, pp. 55–67, Jan 1982.

[11] U. Wachsmann, R. F. H. Fischer, and J. B. Huber, "Multilevel Codes: Theoretical Concepts and Practical Design Rules," *IEEE Transactions on Information Theory*, vol. 45, no. 5, pp. 1361–1391, July 1999.

[12] G. D. Forney and L.-F. Wei, "Multidimensional Constellations - Part I: Introduction, Figures of Merit, and Generalized Cross Constellations," *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 6, pp. 877–892, August 1989.

[13] J. Boutros and E. Viterbo, "Signal Space Diversity: A Power- and Bandwidth-Efficient Diversity Technique for the Rayleigh Fading Channel," *IEEE Transactions on Information Theory*, vol. 44, no. 4, pp. 1453–1467, July 1998.

[14] F. Oggier, E. Bayer-Fluckiger, and E. Viterbo, "New algebraic constructions of rotated cubic lattice constellations for the Rayleigh fading channel," in *Information Theory Workshop, 2003. Proceedings. 2003 IEEE*, March 2003, pp. 263–266.

[15] E. Agrell and M. Karlsson, "Power-efficient modulation formats in coherent transmission systems," *Journal of Lightwave Technology*, vol. 27, no. 22, pp. 5115–5126, Nov 2009.

[16] J. Leibrich and W. Rosenkranz, "Power efficient multidimensional constellations," in *Photonic Networks; 15. ITG Symposium; Proceedings of*, May 2014, pp. 1–6.

[17] L. F. Wei, "Rotationally invariant trellis-coded modulations with multidimensional m-psk," *IEEE Journal on Selected Areas in Communications*, vol. 7, no. 9, pp. 1281–1295, Dec 1989.

[18] G. D. Forney, "Maximum-likelihood sequence estimation of digital sequences in the presence of intersymbol interference," *IEEE Transactions on Information Theory*, vol. 18, no. 3, pp. 363–378, May 1972.

[19] A. Viterbi, "Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, April 1967.

[20] A. Duel-Hallen and C. Heegards, "Delayed Decision-Feedback Sequence Estimation," *IEEE Transactions on Communications*, vol. 37, no. 5, pp. 428–436, May 1989.

[21] K. Shoemake, "Animating Rotation with Quaternion Curves," in *Proc. 12th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH)*, San Francisco, USA, July 1985, pp. 245–254.

[22] I. E. Telatar, "Capacity of Multi-antenna Gaussian Channels," *European Transactions on Telecommunications*, vol. 10, no. 6, pp. 585–595, 1999.

[23] W. H. Gerstacker, F. Obernosterer, R. Meyer, and J. B. Huber, "An Efficient Method for Prefilter Computation for Reduced-State Equalization," in *Proc.IEEE International Symposium on Personal, Indoor and Mobile Radio Communication (PIMRC),*, London, September 2000, pp. 604–609.

[24] M. V. Eyuboglu and S. U. H. Qureshi, “Reduced-State Sequence Estimation with Set Partitioning and Decision Feedback,” *IEEE Transactions on Communications*, vol. 36, no. 1, pp. 13–20, January 1988.

[25] B. Spinnler and J. B. Huber, “Design of Hyper States for Reduced-State Sequence Estimation,” *International Journal of Electronics and Communications (AEUE)*, vol. 50, no. 1, pp. 17–26, 1996.