Boosting Low-Complexity Vector Perturbation Variants Based on Rectangular Shaping

Shanxiang Lyu, Zheng Wang, Ling Liu, Hongliang He, and Guanggang Geng

***Abstract*—We identify vector perturbation (VP) precoding as a special case of nested lattice coding where the fine/coding lattice is adopted from the channel matrices, and the coarse/shaping lattice is in the self-similar form of the fine lattice. Since we still have the freedom to optimize the shaping lattice, in this work, we design an efficient algorithm to construct a rectangular shaping lattice for VP. This design is especially well-suited to the low-complexity approximations of VP, i.e., zero forcing (ZF) and successive interference cancellation (SIC) precoders. Simulations show that rectangular shaping based ZF and SIC precoders exhibit significant performance gains compared with their self-similar shaping based counterparts.**

***Index Terms*—Vector perturbation, shaping, lattice, SIC.**

I. Introduction

Precoding technology has been widely used in Multiple Input Multiple Output (MIMO) communications to eliminate inter-channel interference while transmitting multiple data streams at the same time. Vector perturbation (VP), as a prominent nonlinear precoding technique that achieves near-capacity performance of multiuser MIMO downlink, has attracted extensive attention in the past two decades [1]–[7]. Since the disturbance vector is in a form of tiling the transmitted symbol space, a lattice structure is involved in VP, and the problem of finding the perturbation vector corresponds to solving a computationally hard problem over lattices [1], [2].

Information-theoretic studies on VP have been conducted based on lattice coding [8]–[10]. Lattice coding is reminiscent of joint error-correction-coding and modulation, whose information rate is adapted to fit into the quality of the channel. [8] shows the achievable information rates based on the modulo-lattice-additive-noise technique, while [9] derives expressions for the sum rate using the second moment of the Voronoi region of lattices. We also notice that an information theoretic rate allocation scheme of VP has been investigated in [10], although the scheme is rather sub-optimal.

Due to the high complexity in the part of sphere precoding [1], [2], VP seems not preferred in the industry, especially for the next-generation large-scale communication systems [11]. For this reason, there has been considerable interest in finding low-complexity alternatives for the sphere precoding component inside VP. Exploiting lattice decoding techniques, lattice reduction has be used to reduce the complexity of VP, i.e., an approximate solution is found by zero forcing (ZF) or successive interference cancellation (SIC) on a reduced lattice [3], [4]. A number of recent techniques proposed towards reducing the complexity of VP precoding are also formulated or analyzed based on lattice decoding. For instance, the complexity of the selective perturbation technique [5] is analyzed by counting the number of lattice points inside a sphere; and [4] relies on a "proximity factor" to show that approximate message passing can be concatenated with lattice decoding algorithms to design the perturbation vector.

While the two worlds of lattice coding and decoding are separate from each other (i.e., information theorists design and analyze codes, wireless-communications researchers optimize algorithms to shift the complexity-performance tradeoff), the problem of VP precoding may link the best of both worlds. Specifically, we find that by judiciously assigning different rates to different streams, the quality of lattice basis in VP precoding can be improved. Thus the low-complexity alternatives of sphere precoding may now enjoy satisfactory performance. In this context, we aim to revisit VP from the perspective of lattice shaping (a technique in lattice coding), and design a low-complexity scheme with improved error-rate performance. The contributions and highlights of this work are summarized as follows:

- First, we conceive VP as a a special case of nested lattice coding. To be concise, the coding lattice is adopted from the channel matrix, and the lattice employed to design the perturbation vector is a scalar multiple of the coding lattice; together the process is referred to as self-similar shaping in lattice literature. By changing the size of symbols in each dimension of the perturbation vector, the shaping lattice could be tuned based on our demands.
- Second, we present a simple and efficient algorithm that designs the parallelotopes of the shaping lattices. The method is to approximate the modulation sizes by convex optimization when the problem size is $n = 2$. For a large n, we use the divide-and-conquer principle. This design is extraordinarily suitable for the low-complexity implementations of VP (i.e., SIC and ZF), which fits into large-scale problems which attracted many in the wireless industry. In terms of effectiveness, simulations show that the new design brings us more than 10 dB's performance gain especially for the SIC precoder.

II. Preliminaries

A. Lattices

A lattice generated by n linearly independent vectors $\mathbf{b}_1, \ldots, \mathbf{b}_n \in \mathbb{R}^m$ with $m \geq n$ is represented by

$$\Lambda = \mathcal{L}(\mathbf{b}_1, \ldots, \mathbf{b}_n) = \left\{ \sum_{i=1}^{n} \mathbf{b}_i x_i \mid x_i \in \mathbb{Z} \right\}.$$

The set $[\mathbf{b}_1, \ldots, \mathbf{b}_n] \triangleq \mathbf{B}$ is called a lattice basis.

1) Fundamental Regions and Precoding: A region $\mathcal{F}$ is called a fundamental region for a lattice Λ if shifts of $\mathcal{F}$ by lattice points covers the whole real space. The volume of a fundamental region, denoted as $|\mathcal{F}|$, equals to $\sqrt{\det(\mathbf{B}^\top \mathbf{B})}$. The Voronoi cell is a polytope defined by

$$\mathcal{V}_{\mathcal{L}(\mathbf{B})} = \{\mathbf{y} \mid \|\mathbf{y}\|^2 \leq \|\mathbf{y} - \mathbf{w}\|^2, \forall \mathbf{w} \in \mathcal{L}(\mathbf{B})\}.$$

Manuscript received January 4, 2021; revised April 26, 2021; accepted May 24, 2021. Date of publication May 27, 2021; date of current version July 20, 2021. This work was supported in part by the National Natural Science Foundation of China under Grants 61902149, 61801216, 62001300, 62032009, and 61932010, in part by the Natural Science Foundation of Guangdong Province under Grant 2020A1515010393, in part by the Fundamental Research Funds for Central Universities under Grants 21620350 and 21620438, and the Natural Science Foundation of Jiangsu Province under Grant BK20180420. The review of this article was coordinated by Dr. X. Ge. *(Corresponding author: Hongliang He.)*

Shanxiang Lyu, Hongliang He, and Guanggang Geng are with the College of Cyber Security, Jinan University, Guangzhou 510632, China (e-mail: shanxianglyu@gmail.com; hehongliang@stu.xjtu.edu.cn; guanggang.geng@gmail.com).

Zheng Wang is with National Mobile Communications Research Laboratory, School of Information Science and Engineering, Southeast University, Nanjing 210096, China (e-mail: z.wang@ieee.org).

Ling Liu is with the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China (e-mail: liulingcs@szu.edu.cn).

Digital Object Identifier 10.1109/TVT.2021.3084298

It corresponds to the set of query points that are closer to the origin than any other lattice points. We can either use a Voronoi cell or a parallelotope to define a fundamental region, and each region reflects the decision region of an algorithm solving the closest vector problem (CVP). The CVP is, given a query point $\mathbf{y}$ and a lattice $\mathcal{L}(\mathbf{B})$, to find a vector $\mathbf{v} \in \mathcal{L}(\mathbf{B})$ such that:

$$\|\mathbf{y}-\mathbf{v}\|^2 \leq \|\mathbf{y}-\mathbf{w}\|^2,\ \forall \mathbf{w} \in \mathcal{L}(\mathbf{B}).$$

Since the computational complexity of exactly solving CVP is generally prohibitively high for a random lattice, the low-complexity alternatives, ZF and SIC, are often adopted at the cost of losing certain precision [3], [12]. Their performance can be characterized by the following parallelotopes: The ZF precoding based parallelotope [3] $\mathcal{P}_{ZF,\mathcal{L}(\mathbf{B})}$ is described by vectors $\mathbf{b}_1,\ldots,\mathbf{b}_n$ is

$$\mathcal{P}_{ZF,\mathcal{L}(\mathbf{B})} = \left\{ \sum_{i=1}^{n} \mathbf{b}_i a_i \mid -1/2 \leq a_i < 1/2 \right\}.$$

Define the Gram-Schmidt Orthogonalization (GSO) vectors of $\mathbf{b}_1,\ldots,\mathbf{b}_n$ as $\mathbf{b}_1^*,\ldots,\mathbf{b}_n^*$, the SIC precoding based parallelotope [3] is given by

$$\mathcal{P}_{SIC,\mathcal{L}(\mathbf{B})} = \left\{ \sum_{i=1}^{n} \mathbf{b}_i^* a_i \mid -1/2 \leq a_i < 1/2 \right\}.$$

2) Lattice Shaping: We say that two lattices Λ_c and Λ_s are nested if $\Lambda_s \subset \Lambda_c$. The bigger lattice Λ_c is called the *fine/coding* lattice, and Λ_s is called the *coarse/shaping* lattice. A typical and efficient method to build a set of lattice points for data transmission, called lattice coding, is to use nested lattices based shaping. Specifically, the lattice code $\mathcal{C}$ is defined by

$$\mathcal{C} = \Lambda_c \mod \Lambda_s = \Lambda_c \cap \mathcal{V}_{\mathcal{L}(\mathbf{B})}.$$

For any given $\mathbf{w} \in \Lambda_c$, its shaped constellation point in $\mathcal{C}$ is

$$\mathbf{w} \mod \Lambda_s = \mathbf{w} - \mathcal{Q}_{\Lambda_s}(\mathbf{w}), \tag{1}$$

where $\mathcal{Q}_{\Lambda_s}(\mathbf{w})$ denotes a sphere search algorithm [4] that exactly finds the nearest point of $\mathbf{w}$ in Λ_s. If $\Lambda_s = p\Lambda_c$, then (1) is called *self-similar shaping* [13].

B. Vector Perturbation and CVP

Consider a real-value MIMO broadcast model where the base station is equipped with m transmitting antennas to broadcast messages to n single-antenna users with $m \geq n$ [1], [3]. The observed signal vector can be formulated as

$$\mathbf{r} = \sqrt{\mathrm{SNR}/\beta}\mathbf{H}\mathbf{t} + \mathbf{w}, \tag{2}$$

where $\mathbf{t} \in \mathbb{R}^m$ is the modulated transmission of message $\mathbf{s}$, $\mathbf{w} \in \mathbb{R}^n$ is the additive noise vector with its entries admitting the standard Gaussian distribution $\mathrm{N}(0,1)$, $\mathbf{H} \in \mathbb{R}^{n\times m}$ denotes a channel matrix whose entries also admit $\mathrm{N}(0,1)$, $\beta \triangleq \mathbb{E}\|\mathbf{t}\|^2$ denotes a normalization factor, and SNR is the signal-to-noise ratio (SNR) parameter.

Based on scaling and shifting, let the constellation be $\mathcal{M}^n = \{0,\ldots,p-1\}^n$ and $p > 1$ is a positive integer [2]. With $\mathbf{s} \in \mathcal{M}^n$ and the CSI available at the base station, the unnormalized transmission $\mathbf{t}$ of VP precoding is designed as

$$\mathbf{t} = \mathbf{H}^\dagger \mathbf{s} - p\mathbf{H}^\dagger \mathbf{x}^*, \tag{3}$$

where $\mathbf{H}^\dagger$ denotes the Moore–Penrose pseudoinverse of $\mathbf{H}$, and $\mathbf{x}^* \in \mathbb{Z}^n$ is an integer vector to be optimized based on a chosen lattice decoding algorithm.

At the receivers' side we have

$$\mathbf{r} = (\mathbf{s} - p\mathbf{x})\sqrt{\mathrm{SNR}/\beta} + \mathbf{w}. \tag{4}$$

Then each user i independently performs a quantization and modulo operation to obtain its own estimation:

$$\hat{s}_i = \lfloor r_i \sqrt{\beta/\mathrm{SNR}} \rceil \mod p = \lfloor s_i + w_i \sqrt{\beta/\mathrm{SNR}} \rceil \mod p, \tag{5}$$

where $\lfloor\cdot\rceil$ denotes rounding to the nearest integer. From (5), we can see that if $|w_i\sqrt{\beta/\mathrm{SNR}}| < \frac{1}{2}$ for all i, then $\mathbf{s}$ can be faithfully recovered. Hereby the effective noise per component is $w_i\sqrt{\beta/\mathrm{SNR}} \in \mathrm{N}(0, \beta/\mathrm{SNR})$, where w_i denotes the ith component of $\mathbf{w}$.

To decrease the decoding error probability dominated by β, the optimal $\mathbf{x}^*$ in Eq. (3) can be found by solving CVP based on sphere precoding:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathbb{Z}^n} \left\|\mathbf{H}^\dagger(\mathbf{s}-p\mathbf{x})\right\|^2. \tag{6}$$

The underlying lattice shaping structure in Eq. (3) now becomes evident: the coding lattice is $\mathcal{L}(\mathbf{H}^\dagger)$, and the shaping lattice is $\mathcal{L}(p\mathbf{H}^\dagger)$. Simply put, we have

$$\mathbf{t} = \mathbf{H}^\dagger\mathbf{s} \mod \mathcal{L}(p\mathbf{H}^\dagger) = \mathbf{H}^\dagger\mathbf{s} - \mathcal{Q}_{\mathcal{L}(p\mathbf{H}^\dagger)}(\mathbf{H}^\dagger\mathbf{s}). \tag{7}$$

We observe that this process is exactly self-similar shaping. For this channel-defied lattice code, the information rate per antenna is

$$I = \frac{1}{n}\log_2 |\det(p\mathbf{H}^\dagger)|/|\det(\mathbf{H}^\dagger)| = \log_2 p.$$

III. The Proposed Vector Perturbation With Optimized Shaping

A. The Proposed Rectangular Shaping

Let $\mathbf{M}$ be a diagonal matrix with $p_i \in \mathbb{Z}$ on the diagonal

$$\mathbf{M} = \mathrm{diag}(p_1, p_2, \ldots, p_n). \tag{8}$$

We set $\det(\mathbf{M}) = p^n$ for the sake of comparing with self-similar shaping. We propose to employ a new shaping lattice $\mathcal{L}(\mathbf{H}^\dagger\mathbf{M})$ that enables changing the structure of $\mathcal{L}(\mathbf{H}^\dagger)$ to reduce the effective noise power in VP. The crux is to justify that there exists such a message space $\tilde{\mathcal{M}}^n$ that supports the proposed shaping.

Theorem 1: Define the message space $\tilde{\mathcal{M}}^n$ as the Cartesian product:

$$\tilde{\mathcal{M}}^n = \{0,1,\ldots,p_1-1\} \times \cdots \times \{0,1,\ldots,p_n-1\}.$$

Let the message $\mathbf{s} \in \tilde{\mathcal{M}}^n$ and the transmitted vector be

$$\mathbf{t} = \mathbf{H}^\dagger\mathbf{s} \mod \mathcal{L}(\mathbf{H}^\dagger\mathbf{M}) = \mathbf{H}^\dagger\mathbf{s} - \mathcal{Q}_{\mathcal{L}(\mathbf{H}^\dagger\mathbf{M})}(\mathbf{H}^\dagger\mathbf{s}), \tag{9}$$

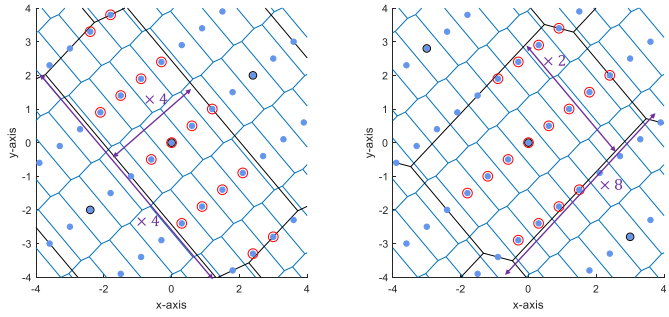
then there exists a bijective mapping between $\mathbf{s}$ and $\mathbf{t}$.

Proof: The main point is to show that the modulo lattice operation does not destroy the uniqueness of $\mathbf{s}$. Hereby we use contradiction to prove that the bijective mapping exists. Specifically, if we can find $\mathbf{s}' \in \tilde{\mathcal{M}}$, $\mathbf{s}' \neq \mathbf{s}$, such that

$$\mathbf{H}^\dagger\mathbf{s} \mod \mathcal{L}(\mathbf{H}^\dagger\mathbf{M}) = \mathbf{H}^\dagger\mathbf{s}' \mod \mathcal{L}(\mathbf{H}^\dagger\mathbf{M}),$$

it implies $\mathbf{s}$ and $\mathbf{s}'$ are two distinct coset representatives of $\{\mathbf{s} - \mathbf{M}\mathbb{Z}^n\}$. This contradicts the fact that the coset representative within $\tilde{\mathcal{M}}^n$ is unique. ■

The operation in (9) is referred to as *rectangular shaping* in lattice literature [13]. Given the CSI $\mathbf{H}$ and a specified information rate of $I = \log_2 p$, we have the freedom to optimize $p_1, p_2, \ldots, p_n$ under the constraint of $\prod_{i=1}^n p_i = p^n$. This advantage helps us to achieve a good



(a) (b)

Fig. 1. Comparing self-similar shaping and rectangular shaping.

shaping lattice even when the coding lattice is not good for shaping. The remaining task is to design $\mathbf{M}$ such that β can be smaller.

We present a 2-dimensional example for further classification. Consider a fine lattice $\mathcal{L}(\mathbf{H}^\dagger)$ with basis

$$\mathbf{H}^\dagger = \begin{bmatrix} 1.5 & 0.6 \\ -1.4 & 0.5 \end{bmatrix}$$

and a specified information rate $I = 2$. The self-similar shaping lattice $\mathcal{L}(4\mathbf{H}^\dagger)$ has the same geometric structure as $\mathcal{L}(\mathbf{H}^\dagger)$. On the contrary, assume that we have found a good shaping lattice of the form $\mathcal{L}(\mathbf{H}^\dagger \mathrm{diag}(2, 8))$. The two shaping lattices are plotted in Fig. 1, where the underlying lattice points for transmission have been marked with red circles. We can observe from the figure that $\mathcal{L}(\mathbf{H}^\dagger \mathrm{diag}(2, 8))$ is much more orthogonal than $\mathcal{L}(4\mathbf{H}^\dagger)$. Moreover, the average power of carved lattice points in rectangular shaping is smaller: we have $\mathbb{E}\|\mathbf{t}\|^2 = 6.93$ for the self-similar case, and $\mathbb{E}\|\mathbf{t}\|^2 = 5.06$ for the proposed rectangular case.

B. The Objective Function

For notation simplicity we denote $\mathbf{B} = \mathbf{H}^\dagger$, and $\mathbf{B}^* = [\mathbf{b}_1^*, \ldots, \mathbf{b}_n^*]$ be the GSO basis of $\mathbf{B}$. A straightforward approach to design the shaping lattice is to minimize the second moment $\sigma^2(\mathcal{V}_{\mathcal{L}(\mathbf{BM})})$ per symbol of the Voronoi region: $\sigma^2(\mathcal{V}_{\mathcal{L}(\mathbf{BM})})$

$$\begin{aligned} &\triangleq \frac{1}{n|\mathcal{V}_{\mathcal{L}(\mathbf{BM})}|} \int_{\mathbf{x}\in\mathcal{V}_{\mathcal{L}(\mathbf{BM})}} \|\mathbf{x}\|^2 d\mathbf{x} \\ &= \frac{1}{np^n\sqrt{\det(\mathbf{B}^\top\mathbf{B})}} \int_{\mathbf{x}\in\mathcal{V}_{\mathcal{L}(\mathbf{BM})}} \|\mathbf{x}\|^2 d\mathbf{x} \\ &= \frac{1}{np^n\sqrt{\det(\mathbf{B}^\top\mathbf{B})}} \int_{\mathbf{x}\in\mathcal{P}_{ZF,\mathcal{L}(\mathbf{BM})}} \|\mathbf{x} - \mathcal{Q}_{\mathcal{L}(\mathbf{BM})}(\mathbf{x})\|^2 d\mathbf{x} \\ &= \frac{1}{nk}\sum_{i=1}^{k} \|\mathbf{x}_i - \mathcal{Q}_{\mathcal{L}(\mathbf{BM})}(\mathbf{x}_i)\|^2. \end{aligned}$$

where $\mathbf{x}_1, \ldots, \mathbf{x}_k$ denote a set of samples taken from $\mathcal{P}_{ZF,\mathcal{L}(\mathbf{BM})}$.

Evaluating $\mathcal{Q}_{\mathcal{L}(\mathbf{BM})}(\mathbf{x}_i)$ still entails high complexity, we propose to minimize the covering radius of the SIC precoding based parallelotope $\mathcal{P}_{SIC,\mathcal{L}(\mathbf{BM})}$ based on choosing $\mathbf{M}$. Since $\mathbf{b}_1^*, \ldots, \mathbf{b}_n^*$ are mutually orthogonal, the covering radius of $\mathcal{P}_{SIC,\mathcal{L}(\mathbf{BM})}$ is formulated as

$$\rho(\mathcal{P}_{SIC,\mathcal{L}(\mathbf{BM})}) = \sqrt{\sum_{i=1}^{n} p_i^2 \|\mathbf{b}_i^*\|^2}.$$

Recall that the desired information rate is set as $\log_2(p)$, so we have a constraint of $\prod_{i=1}^n p_i = p^n$. Therefore we arrive at the objective of

$$\min_{p_1,\ldots,p_n} \sum_{i=1}^{n} p_i^2 \|\mathbf{b}_i^*\|^2 \text{ s.t.}, \prod_{i=1}^{n} p_i = p^n,\ p_i \in \mathbb{Z}^+. \tag{10}$$

The objective function in (10) is feasible for sphere precoding, SIC, ZF, and D2VP [14] (since D2VP scheme is essentially a series of SIC over 2 vectors).

C. The Proposed Algorithm to Solve (10)

1) The Case of $n = 2$: The combinatorial constraint $p_i \in \mathbb{Z}^+, \forall i$ in (10) makes the problem complicated. It is temporarily bypassed and we investigate

$$\min_{p_1,p_2} p_1^2\|\mathbf{b}_1^*\|^2 + p_2^2\|\mathbf{b}_2^*\|^2 \text{ s.t. } p_1p_2 = p^2. \tag{11}$$

for the case of $n = 2$.

By substituting $p_2 = p^2/p_1$ into $p_1^2\|\mathbf{b}_1^*\|^2 + p_2^2\|\mathbf{b}_2^*\|^2$, we obtain a function

$$f(p_2) = p^4 p_2^{-2}\|\mathbf{b}_1^*\|^2 + p_2^2\|\mathbf{b}_2^*\|^2. \tag{12}$$

If the domain of p_2 is taken from $\mathbb{R}$, then $f(p_2)$ is convex with respect to p_2. Letting

$$\frac{\partial f(p_2)}{\partial p_2} = -2p^4 p_2^{-3}\|\mathbf{b}_1^*\|^2 + 2p_2\|\mathbf{b}_2^*\|^2 = 0,$$

we obtain

$$p_2 = p\sqrt{\|\mathbf{b}_1^*\|/\|\mathbf{b}_2^*\|}. \tag{13}$$

Now we consider the discrete nature of p_2 and quantize the values in Eq. (13). The fundamental theorem of arithmetic states that p can be factorized as the product of distinct prime numbers:

$$p = \epsilon_1^{\alpha_1}\epsilon_2^{\alpha_2}\cdots\epsilon_t^{\alpha_t}.$$

It yields a table $\mathcal{T}$ consisting of factors of p. For instance, $\mathcal{T} = 1, 2, 4$ when $p = 4$. Then the estimation of p_1, p_2 can be chosen as

$$p_2' = \min_{x\in\mathcal{T}} |p \times \sqrt{\|\mathbf{b}_1^*\|/\|\mathbf{b}_2^*\|} - x|,\ p_1' = p^2/p_2. \tag{14}$$

By rephrasing Eq. (12) as a convex function about p_1, we also obtain a pair of solutions that may not be the same as p_1, p_2':

$$p_1'' = \min_{x\in\mathcal{T}} |p \times \sqrt{\|\mathbf{b}_2^*\|/\|\mathbf{b}_1^*\|} - x|,\ p_2'' = p^2/p_1. \tag{15}$$

Therefore, the final solutions decided by the following principle: if $f(p_2') \le f(p_2'')$ then we output p_1, p_2'; otherwise we output p_1, p_2''.

2) A general n: When $n \ge 2$, we can employ a "divide-and-conquer" approach to design the diagonal coefficients of $\mathbf{M}$. As shown if Fig. 2, the process of evaluating $p_1, \ldots, p_n$ consists of three steps:

Step 1) Sort $\|\mathbf{b}_1^*\|, \ldots, \|\mathbf{b}_n^*\|$ in descending order.

Step 2) Based on the sorted $\|\mathbf{b}_1^*\|, \ldots, \|\mathbf{b}_n^*\|$, couple the long vectors with short vectors in a pairwise manner. If n is odd, the $(n+1)/2$th vector becomes a standalone vector.

Step 3) Use the formula for the case of $n = 2$ to obtain coefficients for all pairs of $(\mathbf{b}_i^*, \mathbf{b}_j^*)$. If n is odd, we set $p_{(n+1)/2} = p$.

The efficiency of the designed scheme embarks upon the following fact. Since $\mathbf{b}_1^*, \ldots, \mathbf{b}_n^*$ are mutually orthogonal, our divide-and-conquer method would tune $\mathcal{P}_{SIC,\mathcal{L}(\mathbf{BM})}$ towards a hypercube. Then not only the covering radius of $\mathcal{P}_{SIC,\mathcal{L}(\mathbf{BM})}$ is smaller than that of $\mathcal{P}_{SIC,\mathcal{L}(p\mathbf{B})}$, but the gap between the largest and smallest lengths of $\mathbf{b}_i^*$ decreases. This desirable property is vividly shown in the bottom-right graph of Fig. 2. It is noteworthy that reducing $\mathcal{P}_{SIC,\mathcal{L}(\mathbf{BM})}$ also leads to a smaller $\mathcal{P}_{ZF,\mathcal{L}(\mathbf{BM})}$, as [12] has shown that the SIC parallelotope has a smaller outer radius.

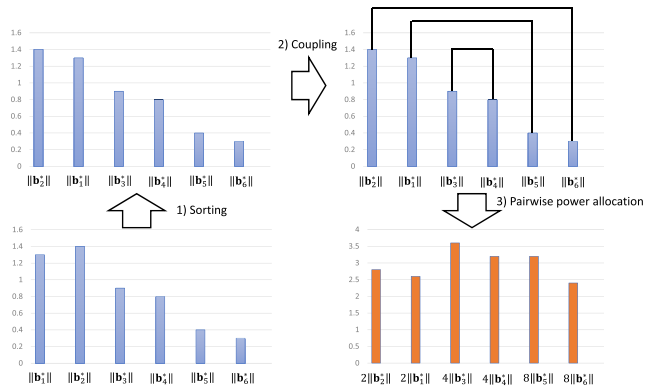


Fig. 2. The three steps to design $p_1, \ldots, p_n$.

D. Wrapping Up

To sum up, we only need to refresh a few components of the conventional VP scheme to arrive at the new design. First, based on the front-end precoder $\mathbf{H}^\dagger$, we calculate its GSO vectors $\mathbf{b}_1^*, \ldots, \mathbf{b}_n^*$. Second, employ the algorithm in Subsection to design $\mathbf{M} = \mathrm{diag}(p_1, p_2, \ldots, p_n)$. The unnormalized transmission $\mathbf{t}$ based on different CVP algorithms is given by

$$\mathbf{t} = \begin{cases} \mathbf{H}^\dagger\mathbf{s} - \mathcal{Q}_{\mathcal{L}(\mathbf{H}^\dagger\mathbf{M})}(\mathbf{H}^\dagger\mathbf{s}), & \text{Sphere Precoding} \\ \mathbf{H}^\dagger\mathbf{s} - \mathrm{SIC}(\mathbf{H}^\dagger\mathbf{s}, \mathbf{H}^\dagger\mathbf{M}), & \text{SIC Precoding} \\ \mathbf{H}^\dagger\mathbf{s} - \mathrm{ZF}(\mathbf{H}^\dagger\mathbf{s}, \mathbf{H}^\dagger\mathbf{M}), & \text{ZF Precoding} \end{cases}$$

in which $\mathrm{SIC}(\mathbf{H}^\dagger\mathbf{s}, \mathbf{H}^\dagger\mathbf{M})$, $\mathrm{ZF}(\mathbf{H}^\dagger\mathbf{s}, \mathbf{H}^\dagger\mathbf{M})$ respectively denotes using SIC [12], ZF [12] to output the closest lattice point based on the query $\mathbf{H}^\dagger\mathbf{s}$ and lattice basis $\mathbf{H}^\dagger\mathbf{M}$. The base station then transmits the power-controlled vector $\mathbf{t}\sqrt{\mathrm{SNR}/\beta}$ with $\beta = \mathbb{E}\|\mathbf{t}\|^2$. At the receivers' side, we have

$$\mathbf{r} = (\mathbf{s} - \mathbf{M}\mathbf{x})\sqrt{\mathrm{SNR}/\beta} + \mathbf{w}. \tag{16}$$

Finally, each user i employs its own p_i to design the modulo operation for its own estimation:

$$\hat{s}_i = \lfloor r_i\sqrt{\beta/\mathrm{SNR}} \rceil \mod p_i = \lfloor s_i + w_i\sqrt{\beta/\mathrm{SNR}} \rceil \mod p_i. \tag{17}$$

E. Complexity

The additional costs: Regarding **Step 1)**, finding the lengths of n vectors requires $O(n^2)$ floating point operations (FLOPS), and the worst case complexity of sorting is known as $O(n^2)$. The calculations in **Steps 2)** and **3)** are also light-weighted: the total number of quantization in $(n-1)/2$ pairs of vectors is $(n-1)|\mathcal{T}|$. Moreover, the preparation of $\mathbf{b}_1^*, \ldots, \mathbf{b}_n^*$ does not incur additionally complexity because SIC precoding generates the availability of these GSO vectors at the beginning. Therefore, the total complexity caused by rectangular shaping is summarized as $O(n^2)$. *Summing up:* As ZF/SIC requires $O(mn^2)$ FLOPS [4], we conclude that the rectangular shaping based ZF/SIC precoders still exhibit the same complexity of $O(mn^2)$.

IV. Simulation Results

In this section we conduct numerical simulations. VP variants based on conventional self-similar shaping is taken as the benchmarks. Associated algorithms to solve CVP include: "ZF," "SIC," "D2VP" [14], and "Sphere precoding" (i.e., conventional VP [1]). Their improved versions based on the proposed rectangular shaping are marked with a prefix "rec," e.g., "rec-SIC". The constellation is chosen as $\mathcal{M} = \{0, \ldots, 7\}$ with $p = 8$, and the results for other sizes of p are similar.

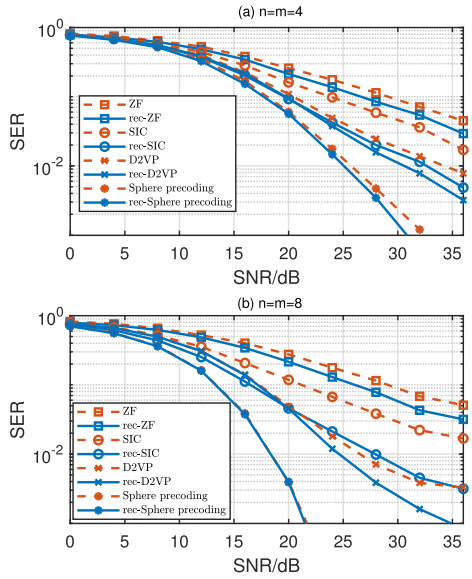


Fig. 3. The SER performance in small-scale systems.

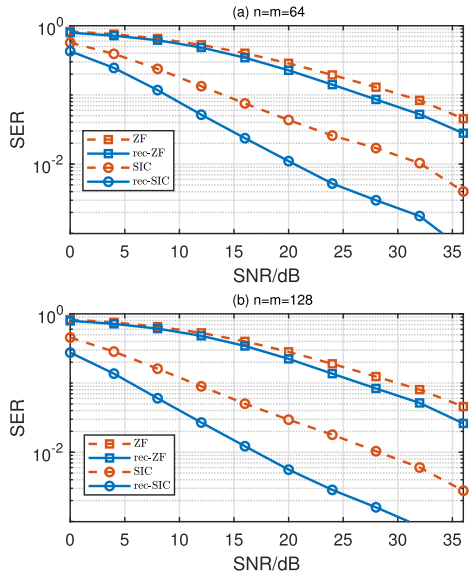


Fig. 4. The SER performance in large-scale systems.

Fig. 3 depicts the symbol-error rate (SER) performance based on different shaping schemes and algorithms with $n = m = 4$ and $n = m = 8$. Some speculations are made: i) rec-ZF, rec-SIC, and rec-D2VP all perform better than their counterparts by at least 3 dBs in the two subfigures. The reason behind the significant improvement of SIC is that, the shaping lattice is designed by minimizing the covering radius of the GSO parallelotope. ii) While we observe about 1 dB gain for rec-Sphere precoding in Fig. 3(a), its gain in Fig. 3(b) is minor, because the objective in (10) is not minimizing the covering radius of lattice $\mathcal{L}(\mathbf{BM})$, and the $\rho(\mathcal{P}_{SIC,\mathcal{L}(\mathbf{BM})})$ based shaping attains better performance for sphere precoding only in small dimensions. Nevertheless, the complexity saving of rec-Sphere precoding is significant, as the lattice basis $\mathbf{H}^\dagger\mathbf{M}$ is much more orthogonal than $p\mathbf{H}^\dagger$.

For the large-scale setting of $n = m = 64$ and $n = m = 128$, we plot the SER performance for the advocated SIC and ZF precoders in Fig. 4. Both subfigures show that the proposed schemes feature evident performance improvement. E.g, in Fig. 4(b), rec-SIC yields around 12 dB gain and rec-ZF provides about 3 dB gain.

V. Conclusion

We have presented an improved design of VP which adaptively changes the shaping lattice. This technique can be attached to the low-complexity ZF and SIC based implementations of VP to achieve significant performance gains. In a high level, we have approximately achieved the goal of lattice reduction (obtaining more orthogonal basis) simply based on rectangular shaping.

References

[1] B. M. Hochwald, C. B. Peel, and A. L. Swindlehurst, "A vector-perturbation technique for near-capacity multiantenna multiuser communication-part II: Perturbation," *IEEE Trans. Commun.*, vol. 53, no. 3, pp. 537–544, Mar. 2005.

[2] C. Windpassinger, R. F. H. Fischer, and J. B. Huber, "Lattice-reduction-aided broadcast precoding," *IEEE Trans. Commun.*, vol. 52, no. 12, pp. 2057–2060, Dec. 2004.

[3] S. Liu, C. Ling, and X. Wu, "Proximity factors of lattice reduction-aided precoding for multiantenna broadcast," in *Proc. IEEE Int. Symp. Inf. Theory*, Cambridge, MA, USA, 2012, pp. 2291–2295.

[4] S. Lyu and C. Ling, "Hybrid vector perturbation precoding: The blessing of approximate message passing," *IEEE Trans. Signal Process.*, vol. 67, no. 1, pp. 178–193, Jan. 2019.

[5] C. Masouros, M. Sellathurai, and T. Ratnarajah, "Maximizing energy efficiency in the vector precoded MU-MISO downlink by selective perturbation," *IEEE Trans. Wireless Commun.*, vol. 13, no. 9, pp. 4974–4984, Sep. 2014.

[6] C. B. Chae, S. Shim, and R. W. Heath, "Block diagonalized vector perturbation for multiuser mimo systems," *IEEE Trans. Wireless Commun.*, vol. 7, no. 11, pp. 4051–4057, Nov. 2008.

[7] A. Li and C. Masouros, "A constellation scaling approach to vector perturbation for adaptive modulation in MU-MIMO," *IEEE Wireless Commun. Lett.*, vol. 4, no. 3, pp. 289–292, Jun. 2015.

[8] Y. Avner, B. M. Zaidel, and S. Shamai, "On vector perturbation precoding for the MIMO Gaussian broadcast channel," *IEEE Trans. Inf. Theory*, vol. 61, no. 11, pp. 5999–6027, Nov. 2015.

[9] D. J. Ryan, I. B. Collings, I. V. L. Clarkson, and R. W. Heath, "Performance of vector perturbation multiuser MIMO systems with limited feedback," *IEEE Trans. Commun.*, vol. 57, no. 9, pp. 2633–2644, Sep. 2009.

[10] A. Razi, D. J. Ryan, I. B. Collings, and J. Yuan, "Sum rates, rate allocation, and user scheduling for multi-user MIMO vector perturbation precoding," *IEEE Trans. Wireless Commun.*, vol. 9, no. 1, pp. 356–365, Jan. 2010.

[11] F. Rusek *et al.*, "Scaling up MIMO: Opportunities and challenges with very large arrays," *IEEE Signal Process. Mag.*, vol. 30, no. 1, pp. 40–60, Jan. 2013.

[12] C. Ling, "On the proximity factors of lattice reduction-aided decoding," *IEEE Trans. Signal Process.*, vol. 59, no. 6, pp. 2795–2808, Jun. 2011.

[13] B. M. Kurkoski, "Encoding and indexing of lattice codes," *IEEE Trans. Inf. Theory*, vol. 64, no. 9, pp. 6320–6332, Sep. 2018.

[14] Y. Ma, A. Yamani, N. Yi, and R. Tafazolli, "Low-complexity MU-MIMO nonlinear precoding using degree-2 sparse vector perturbation," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 3, pp. 497–509, Mar. 2016.