A Clustering Approach to Optimizing Beam Steering Directions in Wireless Systems

Ahmed Elshafiy*, Ashwin Sampath†, and Kenneth Rose*
*University of California, Santa Barbara, CA 93106, USA
Email: {a_elshafiy, rose}@ece.ucsb.edu
†Qualcomm, Bridgewater Township, NJ 08807, USA
Email: asampath@qualcomm.com

2021 IEEE Wireless Communications and Networking Conference (WCNC) | 978-1-7281-9505-6/20/$31.00 ©2021 IEEE | DOI: 10.1109/WCNC49053.2021.9417274

***Abstract*—Directional beamforming with large antenna arrays is key to mitigating the substantial signal loss experienced at the millimeter wave frequency band, important to fifth generation (5G) cellular systems and next generation wireless local area networks, where it entails a significant increase in the number of beams. This paper is motivated by the realization that the underlying problem of finding the optimal set of beam steering directions will benefit from fundamental signal processing methodologies, and specifically from basic principles and algorithms for cluster analysis. Earlier work by authors established the equivalence between the problem of optimizing a set of beam steering directions and the classical problems of clustering and quantizer design, albeit with an unusual distortion measure. Subsequently in [1], a k-means-like approach was derived to optimize beam steering directions and guarantee convergence to at least locally optimal solution. The main contribution of this work is the derivation of a global optimization approach within the deterministic annealing framework, to circumvent poor local optima that riddle the cost surface. Simulation results show that the approach delivers considerable gains over the baseline uniform beam steering technique, specifically, up to 6 dB and 13 dB gains, in terms of average and 10th percentile of the power array factor, respectively, as well as up to 6.5 dB gain in the average Signal-to-Noise Ratio (SNR).**

***Index Terms*—Wireless systems, Beam Steering, Deterministic Annealing, Clustering, Non-convex Optimization.**

I. Introduction

Multiple-Input Multiple-Output (MIMO) systems in conjunction with millimeter-wave frequencies have been recognized as a promising tool in the effort to satisfy the ever-growing demand for higher data-rates. Given that physical layer technologies already operate at, or close to, Shannon capacity, the main focus must be on the system bandwidth [2], [3]. Studies have shown that considerable rate gains can be achieved through millimeter-wave communications by exploiting the substantial bandwidth available at these frequencies. However, a number of significant challenges arise as well [4], [5], including increased path-loss, shadowing losses, signal attenuation, and atmospheric absorption at some frequencies, which cause considerable decrease in link budget and result in considerable reduction in cell coverage area. To meet this challenge, larger transmit/receive arrays, and hence increased array factors, are employed to boost the link budget. Considering a transmit linear-array of length N_{tx}, the increase in effective isotropic radiated power (EIRP) due to beamforming is proportional to N_{tx} [6], yielding a corresponding increase in the receiver signal-to-noise ratio (SNR). However, the half-power beam-width (HPBW) is inversely proportional to N_{tx}. Thus, large arrays offer EIRP gains in the steering direction, but at the cost of narrower beams which in-turn require an increase in the number of beams needed to maintain acceptable spatial coverage. Both transmitter and receiver typically operate with predefined "*codebooks*" of beamforming vectors, wherein each codebook entry corresponds to a beam steering direction. An increase in codebook size hinders beam tracking and beam alignment due to the inherent increase in beam measurement time (sweep time) and thus compromises the system responsiveness to user and environment dynamics.

Beam-broadening was proposed as a counter measure to allow a tradeoff between the requirements of high EIRP and low beam management complexity, especially in conjunction with user tracking and initial access [6]–[9]. Additionally, enhanced robustness to user dynamics can be achieved by employing a more efficient beam search or beam alignment algorithms for a given codebooks of beam steering directions, as has been pursued in [10], [11]. A central motivation for this paper is the realization that, regardless of the beam width or the beam alignment algorithm, the overall performance can be improved by optimal design of the beam steering directions, to match the observed or estimated user statistics. It is intuitively obvious that an optimal design of beam steering directions will jointly consider the distribution of users as well as the direction-dependent beam width. The objective of this paper is to develop a sound methodology, from basic signal processing principles, for finding the optimal set of beam steering directions, i.e., designing the beam steering codebook, given a codebook size budget.

The problem of finding the optimal beam steering angles can, in fact, be viewed as a clustering problem, where the two-dimensional angular space (azimuth and elevation angles) is partitioned into N_b sub-cells each represented by a pointing angle [1]. As the number of pointing angles is increased, the average link performance over the angular space increases, but so does the rate of beam updates, and the system becomes less robust to dynamics. This tradeoff is analogous to the classical rate-distortion tradeoff considered in quantizer design for data compression. As the quantizer design or, more generally, the clustering problem, appears with various flavors in many diverse applications, solution methods have been developed in multiple disciplines. In the communications or information-theory literature, an early clustering method was suggested for scalar quantization, variants of which are known as the Lloyd algorithm [12] or the Max quantizer [13]. This method was

978-1-7281-9505-6/21/$31.00 ©2021 IEEE

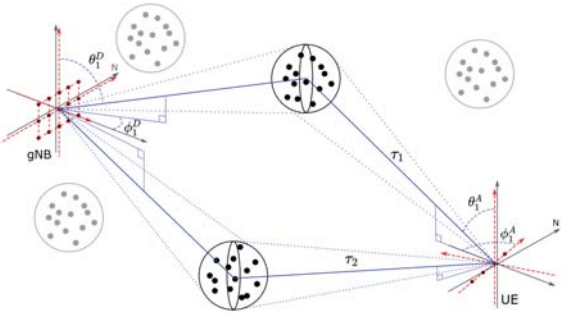


Fig. 1: Snapshot of the 3D CDL channel model in [20].

later generalized to vector quantization (VQ), and to a large family of distortion measures [14], and the resulting algorithm is commonly referred to as the generalized Lloyd algorithm (GLA). In the pattern-recognition literature, similar algorithms have been introduced including the ISODATA [15] and the k-means [16] algorithms. We note in passing that VQ techniques have been pursued in the wireless communications literature for a different problem, namely, beamforming adaptation to estimated channel coefficients [17], [18].

All the above iterative methods alternate between two complementary steps (often referred to as the Lloyd iteration): optimization of the partition into clusters given the current codebook entries, and optimization of the codebook entries for their respective clusters. It is easy to show that such an iterative procedure is monotone non-increasing in the distortion, and convergence to a local minimum of the distortion is guaranteed. The Deterministic Annealing (DA) approach, for conventional distortion measures, has been proposed as a powerful algorithm for avoiding poor local minima [19]. The optimal solution can be tracked in a deterministic annealing framework, starting at the global optimum for high distortion (where the cluster means all coincide at a single point, i.e., we have a single effective mean for the entire training set) and tracking the minimum as the temperature (the Lagrangian parameter controlling the tradeoff between distortion and entropy) is lowered. During this annealing process, the system undergoes a sequence of "phase transitions" whereby the cardinality of the set of effective means increases. Inspired by principles of statistical physics and derived in terms of information theory, DA was proposed as a powerful non-convex optimization tool for compression, clustering, classification and related problems.

The overall novelty of this work is the derivation of an approach within a powerful optimization framework, namely, deterministic annealing to circumvent poor local optima (that might result from the k-means algorithm in [1]), solve the clustering problem at hand, and achieve significant performance gains. The remainder of this paper is organized as follow: Section II provides relevant background, including definition of system model and review of beamforming techniques. The novel DA-based beam steering algorithm, is presented in Section III. The experimental evaluation is detailed in Section IV. Conclusions are drawn in Section V.

II. Relevant Background

A. System Model

Consider the downlink transmission direction. For outdoor settings, the 5G base station (also called gNB) is equipped with a planar array consisting of N_{tx} antennas, while the user equipment (UE) comprises a linear array consisting of N_{rx} antennas. Let $s \in \{1, 2, \ldots, N_{\text{tx}}\}$ and $u \in \{1, 2, \ldots, N_{\text{rx}}\}$ denote the transmit and receive antenna indices, respectively. The downlink channel we consider, is modeled as the 3GPP Cluster Delay Line (CDL) channel [20], which is depicted in Fig. 1. Let N_{c} denote the number of detected clusters, and M_{r} the number of rays within a single cluster. Let $m \in \{1, 2, \ldots, M_{\text{r}}\}$ be the ray index, and $n \in \{1, 2, \ldots, N_{\text{c}}\}$ be the cluster index. The $(N_{\text{rx}} \times N_{\text{tx}})$ channel matrix is denoted by $\mathbf{H}_{n,m}(t)$, where t is the time index. Next, the unit-norm phase-control $(N_{\text{tx}} \times 1)$ transmit beamforming vector and, similarly, the $(N_{\text{rx}} \times 1)$ receive beamforming vector are denoted by $\mathbf{b}_{\text{tx}}(\boldsymbol{\varphi})$ and $\mathbf{b}_{\text{rx}}(\boldsymbol{\vartheta})$, respectively, where $\boldsymbol{\varphi}$, and $\boldsymbol{\vartheta}$ are the transmit and receive vectors of the beamforming phases. Beamforming can be performed using amplitude control, phase control, or both. Without loss of generality, we will focus herein on phase-control based beamforming which is more power efficient than amplitude-based beamforming [9]. The system model in this setting is given by,

$$y(t, f_r) = (\mathbf{b}_{\text{rx}}(\boldsymbol{\vartheta}))^{\text{H}} \sum_{n=1}^{N_{\text{c}}} \sum_{m=1}^{M_{\text{r}}} \left\{ \left(\mathbf{H}_{n,m}(t) e^{-j2\pi f_r \tau_n(t)} \right) \mathbf{b}_{\text{tx}}(\boldsymbol{\varphi}) \right. \\ \left. x(t, f_r) \right\} + (\mathbf{b}_{\text{rx}}(\boldsymbol{\vartheta}))^{\text{H}} \mathbf{n}(t, f_r), \tag{1}$$

where f_r is the rth sub-carrier frequency, $\tau_n(t)$ is the nth cluster delay, $x(t, f_r)$ is the complex frequency domain transmit symbol with $\mathbb{E}\left[|x(t, f_r)|^2\right] = 1$, and $\mathbf{n}(t, f_r) \sim \mathcal{CN}(0, \sigma_n^2)$ is the complex AWGN vector, with $\sigma_n^2 = k_{\text{B}} T B$, where k_{B} is the Boltzmann constant, T is the temperature and B is the transmission bandwidth. We employ the standard notation $(\cdot)^{\text{T}}$ and $(\cdot)^{\text{H}}$ to denote transposition and the conjugate transposition operations, respectively. The (u, s) element of the channel matrix $\mathbf{H}_{n,m}(t)$ is denoted by $h_{n,m}^{u,s}(t)$, and detailed description of the channel coefficients model can be found in [9], [20], [21]. The perceived channel coefficients, after beamforming for each cluster and ray, are obtained as,

$$h_{n,m}(t, f_r) = (\mathbf{b}_{\text{rx}}(\boldsymbol{\vartheta}))^{\text{H}} \mathbf{H}_{n,m}(t) e^{-j2\pi f_r \tau_n(t)} \mathbf{b}_{\text{tx}}(\boldsymbol{\varphi}), \tag{2}$$

and the aggregate channel transfer function, due to all clusters and rays, is

$$h(t, f_r) = \sum_{n=1}^{N_{\text{c}}} \sum_{m=1}^{M_{\text{r}}} h_{n,m}(t, f_r). \tag{3}$$

The SNR at sub-carrier r with frequency f_r is given by,

$$\gamma_r(t) = \frac{P_{\text{tx}} G_{\text{tx}} |h(t, f_r)|^2 G_{\text{rx}}}{P_{\text{L}}(t) F_n \ \sigma_n^2}, \tag{4}$$

where P_{tx} is the average transmit power, $P_{\text{L}}(t)$ is the path-loss, F_n is the receiver noise factor, and where G_{tx} and G_{rx} are the maximum gains of the transmit and receive antenna elements

relative to an isotropic antenna element, respectively.

B. Beamforming Techniques

This subsection provides an analysis for phase-control transmit beamforming, noting that the corresponding analysis for receive beamforming is similarly obtained in a straightforward manner. Consider a planar antenna array with uniform spacing between horizontal and vertical elements, i.e., $d_x = d_y = \frac{\lambda_c}{2}$, where λ_c is the carrier wavelength. Define the beam-space transformation on the x-axis and y-axis as $\Omega_x = kd_x \sin(\theta)\cos(\phi) = \pi \sin(\theta)\cos(\phi)$, and $\Omega_y = kd_y \sin(\theta)\sin(\phi) = \pi \sin(\theta)\sin(\phi)$, where $k = \frac{2\pi}{\lambda_c}$ is the wave number, θ is the elevation angle, and ϕ is the azimuth angle. The conventional planar array setup is shown in [1, Fig. 1]. Hence, the transmit power array factor simplifies to [22],

$$\begin{aligned} \mathbf{a}_{\text{tx}}(\Omega_x, N) &\triangleq \left[1 \;\; e^{-j\Omega_x} \;\; \ldots \;\; e^{-j\Omega_x(N-1)}\right]^{\text{T}}, \\ \mathbf{a}_{\text{tx}}(\Omega_y, N) &\triangleq \left[1 \;\; e^{-j\Omega_y} \;\; \ldots \;\; e^{-j\Omega_y(N-1)}\right]^{\text{T}} \end{aligned} \tag{5}$$

$$\mathbf{b}_{\text{tx}}(\varphi) \triangleq \mathbf{b}_{\text{tx}}^{(x)}(\varphi_x) \otimes \mathbf{b}_{\text{tx}}^{(y)}(\varphi_y), \tag{6}$$

$$\begin{aligned} A_{\text{tx}}(\Omega_x, \Omega_y, \varphi_x, \varphi_y) &= \left(\mathbf{a}_{\text{tx}}(\Omega_x)^{\text{H}} \mathbf{b}_{\text{tx}}^{(x)}(\varphi_x)\right) \cdot \\ &\quad \left(\mathbf{a}_{\text{tx}}(\Omega_y)^{\text{H}} \mathbf{b}_{\text{tx}}^{(y)}(\varphi_y)\right), \end{aligned} \tag{7}$$

$$A_{\text{tx}}(\Omega_x, \Omega_y, \varphi_x, \varphi_y) = A_{\text{tx}}^{(x)}(\Omega_x, \varphi_x) A_{\text{tx}}^{(y)}(\Omega_y, \varphi_y),$$

where $\mathbf{b}_{\text{tx}}^{(x)}(\varphi_x)$ and $\mathbf{b}_{\text{tx}}^{(y)}(\varphi_y)$ are the beamforming vectors along the x and y coordinates of the planar array in [1, Fig. 1], respectively. The Kronecker product operation is denoted by $\otimes$. The array factor can be maximized at a given steering direction by using the conventional Constant Phase Offset (CPO) beamforming technique [6], [8], [9], [22], yielding the beamforming vectors:

$$\mathbf{b}_{\text{tx}}^{(x)}(\omega_x) = \frac{1}{\sqrt{N_{\text{x}}}} \left[1 \quad e^{-j\omega_x} \quad \ldots \quad e^{-j\omega_x(N_{\text{x}}-1)}\right]^{\text{T}}, \tag{8}$$

$$\mathbf{b}_{\text{tx}}^{(y)}(\omega_y) = \frac{1}{\sqrt{N_{\text{y}}}} \left[1 \quad e^{-j\omega_y} \quad \ldots \quad e^{-j\omega_y(N-1)}\right]^{\text{T}}, \tag{9}$$

where $\omega_x = \pi \sin(\theta_0)\cos(\phi_0)$ and $\omega_y = \pi \sin(\theta_0)\sin(\phi_0)$ are the beam space transformation of the elevation and azimuth steering angles θ_0 and ϕ_0, respectively. In this setting, the highest possible array factor, $10\log_{10}(N_{\text{tx}})$ dB, is guaranteed at the steering direction. It is worthwhile to note in passing that if a single beam is scheduled by the base station to serve a user, then maximizing the array factor at the dominant channel direction between the UE and the gNB will consequently boost the perceived user SNR, defined in (4). Hence, the average beamforming array factor across users is the objective function of choice for the beam steering design approaches introduced in Section III. Additionally, in [8], authors showed that the low-complexity dominant directional beamforming scheme suffers only a minimal SNR loss (less than a dB loss for over 50% of the users in channels with up to $N_c = 5$ clusters) relative to even the best beamforming scheme. Consequently, single serving beam per user with CPO beamforming at the channel dominant direction has been widely employed in practice [6], [8], [23].

III. Optimal Beam Steering Directions

This section covers our main contribution, namely, the development of DA-based codebook design method to approach beam steering optimality. Note that we only focus on designing the pointing angles of the codebook for CPO beams, without recourse to other design aspects such as beam shape, side lobes level, etc. The beamforming vectors are stored as codebook entries, such that each entry corresponds to an angular direction. Specifically, each codebook entry corresponds to an elevation and azimuth angle pair. The simplest (and most common) beam steering approach is to quantize the elevation and azimuth field-of-view *uniformly* into N_b pointing directions, similar to [7], [24], where N_b is the number of beams (entries) in the codebook. A somewhat more sophisticated approach quantizes the beam-space field-of-view $\overline{\Omega}_x$, and $\overline{\Omega}_y$ uniformly, which is known as the Discrete Fourier Transform (DFT) codebook [23], [25].

It is important to note that the beam shape is direction-dependent, i.e., different beam steering angles result in wider or narrower beams, as was shown in [1, Fig. 2]. Moreover, a common simplifying assumption is that the UE positions are uniformly distributed on the horizontal plane [20], which nevertheless results in a non-uniform distribution of user angles ϕ_i and θ_i across the angular space, where i is the user index. Uniform distribution of steering angles implies that the beams' density across the angular space remains unchanged in the regions of space at which the beams are wider or in the regions of space at which there is low or no user density. Hence, we conclude that uniform distribution of beam steering angles across the field-of-view is virtually always suboptimal, even under simplistic assumptions such as uniform user distribution on a plane. Earlier work by authors in [21] proposed a heuristic beam steering method. Although this technique accounts for the non-uniform beam width, it does not account for users location distribution, which is potentially time-varying. However, this heuristic non-uniform beam steering technique aims to maximize the beam coverage across the field of view as shown in [21, Fig. 5], and is useful in cases where user statistics are unknown or hard to obtain. Hence, our second approach to this problem was to pursue an iterative framework that guarantees convergence to (at least locally) optimal performance (results appeared in [1]). The first key realization is that the beam steering problem at hand is effectively equivalent to a generalized clustering problem (albeit with an unusual distortion measure). The space to be divided into regions is the 2-dimensional angular space, with boundaries specified by the transmitter field-of-view. The data vectors to be clustered are the users' angle vectors as seen from transmitter local coordinate system, which are denoted as $\psi_i = [\phi_i \;\; \theta_i]^{\text{T}}$. For each cluster, a single beam steering direction, which we will also refer to as the cluster centroid, is chosen to serve any of the users in the cluster. A novel distortion function between the ith vector ψ_i and the jth codebook entry χ_j was defined as:

$$d(\psi_i, \chi_j) = \sqrt{N_x N_y} - |A_{\text{tx}}^{(x)}(\psi_i, \chi_j)||A_{\text{tx}}^{(y)}(\psi_i, \chi_j)|, \quad (10)$$

where $|A_{\text{tx}}^{(x)}(\psi_i, \chi_j)|$ and $|A_{\text{tx}}^{(y)}(\psi_i, \chi_j)|$ are the per-dimension absolute array factors. For example, if the CPO technique (directional beam) is employed, the per-dimension array factors are

$$\begin{aligned}
\left|A_{\text{tx}}^{(x)}(\psi, \chi)\right| &= \left|A_{\text{tx}}^{(x)}([\phi\ \theta]^{\text{T}}, [\phi_0\ \theta_0]^{\text{T}})\right| \\
&= \frac{1}{\sqrt{N_x}}\left[\frac{\sin\left(\frac{N_x\pi}{2}\left(\cos(\phi)\sin(\theta) - \cos(\phi_0)\sin(\theta_0)\right)\right)}{\sin\left(\frac{\pi}{2}\left(\cos(\phi)\sin(\theta) - \cos(\phi_0)\sin(\theta_0)\right)\right)}\right], \\
\left|A_{\text{tx}}^{(y)}(\psi, \chi)\right| &= \left|A_{\text{tx}}^{(y)}([\phi\ \theta]^{\text{T}}, [\phi_0\ \theta_0]^{\text{T}})\right| \\
&= \frac{1}{\sqrt{N_y}}\left[\frac{\sin\left(\frac{N_y\pi}{2}\left(\sin(\phi)\sin(\theta) - \sin(\phi_0)\sin(\theta_0)\right)\right)}{\sin\left(\frac{\pi}{2}\left(\sin(\phi)\sin(\theta) - \sin(\phi_0)\sin(\theta_0)\right)\right)}\right].
\end{aligned} \quad (11)$$

In other words, the distortion between ith user and jth beam steering angle is defined as *the decrease in absolute array factor*, relative to the maximum achievable value (in the ideal setting). This will subsequently take into account the direction-dependent beam width, and thus users are assigned to clusters at which the transmit array factor is maximized. Next, a variant of the k-means algorithm (or the GLA) was derived to optimize the codebook of beam steering angles (see [1] for details). It was shown that the two steps of the k-means algorithm's main iteration guarantee that average distortion across users D is monotonically non-increasing, and in fact monotonically decreasing until convergence (under mild assumptions regarding treatment of ties in the nearest neighbor step).

One major drawback of the classical k-means clustering algorithm, is that it only guarantees convergence to a locally optimal solution, while in many cases of interest the cost surface is riddled with poor local minima. A variety of heuristic approaches have been proposed to tackle this difficulty, and they range from repeated optimization with different initialization, and heuristics to obtain good initialization, to heuristic rules for cluster splits and merges, etc. Nevertheless, there is a substantial gain to be recouped by a principled attack on the problem. This motivates the use of powerful optimization tools. Deterministic annealing has been demonstrated to be highly effective in avoiding poor local minima, when conventional distortion measures are used, and has become the method of choice in numerous disciplines [19]. DA is motivated by the annealing process in physical chemistry, where certain chemical systems are driven to their low energy states by annealing, i.e., via gradual cooling of their temperature. Additional non-convex optimization tools have been also inspired by the annealing process of chemical systems such as stochastic relaxation [26] or simulated annealing [27]. However these optimization methods can only reach the global minimum if the rate of lowering the temperature follows $T \propto 1/\log(n)$, where n is the iteration index [26]. This slow annealing schedule is often unrealistic in many practical applications. As its name suggests, DA tries to enjoy the best of two scenarios. On the one hand it is deterministic, meaning that random motion on the energy surface while making incremental progress on the average, as is the case for stochastic relaxation, is discouraged due to its slow convergence. On the other hand, it is still an annealing method and aims at the global minimum, instead of getting greedily attracted to a nearby local minimum.

DA introduces a controlled amount of randomness in the optimization, measured by the Shannon entropy, and controlled by a Lagrange multiplier T, analogous to "temperature" in the physical system. The resulting Lagrangian, an expectation function accounting for the tradeoff between distortion and entropy, is in fact exactly the Helmholtz free energy in physics, and is *deterministically* minimized at successive temperatures, thus circumventing the high computational complexity of stochastic simulated annealing. However, utilization of the DA in the problem at hand is challenging due to the existence of the irregular distortion function defined in (10). Hence, in this work, a variant of DA is derived and employed in order to optimize the codebook of beam steering directions.

Unlike the k-means algorithm, DA considers a probabilistic assignment between the users' angular vectors $\{\psi_i\}$ and codebook entries or cluster centroids $\{\chi_j\}$. Let the association probabilities be denoted as $p(j|i)$. In this case, the overall average distortion in the system due to quantization of beam pointing angle is given by the expectation,

$$D = \sum_i \sum_j p(j|i)p(i)d(\psi_i, \chi_j), \quad (12)$$

where $p(i)$ is the prior probability of a user positioned at angular vector ψ_i. Note that minimizing the distortion with respect to the free parameters $\{\chi_j, p(j|i)\}$ would immediately lead to hard association between the user and the nearest codebook entry, where the term "nearest" is used in the sense of the distortion measure. Instead, the distortion is minimized subject to an imposed level of randomness which is naturally measured by Shannon's entropy H. Hence, the Lagrangian function to be minimized can be written as,

$$\mathcal{L} = D - TH, \quad (13)$$

where,

$$H = -\sum_i \sum_j p(j|i)p(i)\log\left(p(j|i)p(i)\right), \quad (14)$$

and T ("temperature") is the Lagrangian parameter. Next, an iterative approach, which is an appropriately designed random relative of the k-means algorithm, is employed to minimize the Lagrangian function:

1) Initialize temperature, $T = T_{\max}$ and beam steering angles codebook $\{\chi_j\}$.
2) Fix the codebook $\{\chi_j\}$ and find the random clustering partition (i.e., probabilistic assignment of users to steering angles) which minimizes the Lagrangian cost:

$$\{p(j|i)\} = \underset{\{p(j|i)\}}{\arg\min}\ \mathcal{L}, \qquad \forall i, \forall j \quad (15)$$

Note that the solution must further impose the constraint $\sum_j p(j|i) = 1, \forall i$, which directly yields a random relative of the nearest neighbor rule, given by the Gibbs distribution:

$$p(j|i) = \frac{\exp\left(-\frac{d(\psi_i, \chi_j)}{T}\right)}{Z_i}, \tag{16}$$

where the normalization constant is

$$Z_i = \sum_j \exp\left(-\frac{d(\psi_i, \chi_j)}{T}\right), \tag{17}$$

sometimes called the partition function in physics.

3) Fix the random clustering partition, $\{p(j|i)\}$ and optimize the steering angles codebook to minimize the Lagrangian cost. Specifically,

$$\{\chi_j\} = \arg\min_{\{\chi_j\}} \mathcal{L} = \arg\min_{\{\chi_j\}} D, \tag{18}$$

where we used the fact that the entropy is determined by the (fixed) clustering partition, and hence can be discarded from $\mathcal{L}$ in this step. Noting further that D is additive in the contributions of individual steering angles we obtain:

$$\chi_j = \arg\min_{\chi} \sum_i p(j|i)p(i)d(\psi_i, \chi), \tag{19}$$

or as necessary condition for optimality, the random relative of the centroid rule:

$$\sum_i p(j|i)p(i)\frac{\partial}{\partial \chi} d(\psi_i, \chi) = 0 \quad , j = 1, 2, \ldots, N_b, \tag{20}$$

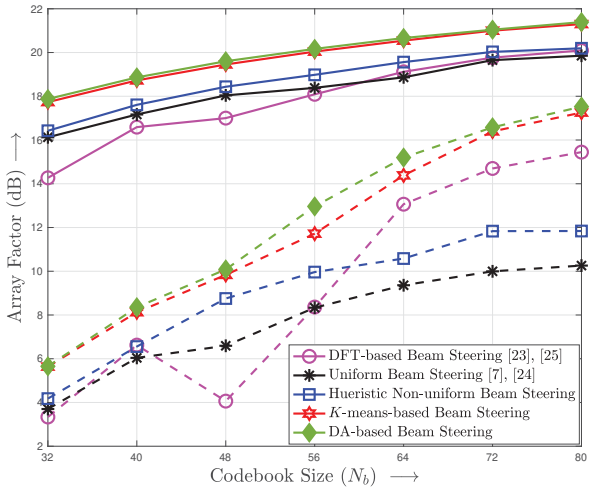
Numerical search with finite resolution in the 2D angular space or gradient descent algorithms with multiple initialization points or both can be employed to solve the minimization problem of (19).

4) Check if convergence condition satisfied, else go to step 2.
5) Cool the system, e.g., $T = \alpha T$, with $\alpha < 1$. If the prescribed minimum temperature is reached then terminate the algorithm.
6) Perturb codebook entries to check for possible splitting of codebook centroids (or phase transitions) then go to step 2.

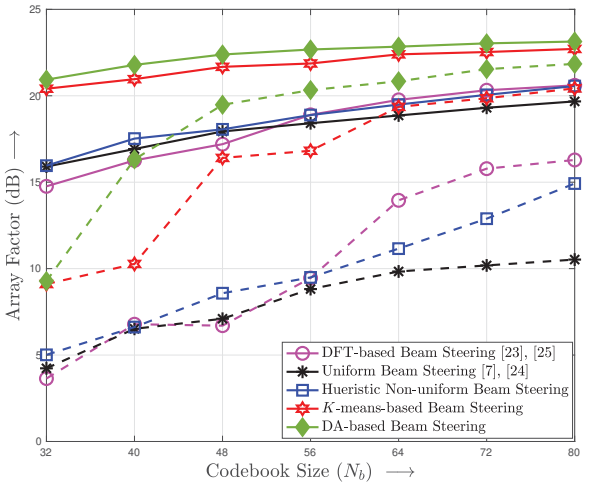
At $T = 0$, the DA algorithm degenerates to the k-means algorithm, however the annealing process until then eliminates the sensitivity to initialization. In step 4, convergence can be checked by comparing $\frac{\Delta \mathcal{L}}{\mathcal{L}}$ to a convergence threshold. It is important to note that by gradual cooling, the system undergoes a series of phase transitions at corresponding "critical temperatures", in analogy to physical systems, wherein the cardinality of the codebook grows. See [19] for extensive analysis of DA's sequence of phase transitions through which the cardinality of the codebook grows, as well as for demonstration that the algorithm is invariant to initialization.

IV. Experimental Results

The beam steering angles optimization algorithms are first evaluated in terms of the average and the 10th percentile of the array factor seen across all users. The competing beam placement schemes are: *i*) DFT-based beam steering as defined in [23], [25] *ii*) Uniform beam steering as employed in [7], [24], *iii*) Heuristic non-uniform beam steering in [21], *iv*) k-means-based beam steering in [1], and finally *v*) DA-based beam steering proposed in Section III. The former two (DFT-based and uniform beam steering) serve as baseline reference for the comparison, and the latter three are the proposed schemes presented in this paper and earlier work of the authors. For uniform beam steering and heuristic non-uniform beam steering algorithms, the number of elevation beams $N_b^\theta \in \{1, 2, \ldots, 16\}$, and the selected value of N_b^θ is numerically optimized for each codebook size N_b, so as to maximize the average array factor. The gNBs are assumed to be equipped with 32×8 planar arrays. The performance is evaluated for a variety of UE distributions. First, the UE angles, seen from the gNB local co-ordinate system, are assumed to be uniformly distributed over the field-of-view ($\overline{\phi} = 180°, \overline{\theta} = 90°$). Fig. 2a depicts the average power array factor and its 10th percentile in



(a) The UEs' angles are assumed uniformly distributed.



(b) The UEs' angles are assumed to be distributed as a mixture of bi-variate Gaussians.

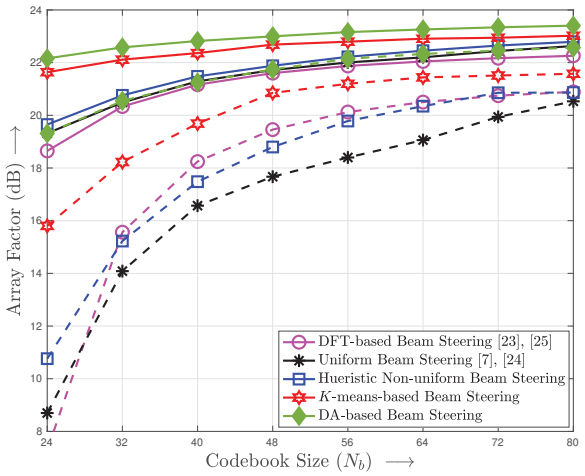
Fig. 2: Average (sold lines) and 10th percentile (dashed lines) of power array factor for competing beam steering design methods.

dB versus the beam codebook size. The proposed DA-based beam steering approach offers gains of up to 4 dB and 7.2 dB, in the average power array factor and its 10th percentile, respectively, when compared with the baseline methods. Note that the codebooks are designed to maximize the average power array factor over all users, which sometimes results in a degraded 10th percentile performance as seen for DFT-based codebook at $N_b = 48$.

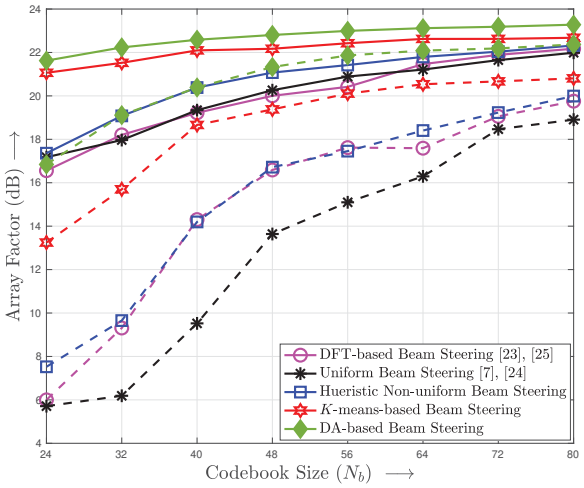
Next, to test the approaches in a less simplistic scenario, the users' angles were distributed as a mixture of bi-variate Gaussians in the angular field-of-view ($\overline{\phi} = 180°, \overline{\theta} = 90°$). The underlying premise of this model is that users often tend to cluster around certain locations such as shops, traffic lights, bus stops, etc. The average power array factor and its 10th percentile are plotted for this scenario in Fig. 2b. Note that in this case, the proposed DA-based codebook design offers up to 6 dB and 12.5 dB improvements in the average power array factor and its 10th percentile, respectively, when compared with uniform or DFT-based beam steering approaches.

We next consider the simple UE distribution suggested in [20] for outdoor urban Micro (UMi) system scenarios, where UE positions are uniformly distributed on the horizontal plane. Under this UE distribution assumption, two network layouts were simulated: *i*) The gNBs are placed in a Manhattan-like grid, and sectorized into 4 sectors, or *ii*) The gNBs are placed in a hexagonal grid, and sectorized into 3 sectors. The inter-site distance for both network layouts is 200m. The average power array factor and its 10th percentile are plotted for this scenario in Fig. 3a and Fig. 3b. The proposed DA-based design method outperforms the baseline methods by up to 5.5 dB and 13 dB in the average power array factor and its 10th percentile, respectively. It is noteworthy that the DA algorithm offers larger gains over the baseline schemes when the UE angles are non-uniformly distributed. This is to be expected because DA can adapt and exploit irregularities in the UE distribution, for example by placing more beams at the angular directions pointing at areas that are more densely populated by UEs. This flexibility is not available to the uniform beam steering method or the DFT-based beam steering method, thus putting them at significant disadvantage in likely scenarios of non-uniform UE distribution.

To provide further evidence for the practical benefits of the proposed beam placement algorithms, a full-fledged system simulation was carried out for outdoor cellular 5G settings. The simulation assumptions are summarized in Table I. Random TDM scheduler is employed per base station sector, where each sector schedules randomly one of the active users. For each gNB-UE link, the transmit beam that maximize the received SNR is enabled, where beams are selected from a pre-defined beamforming codebook that is designed offline. The average SNR performance, calculated using (4), is depicted in Fig. 4. The proposed beam steering algorithms offer up to 4.5 dB and 6.5 dB improvements in the average SNR seen over all users for Manhattan-like, or hexagonal network grids, respectively. Note that while the simulation is for the simple channel (consisting of one ray), the results and conclusions



(a) The UEs' positions are uniformly distributed across the horizontal plane in a Manhattan-like network grid.



(b) The UEs' positions are uniformly distributed across the horizontal plane in a hexagonal network grid.

Fig. 3: Average (sold lines) and 10th percentile (dashed lines) of power array factor for competing beam steering design methods.

are readily extendable to more complex channels. It is further important to emphasize that the performance gains are achieved at no operational cost, because typical beam steering codebooks are designed offline and stored in memory. Thus, the operational complexity of deploying any of the competing codebooks is the same. On the other hand, during their design phase, both the k-means and DA-based algorithms require prior information (or assumptions) on user statistics, which is implicit in the training data used. If the system experiences a dynamic user distribution, DA-based and k-means algorithm would require additional operational complexity in order to track user statistics and update codebooks accordingly.

V. Conclusion

This paper investigates the problem of finding the optimal beam steering codebook to match user statistics. Ultimately, a powerful non-convex optimization technique is derived within

TABLE I: Summary of System Simulation Assumptions.

Metric	Value
System Scenario	UMi
Direction of Transmission	Downlink (gNB to UE)
Carrier Frequency & Bandwidth	f_c = 28 GHz, B = 100 MHz
Sub-carrier Spacing	Δf = 120 kHz
Number of Clusters & Rays	N_{c} = 1 and M_{r} = 1
Path-loss Model	3GPP model in [20]
Network Layout	Manhattan-like or Hex. grid
Inter-site Distance	D = 200 meters
Number of gNBs	25 sites or 19 sites
Number of UEs per site	10 UEs
Avg. TX Power Per PA	23 dBm [21], [28]
gNB Antenna Array Size	N_{tx} = 256 elements
gNB Element Power Model	According to [20]
gNB Max. Element Gain	G_{tx} = 8 dBi
UE Antenna Array Size	N_{rx} = 1 element
UE Element Power Model	Omni Antenna Element
UE Max. Element Gain	G_{rx} = 0 dBi
UE Noise Figure	$10\log_{10}(F_n)$ = 8 dB

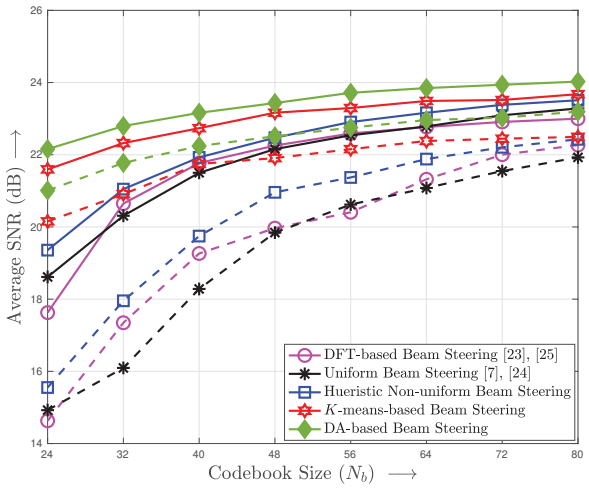


Fig. 4: System SNR performance for competing beam steering design methods in Manhattan-like network gird (solid lines) or hexagonal network grid (dashed lines). The UEs' positions are uniformly distributed across the horizontal plane.

the framework of deterministic annealing, to avoid poor local minima on the cost surface (that might result from the state-of-the-art k-means approach in [1]). The proposed DA-based beam steering algorithm outperforms the baseline uniform steering approaches by up to 6 dB and 12.5 dB in the average and the 10th percentile of power array factor, respectively. Additionally, in a full-fledged system simulation for an outdoor cellular 5G setting, the DA-based algorithms yields SNR gains of up to 6.5 dB. It is noted that the gains in power array factor or in SNR can be traded for significantly reduced codebook size. This would, in turn, reduce the beam management complexity, and hence enhance robustness to user dynamics.

REFERENCES

[1] A. Elshafiy, K. Rose, and A. Sampath, "On Optimal Beam Steering Directions in Millimeter Wave Systems," in *International Conference on Acoustics, Speech, and Signal Processing*, April 2019.

[2] M. Agiwal, A. Roy, and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey," *IEEE Commun. Surveys and Tut.*, vol. 18, pp. 1617–1655, 3rd Quart. 2016.

[3] M. Shafi *et al.*, "5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice," *IEEE J. on Sel. Areas in Commun.*, vol. 35, pp. 1201–1221, 2017.

[4] T. S. Rappaport *et al.*, "Millimeter Wave Mobile Communications for 5G Cellular: It Will Work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[5] J. G. Andrews *et al.*, "What Will 5G Be?" *IEEE J. on Sel. Areas in Commun.*, vol. 32, pp. 1065–1082, 2014.

[6] S. Rajagopal, "Beam Broadening for Phased Antenna Arrays using Multi-beam Subarrays," in *IEEE Int. Conf. on Commun.*, June 2012.

[7] M. Giordani *et al.*, "A Tutorial on Beam Management for 3GPP NR at mmWave Frequencies," *IEEE Commun. Surveys and Tut.*, vol. 21, pp. 173–196, 2018.

[8] V. Raghavan *et al.*, "Beamforming Tradeoffs for Initial UE Discovery in Millimeter-Wave MIMO Systems," *IEEE J. on Sel. Topics Signal Process.*, vol. 10, pp. 543–559, 2016.

[9] A. Elshafiy and A. Sampath, "Beam Broadening for 5G Millimeter Wave Systems," in *IEEE Wireless Commun. and Net. Conf.*, April 2019.

[10] V. Desai *et al.*, "Initial Beamforming for mmWave Communications," in *48th Asilomar Conf. on IEEE Sig., Systems and Comp.*, Nov. 2014.

[11] C. N. Barati, S. A. Hosseini, S. Rangan, P. Liu, T. Korakis, S. S. Panwar, and T. S. Rappaport, "Directional Cell Discovery in Millimeter Wave Cellular Networks," *IEEE Trans. on Wireless Communications*, vol. 14, pp. 6664–6678, 2015.

[12] S. P. Lloyd, "Least Squares Quantization in PCM," *IEEE Trans. on Inform. Theory*, vol. 28, pp. 129–137, 1982.

[13] J. Max, "Quantizing for Minimum Distortion," *IEEE Trans. on Inform. Theory*, vol. 6, pp. 7–12, 1960.

[14] Y. Linde, A. Buzo, and R. M. Gray, "An Algorithm for Vector Quantizer Design," *IEEE Trans. on Commun.*, vol. 28, pp. 84–95, 1980.

[15] G. Ball and D. Hall, "A Clustering Technique for Summarizing Multivariate Data," *Behavioral Science*, vol. 12, pp. 153–155, 1967.

[16] J. MacQueen, "Some Methods for Classification and Analysis of Multivariate Observations," *Proc. 5th Berkeley Symp. Math. Statistics and Probability*, vol. 1, pp. 281–297, 1967.

[17] J. C. Roh and B. D. Rao, "Transmit Beamforming in Multiple-Antenna Systems with Finite Rate Feedback: A VQ-Based Approach," *IEEE Trans. on Inform. Theory*, vol. 52, pp. 1101–1112, 2006.

[18] P. Xia and G. B. Giannakis, "Design and Analysis of Transmit-Beamforming based on Limited-Rate Feedback," *IEEE Trans. on Sig. Processing*, vol. 54, pp. 1853–1863, 2006.

[19] K. Rose, "Deterministic Annealing for Clustering, Compression, Classification, Regression, and Related Optimization Problems," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2210–2239, 1998.

[20] 3GPP, "Study on Channel Model for Frequency Spectrum Above 6 GHz," 3rd Generation Partnership Project (3GPP), TR 38.900 V15.0.0, Tech. Rep., 2018. [Online]. Available: http://www.3gpp.org/DynaReport/38900.htm

[21] A. Elshafiy and A. Sampath, "System Performance of Indoor Office Millimeter Wave Communications," in *IEEE Wireless Communications and Networking Conference*, April 2019.

[22] C. Balanis, *Antenna Theory, Analysis, and Design*, 3rd ed. New Jersey: Wiley, 2005.

[23] M. Cheng, J.-B. Wang, J.-Y. Wang, M. Lin, Y. Wu, and H. Zhu, "A Fast Beam Searching Scheme in mmWave Communications for High-Speed Trains," in *IEEE International Conf. on Communications*, May 2019.

[24] M. Giordani *et al.*, "Initial Access Frameworks for 3GPP NR at mmWave Frequencies," in *2018 17th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*, June 2018.

[25] D. Yang, L.-L. Yang, and L. Hanzo, "DFT-Based Beamforming Weight-Vector Codebook Design for Spatially Correlated Channels in the Unitary Precoding Aided Multiuser Downlink," in *IEEE International Conference on Communications*, May 2010.

[26] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images," *IEEE Trans. on Pattern Analysis Machine Intelligence*, vol. 6, pp. 721–741, 1984.

[27] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by Simulated Annealing," *Science*, vol. 220, pp. 671–680, 1983.

[28] A. Chakrabarti and H. Krishnaswamy, "High power, High Efficiency Stacked mmWave Class-E-like CMOS Power Amplifiers: Theory and Implementation," *IEEE Trans. on Microwave Theory and Techniques*, vol. 62, pp. 1686–1704, 2014.