Learning to Beamform for Minimum Outage

Yunmei Shi, Aritra Konar, *Member, IEEE*, Nicholas D. Sidiropoulos, *Fellow, IEEE*,
Xing-Peng Mao, *Member, IEEE*, and Yong-Tan Liu, *Senior Member, IEEE*

***Abstract*—Acquiring channel state information (CSI) at the base station (BS) is a critical requirement for successfully employing transmit beamforming in multiantenna systems. In practice, channel estimation/quantization errors, feedback delays, and fast fading can make it difficult to obtain accurate CSI at the BS. In this paper, we consider an outage-based approach for transmit beamforming in order to deal with the channel uncertainty at the BS. Our formulation is applicable to both point-to-point transmit beamforming as well as single-group multicasting scenarios. A key difference from prior works is that we do not assume knowledge of the underlying channel distribution; instead, stochastic approximation is used for computing approximate solutions of a nonconvex stochastic optimization problem via simple first-order methods (FOMs). We evaluate the performance of our FOMs in two settings: First) where we design a beamformer at the BS based on historical channel realizations collected over a relatively long time window before deployment, and second) in a post-deployment phase where we perform incremental updates of our beamformer based on intermittent, delayed, or peer feedback. Simulation results reveal the effectiveness of FOMs for our problem compared to other alternatives.**

***Index Terms*—Downlink beamforming, outage minimization, stochastic approximation, stochastic gradient methods.**

I. Introduction

Linear transmit beamforming is a simple but potent technique for enhancing the throughput of multiple-input multiple-output (MIMO) systems by exploiting channel state information (CSI) at the base station (BS) [2]–[7]. In practical wireless systems, acquiring accurate downlink channel estimation at the BS in a timely manner can be a challenging task. This is particularly the case in systems equipped with large transmit antenna arrays, where the overhead in channel estimation can be significant. Reducing the burden associated with CSI acquisition in multi-antenna systems is currently an active area of research [8]–[16].

Considering time-division duplex (TDD) systems, where uplink and downlink channels occupy the same frequency band but operate in different time slots, channel reciprocity is typically assumed in order to obtain downlink CSI via uplink channel estimation. However, reciprocity is only an approximation, which inevitably introduces errors in the channel estimates. Furthermore, the downlink estimate obtained at the BS is based upon the uplink channel in the previous time slot, which gives rise to errors in high-mobility scenarios. Meanwhile, in frequency-division duplex (FDD) systems, users estimate the downlink channel and provide low-rate quantized feedback to the BS, which again results in errors. Feedback delays arising in high mobility scenarios are an additional concern.

In order to mitigate the performance degradation stemming from imperfect CSI, the design of robust approaches which take into account channel uncertainty is well motivated and has received considerable attention. Depending on the channel uncertainty model and robust performance metric employed, robust designs for downlink transmit beamforming can be broadly classified into the following categories: i) the channel model is described by a set of nominal channel vectors subject to unknown, bounded perturbations, and the performance metric used aims to minimize the worst-case Quality-of-Service (QoS) with respect to (w.r.t.) all possible perturbations [7], [17]–[24]; ii) the channel uncertainties are modeled as random variables, and the expected value of QoS (expectation taken w.r.t. the distribution) is used as the performance metric; and iii) outage based approaches [25]–[36], where the metric corresponds to the actual QoS satisfying a desired performance threshold with high probability. The worst-case approach in i) may result in a very conservative design compared to ii), which can yield, on average, higher throughput. The drawback of the latter approach is that it can perform very poorly in instances of persistent deep fading, which makes it unsuitable for use in delay-sensitive applications. Outage based approaches strike a favorable balance between the first two, with the system designer being able to vary the level of pessimism by changing the outage threshold.

In this paper, we adopt an outage based design approach for transmit beamforming. Our formulation can be applied to describe the following two downlink transmission scenarios: a

Manuscript received September 29, 2017; revised March 26, 2018, June 11, 2018, and August 6, 2018; accepted August 6, 2018. Date of publication August 14, 2018; date of current version September 5, 2018. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Marios Kountouris. This work of Y. Shi was supported in part by the China Scholarship Council while a visiting exchange student with the Department of Electrical and Computer Engineering, University of Minnesota, from Fall 2016 to Summer 2017, and the work of A. Konar was supported in part by a Doctoral Dissertation Fellowship from the University of Minnesota. The work of A. Konar and N. Sidiropoulos was also supported in part by NSF CIF-1525194. This paper was presented in part at the 7th IEEE International Workshop on Computational Advances in MultiSensor Adaptive Processing, Dutch Antilles, Dec. 10–13, 2017. *(Corresponding author: Nicholas D. Sidiropoulos.)*

Y. Shi was a visiting student with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA. She is now with the Department of Electronics and Instrumentation Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: shiyunmeidamei@hit.edu.cn).

A. Konar and N. D. Sidiropoulos were with the Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455 USA. They are now with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA 22903 USA (e-mail: aritra@virginia.edu; nikos@virginia.edu).

X.-P. Mao and Y.-T. Liu are with the Department of Electronics and Instrumentation Engineering, Harbin Institute of Technology, Harbin 150001, China (e-mail: mxp@hit.edu.cn; liuyt@hit.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2018.2865408

point-to-point multiple-input single-output (MISO) setting and a multi-user MISO setting where the users form a single multicast group. For the purposes of exposition, we explain our approach from the perspective of the former setting, with relevant remarks for the multicasting case where necessary.[1] Outage based MISO beamforming has applications in satellite and military communications [38], radar settings [39], and is also a potential component of ultra-reliable low latency communication (URLLC) systems in future 5G networks [40].

Our design formulation entails minimizing the outage probability (i.e., the probability that the received SNR falls below a certain threshold) subject to transmit power constraints. Such a model was previously considered in [30], where the channel vectors were modeled as being drawn from a Gaussian mixture model (GMM), whose parameters are perfectly known at the BS. In a major departure from this work[2], in the present paper, we do not assume knowledge of the channel distribution. Rather than fitting a model for the channel distribution first, here we seek to minimize outage *directly* from available channel realizations without using any prior knowledge. Using the fact that the probability of an event can be equivalently expressed as the expectation of the indicator function of said event, we equivalently reformulate our problem as a stochastic optimization (SO) problem. Given the lack of knowledge of the underlying channel distribution, we aim to bring to bear tools from stochastic approximation [41]–[43] upon our problem. Additional examples of stochastic optimization approaches in the wireless communication literature include [37] which deals with optimization of long-term ergodic rates under queue stability constraints; and [36], which considers minimizing power subject to outage constraints in a multiuser interference setting with single transmit and receive antennas. [36] uses channel realizations to construct many instantaneous inequalities in lieu of the outage constraints, the idea being that if one satisfies all these inequalities then one will avoid outage with a high probability. As the instantaneous inequalities are linear in the link powers, the resulting problem is convex[3].

A formidable challenge in pursuing our approach, which is geared towards approximating the outage *cost function*, arises from the fact that the indicator function of the desired event is non-convex and discontinuous, which prevents direct application of algorithmic techniques developed for stochastic approximation. In order to circumvent this issue, we design two judicious smooth approximations of the indicator function, which result in two surrogate formulations which are amenable to stochastic gradient type methods. Both formulations are nonconvex however, which still makes it challenging to obtain high quality sub-optimal solutions in polynomial-time. We consider computing such approximate solutions for our problems in the context of the following two settings. In the first setting, we assume that the BS has collected a set of channel samples before deployment, using which we construct a sample average approximation of our SO problems, followed by application of simple first-order methods (FOMs) for efficiently computing solutions. In this case, no CSI is required at the BS post-deployment. We also consider an alternate scenario corresponding to the post-deployment phase, where assuming the availability of intermittent, possibly outdated channel estimates at the BS provided by the user or even other 'peer' users (with the same channel distribution), we use streaming FOMs to compute solutions for the SO problems in an online fashion.

As a baseline for comparison against the FOM based approach, we also devise a set of algorithms based on an approximation of the cost function of the original SO problem using Markov's inequality. While this approach was originally proposed in [30] for the Gaussian mixture distribution, here we extend it to the case where the channel distribution is unknown for both of the aforementioned settings. An extensive set of simulations is then carried out to compare the performance of all methods in various settings.

Relative to the conference version [1] which only describes the online setting, the journal version adds the offline case, a detailed discussion on the convergence of the FOMs, additional experiments for both offline and online cases, while also featuring more comprehensive exposition.

The rest of the paper is organized as follows. Section II contains a detailed description of our formulation and outlines our proposed approach. Section III provides a description of the FOMs we use for computing solutions of our SO problems, while Section IV describes the algorithms based on approximation via Markov's inequality. Experimental results are provided in Section V and conclusions are drawn in Section VI.

The following notations are used throughout the paper. Matrices and vectors are represented by bold upper-case and bold lower-case characters, respectively, while calligraphic notation is used to denote sets. Superscripts $(\cdot)^T$ and $(\cdot)^H$ stand for the transpose and conjugate transpose, respectively. The shorthand $[M]$ is used for the set $\{1, 2, \cdots, M\}$. We denote the N-dimensional real and complex Euclidean space by $\mathbb{R}^N$ and $\mathbb{C}^N$, respectively. The operators $\Re[\cdot]$ and $\Im[\cdot]$ denote the real and imaginary parts of a complex number, respectively, and $\lceil\cdot\rceil$ stands for the ceiling operator. $\|\cdot\|_2$ and $|\cdot|$ represent the Euclidean norm and absolute value, respectively, while $\mathbb{E}[\cdot]$ denotes the mathematical expectation operator. The set of natural numbers is denoted by $\mathbb{N}$.

II. Model and Problem Formulation

A. Model

We consider a single cell, MISO downlink beamforming scenario, where a BS equipped with N transmit antennas serves a single user with one receive antenna. The downlink channel can be described as

$$y = \mathbf{h}^H \mathbf{w} s + n \tag{1}$$

where $y \in \mathbb{C}$ is the received signal at the user, $\mathbf{h} \in \mathbb{C}^N$ denotes the instantaneous channel vector, and $\mathbf{w} \in \mathbb{C}^N$ is the beamformer employed by the BS. In addition, $s \in \mathbb{C}$ is the transmitted signal satisfying $\mathbb{E}\{|s|^2\} = 1$ and $n \sim \mathcal{CN}(0, 1)$ represents

[1] The multicasting perspective has been explained at length in [30].

[2] *and* from the majority of previously considered outage based approaches.

[3] In our case, the approach of [36] would lead to nonconvex instantaneous inequalities.

complex, circularly symmetric, Gaussian noise of zero mean and unit variance.

Assuming the availability of perfect, instantaneous CSI at the BS, the beamformer which maximizes the instantaneous received SNR subject to a transmit sum power constraint can be computed as

$$\mathbf{w}^* = \arg \max_{\|\mathbf{w}\|_2^2 \leq P} |\mathbf{w}^H \mathbf{h}|^2 = \sqrt{P}(\mathbf{h}/\|\mathbf{h}\|_2) \tag{2}$$

In this paper, we completely forego the assumption of the availability of instantaneous CSIT. While it is possible, in principle, to estimate the downlink channel at the BS using estimation techniques based upon either FDD or TDD modes of operation, this comes at the expense of significant signaling and system hardware overhead. Moreover, in practice, as a result of estimation errors and feedback delays, exact CSI cannot be acquired at the BS. If, however, the overhead associated with acquiring CSI in a timely fashion is affordable at the BS, and the channel varies slowly, then there is no need to consider a different criterion. On the other hand, situations may arise where the channel changes abruptly and cannot be tracked by the BS. This is the case, for example, when the receiver suddenly turns a corner, or enters an elevator, or in high mobility / Doppler scenarios, or in intermittent communication scenarios (arising in IoT settings), when phase/carrier/Doppler synchronization is suddenly lost. Another example where the channel can change abruptly arises at the boundary between cells, where a user enters a cell, wanders over to the other, and returns again to the initial cell. In such cases, it is undesirable to employ a beamformer design criterion based on instantaneous SNR.

In order to account for such situations, we employ a probabilistic model where the temporal variations of the downlink channel $\{\mathbf{h}_t\}_{t=0}^{\infty}$ correspond to different realizations drawn from an unknown underlying distribution; i.e., we model $\mathbf{h}$ as a random vector defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega \subseteq \mathbb{C}^N$ denotes the sample space, $\mathcal{F}$ is a collection of subsets of Ω forming a $\sigma-$ field, and $\mathbb{P}$ is an unknown probability measure defined on $\mathcal{F}$. Independence of the channel realizations can be assumed although this is not necessary. As we show later on, this condition can be weakened provided the channel process satisfies appropriate ergodic mixing conditions. Under this model, we propose to formulate our design criterion based on minimizing the outage probability

$$\min_{\mathbf{w} \in \mathcal{W}} \left\{ F(\mathbf{w}) := \Pr\left(|\mathbf{w}^H \mathbf{h}|^2 < \gamma \right) \right\} \tag{3}$$

where $\gamma > 0$ denotes the outage threshold and $\mathcal{W} \subset \mathbb{C}^N$ is a simple[4], convex, compact set. Particular examples of $\mathcal{W}$ include sum power constraints (SPCs) and per-antenna constraints (PACs). For PACs, the constraint set $\mathcal{W}$ is defined as

$$\mathcal{W} = \{\mathbf{w} \in \mathbb{C}^N \,|\, |w(n)|^2 \leq P_n, \forall\, n \in [N]\} \tag{4}$$

Under ergodic mixing conditions on the channel process, minimizing outage approximately maximizes the fraction of time the downlink channel satisfies a desired SNR requirement.

Clearly, (3) is a more conservative design criterion compared to (2). However, in the special case where the distribution is known *a priori* at the BS, we point out that no instantaneous CSIT is required to design $\mathbf{w}$.

An important consideration in outage minimization is the choice of distribution employed to describe the channel uncertainty. A popular choice is to use a complex, circularly symmetric, Gaussian distribution; i.e., $\mathbf{h} \sim \mathcal{CN}(\mathbf{m}, \mathbf{C})$. This is a model that is relevant as long as the large scale channel parameters remain constant; i.e., for time scales on the order of seconds. In order to perform outage minimization with this particular model, we refer the reader to Appendix A. For channels which can change abruptly due to situations mentioned earlier, a multi-modal distribution is a better description of the channel uncertainty. The main focus of this manuscript is on performing outage minimization with such distributions, which are valid for longer time scales. In prior work [30], a minimum outage criterion was proposed where the distribution was modeled as a GMM whose parameters are known perfectly at the BS. In this case, (3) admits the following interpretation: assuming J is the number of kernels of the GMM, there are J possible channel states where each state $j \in [J]$ is associated with a certain Gaussian distribution of the channel vector and occurs with a certain probability π_j, where $\pi_j \geq 0, \forall\, j \in [J]$ and $\sum_{j=1}^{J} \pi_j = 1$.

Remark 1: We briefly describe the same outage minimization problem from the perspective of single-group multicasting as detailed in [30]. Consider a scenario where a BS with N antennas transmits a common information-bearing signal to a group of K single receive antenna subscribers. In this context, the outage formulation (3) can be motivated by the fact that ideally, a subscriber should be able to join or leave the multicast group without notifying the BS. Hence, the BS has to operate without knowing the users' instantaneous channels as well as the number of users participating in the current multicast session. If a new subscriber wishes to join the current group, then employing the criterion (3) maximizes the probability that he/she will be served. Here again, the subscriber channel vectors $\{\mathbf{h}_k\}_{k=1}^{K}$ are modeled as random vectors drawn from an underlying distribution; i.e., all subscriber channels are assumed to be statistically equivalent. In multicasting, subscribers are often geographically clustered together in one of several service areas (e.g., town squares, malls, campuses, etc.), which naturally motivates using a multi-modal distribution like a GMM to model the channel uncertainty. Furthermore, if a large number of realizations are drawn from this distribution, then minimizing the outage probability approximately corresponds to maximizing the fraction of users that will be served.

B. Proposed Approach

Given channel realizations (e.g., historical data from measurement campaigns), it is theoretically possible to accurately approximate any distribution via a GMM. However this approach requires the BS to identify the parameters of the model, which cannot be guaranteed using practical algorithms like Expectation-Maximization (EM). Furthermore, even when the parameters are perfectly known, algorithmic approaches for (3)

[4] By simple, we mean that the Euclidean projection of a point onto $\mathcal{W}$ is easy to compute.

may require evaluation of computationally intensive integrals associated with the cost function and its higher-order derivatives. Indeed, in [30], algorithmic solutions are proposed and tested only for specific instances of the GMM.

In order to circumvent the aforementioned issues, in this work, we do not explicitly assume (or attempt to fit) a GMM to describe (approximate) the underlying distribution. Instead, we rely upon stochastic approximation for obtaining solutions of (3) via simple iterative methods based on available channel realizations; i.e., we adopt a *data-driven* approach for designing a beamformer that minimizes outage. More precisely, we consider the following two settings.

First, consider a scenario where a collection of channel realizations $\mathcal{H}_T := \{\mathbf{h}_t\}_{t=1}^T$ is made available at the BS via measurement campaigns conducted over a certain time period before deployment. In the absence of the distributional information of $\mathbf{h}$, we propose to use the set of sample realizations $\mathcal{H}_T$ to approximate the cost function of (3). Towards this end, note that we can equivalently express (3) as

$$\min_{\mathbf{w}\in\mathcal{W}} \Pr\left(|\mathbf{w}^H\mathbf{h}|^2 < \gamma\right) \Leftrightarrow \min_{\mathbf{w}\in\mathcal{W}} \mathbb{E}_{\mathbf{h}}[\mathbb{1}_{\{|\mathbf{w}^H\mathbf{h}|^2<\gamma\}}] \tag{5}$$

Define

$$f(\mathbf{w};\mathbf{h}) := \mathbb{1}_{\{|\mathbf{w}^H\mathbf{h}|^2<\gamma\}} = \begin{cases} 1, \text{ if } |\mathbf{w}^H\mathbf{h}|^2 < \gamma \\ 0, \text{ otherwise} \end{cases} \tag{6}$$

as the indicator function of the event $|\mathbf{w}^H\mathbf{h}|^2 < \gamma$. Utilizing $\mathcal{H}_T$, we construct the following sample average estimate of $\mathbb{E}_{\mathbf{h}}[f(\mathbf{w};\mathbf{h})]$

$$\hat{F}(\mathbf{w};\mathcal{H}_T) := \frac{1}{T}\sum_{t=1}^{T}\left\{f(\mathbf{w};\mathbf{h}_t) := \mathbb{1}_{\{|\mathbf{w}^H\mathbf{h}_t|^2<\gamma\}}\right\} \tag{7}$$

Thus, given access to a window of channel realizations $\{\mathbf{h}_t\}_{t=1}^T$ alone (but not the distribution), the interpretation is that we minimize the total number of outages over ("recent") channel "history" - which is very reasonable when the channel can change abruptly between different states.

It is worthwhile to point out that if the channel is constant, then all channel samples will be the same, and minimizing the term on the right hand side, e.g., under a sum power constraint, will reduce to matched filtering to the fixed channel. If the channel is changing slowly, then, keeping a short window of samples in the stochastic approximation of the cost function on the right hand side above, the solution will be naturally close to the latest channel, but it will in fact take more than first and second order channel statistics into account, which is a good thing. If the channel does change more rapidly, we will obtain a beamforming vector that minimizes the number of Quality-of-Service (QoS) violations for the channel history considered. It is up to the designer to decide what is the time window used (i.e., of the order of minutes, hours, or days) to collect and use channel data. If that window is long, then indeed what one minimizes is long-term outage, which is not appropriate for slowly varying channels; but it is appropriate for rapidly varying ones, which are hard to track instantaneously.

Under appropriate ergodic mixing conditions on the channel process [44, p. 171], we have

$$\lim_{T\to\infty} \hat{F}(\mathbf{w};\mathcal{H}_T) = \mathbb{E}_{\mathbf{h}}[f(\mathbf{w};\mathbf{h})] = F(\mathbf{w}), \forall\ \mathbf{w} \in \mathcal{W} \tag{8}$$

almost surely; i.e., sample averages converge to ensemble averages with probability (w.p.) 1. On replacing $F(\mathbf{w})$ by $\hat{F}(\mathbf{w};\mathcal{H}_T)$ in (3), we obtain the problem

$$\min_{\mathbf{w}\in\mathcal{W}} \hat{F}(\mathbf{w};\mathcal{H}_T) \tag{9}$$

which corresponds to an empirical approximation of its ensemble counterpart (5).

Remark 2: Stronger results pertaining to convergence of the optimal value and optimal solutions of (9) to their counterparts of (5) can be obtained under additional assumptions. Applying [45, Proposition 2.1], we have that the functions $F(\mathbf{w})$ and $\hat{F}(\mathbf{w};\mathcal{H}_T)$ are lower semicontinuous. Assuming that the channel realizations $\mathcal{H}_T$ are independent, and, for every $\mathbf{w} \in \mathcal{W}$, we have $|\mathbf{w}^H\mathbf{h}|^2 \neq \gamma$ w.p. 1, then [45, Proposition 2.1] asserts that $F(\mathbf{w})$ is continuous and $\hat{F}(\mathbf{w},\mathcal{H}_T)$ converges to $F(\mathbf{w})$ uniformly w.p. 1 on $\mathcal{W}$. Under some additional technical conditions listed in [46, Theorem 2.3], uniform convergence is sufficient to guarantee that i) the optimal value, and ii) the set of optimal solutions of (9) converge to their counterparts of (5) as $T \to \infty$.

Note that the above scenario entails solving a batch problem in an offline setting; i.e., the samples $\mathcal{H}_T$ have to be collected and stored at the BS before deployment in order to design $\mathbf{w}$. While such a setup alleviates the need for instantaneous CSIT, in practice, in the post-deployment phase, assuming the user provides intermittent feedback to the BS regarding his/her estimate of the downlink channel, it is desirable to develop approaches which can exploit this information for computing a solution of (3) in an *online* fashion. In this setting, we seek to minimize $F(\mathbf{w})$ by updating $\mathbf{w}$ as channel estimates arrive at the BS in a streaming fashion. Note that our minimum outage criterion can naturally cope with intermittent/delayed feedback from the user due to our assumption that all channels are drawn from the same underlying distribution, and hence, are statistically equivalent. As a result, feedback requirements are considerably relaxed compared to other setups reliant on instantaneous CSIT. Furthermore, online algorithms are appealing for their simplicity and low storage requirements since computing updates only requires storing the most recent channel estimate; i.e., updates are only based on $f(\mathbf{w};\mathbf{h}_t)$, where t denotes the streaming index. Consequently, such a setting is well suited for large-scale antenna systems operating in FDD [5], where N is potentially large and feedback delays are inevitable. We also point out that it is not necessary to receive feedback from the same user; feedback from different 'peer' users with the same channel distribution as the user of interest can also be utilized (due to statistical equivalence). This becomes useful in situations where channel estimates are stale (due to fast channel variation relative to estimation and feedback delays to get the estimate back to the transmitter), or intermittent, or unavailable

[5] This is indeed the case in most practical systems.

from a particular user of interest–but may be available from peer users.[6]

Before further elaborating upon our algorithmic approaches for the outlined settings, we discuss the computational tractability of (3). In [30], under the GMM assumption and using sum-power constraints (SPCs), it was established that (3) is NP–Hard in the worst-case when $J > N$. Although we resort to stochastic approximation in this work, we point out that irrespective of the distribution of $\mathbf{h}$, the indicator function $f(\mathbf{w};\mathbf{h})$ is always non-convex *and* discontinuous, which poses a formidable challenge to our approach. Indeed, the analysis of most stochastic optimization algorithms are based on the assumption that $f(\mathbf{w};\mathbf{h})$ is continuous in $\mathbf{w}$ for every $\mathbf{h}$. We partially address this problem by constructing smooth approximations of $f(\mathbf{w};\mathbf{h})$, as we explain in the following section. While one may question the merit of using tools developed for continuous optimization for a problem which is not continuous, we demonstrate that applying these simple algorithms on smoothed surrogates of the indicator function can perform surprisingly well.

C. Smooth Surrogates for the Indicator Function

First, since $f(\mathbf{w};\mathbf{h})$ is a real function of complex variables, we can equivalently express it in terms of real variables as

$$f(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) := \mathbb{1}_{\{\|\tilde{\mathbf{H}}^T\tilde{\mathbf{w}}\|_2^2<\gamma\}} \tag{10}$$

where we have used the definitions $\tilde{\mathbf{w}} := [\Re[\mathbf{w}]^T, \Im[\mathbf{w}]^T]^T \in \mathbb{R}^{2N}, \tilde{\mathbf{h}} := [\Re[\mathbf{h}]^T, \Im[\mathbf{h}]^T]^T \in \mathbb{R}^{2N}$ and

$$\tilde{\mathbf{H}} := \begin{bmatrix} \Re[\mathbf{h}] & \Im[\mathbf{h}] \\ \Im[\mathbf{h}] & -\Re[\mathbf{h}] \end{bmatrix} \in \mathbb{R}^{2N\times 2}$$

We now consider the following smooth surrogates of $f(\tilde{\mathbf{w}};\tilde{\mathbf{h}})$.

1) *Sigmoid Approximation:* Consider the function

$$u(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) := \frac{1}{1+\exp\left(\|\tilde{\mathbf{H}}^T\tilde{\mathbf{w}}\|_2^2-\gamma\right)} \tag{11}$$

Note that when $\|\tilde{\mathbf{H}}^T\tilde{\mathbf{w}}\|_2^2 << \gamma$, $u(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) \approx 1$ and when $\|\tilde{\mathbf{H}}^T\tilde{\mathbf{w}}\|_2^2 >> \gamma$, $u(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) \approx 0$. Hence, the function $u(\tilde{\mathbf{w}};\tilde{\mathbf{h}})$ serves as a continuous approximation of $f(\tilde{\mathbf{w}};\tilde{\mathbf{h}})$. Furthermore, $u(\tilde{\mathbf{w}};\tilde{\mathbf{h}})$ is continuously differentiable with gradient

$$\nabla_{\tilde{\mathbf{w}}} u(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) = -\frac{2\exp\left(\|\tilde{\mathbf{H}}^T\tilde{\mathbf{w}}\|_2^2-\gamma\right)\tilde{\mathbf{H}}_i\tilde{\mathbf{H}}_i^T\tilde{\mathbf{w}}}{\left[1+\exp\left(\|\tilde{\mathbf{H}}^T\tilde{\mathbf{w}}\|_2^2-\gamma\right)\right]^2} \tag{12}$$

2) *Smoothed Point-Wise Maximum Approximation:* Consider the point-wise maximum (PWM) function

$$v(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) := \max\left\{0, 1-\frac{\|\tilde{\mathbf{H}}^T\tilde{\mathbf{w}}\|_2^2}{\gamma}\right\} \tag{13}$$

Again, for $\|\tilde{\mathbf{H}}^T\tilde{\mathbf{w}}\|_2^2 << \gamma$, $v(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) \approx 1$ and when $\|\tilde{\mathbf{H}}^T\tilde{\mathbf{w}}\|_2^2 >> \gamma$, $v(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) = 0$. While $v(\tilde{\mathbf{w}};\tilde{\mathbf{h}})$ can serve as a continuous approximation of $f(\tilde{\mathbf{w}};\tilde{\mathbf{h}})$, it is non-differentiable in $\tilde{\mathbf{w}}$ in contrast to (11). Hence, we propose to construct a differentiable approximation of $v(\tilde{\mathbf{w}};\tilde{\mathbf{h}})$ using Nesterov's smoothing approach [48]. Towards this end, note that $v(\tilde{\mathbf{w}};\tilde{\mathbf{h}})$ admits the following equivalent representation

$$v(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) = \max_{0\leq y\leq 1}\left\{y\left(1-\frac{\|\tilde{\mathbf{H}}^T\tilde{\mathbf{w}}\|_2^2}{\gamma}\right)\right\} \tag{14}$$

This step allows us to construct a smooth surrogate of (13) by applying the following modification to (14)

$$v^{(\mu)}(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) = \max_{0\leq y\leq 1}\left\{y\left(1-\frac{\|\tilde{\mathbf{H}}^T\tilde{\mathbf{w}}\|_2^2}{\gamma}\right)-\frac{\mu}{2}y^2\right\} \tag{15}$$

where $\mu > 0$ is a smoothing parameter. Note that for a given $\tilde{\mathbf{w}}$, the maximization problem is strongly concave in y over the unit interval, and thus admits a unique solution which can be computed in closed form. Using the definition

$$g(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) := 1-\frac{\|\tilde{\mathbf{H}}^T\tilde{\mathbf{w}}\|_2^2}{\gamma} \tag{16}$$

allows us to ultimately represent $v^{(\mu)}(\tilde{\mathbf{w}};\tilde{\mathbf{h}})$ in the following form

$$v^{(\mu)}(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) = \begin{cases} 0, & g(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) < 0 \\ \dfrac{1}{2\mu}\left(g(\tilde{\mathbf{w}};\tilde{\mathbf{h}})\right)^2, & 0 \leq g(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) \leq \mu \\ g(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) - \dfrac{\mu}{2}, & g(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) > \mu \end{cases} \tag{17}$$

Note that (17) is continuously differentiable in $\tilde{\mathbf{w}}$ with derivatives given by

$$\nabla_{\tilde{\mathbf{w}}} v^{(\mu)}(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) = \begin{cases} 0, & g(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) < 0 \\ \dfrac{g(\tilde{\mathbf{w}};\tilde{\mathbf{h}})}{\mu}\nabla_{\tilde{\mathbf{w}}} g(\tilde{\mathbf{w}};\tilde{\mathbf{h}}), & 0 \leq g(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) \leq \mu \\ \nabla_{\tilde{\mathbf{w}}} g(\tilde{\mathbf{w}};\tilde{\mathbf{h}}), & g(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) > \mu \end{cases} \tag{18}$$

where

$$\nabla_{\tilde{\mathbf{w}}} g(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) := -\frac{2}{\gamma}\tilde{\mathbf{H}}\tilde{\mathbf{H}}^T\tilde{\mathbf{w}} \tag{19}$$

In addition, regarding the quality of approximation, it is possible to establish the following bounds [48, p. 132]

$$v^{(\mu)}(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) \leq v(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) \leq v^{(\mu)}(\tilde{\mathbf{w}};\tilde{\mathbf{h}}) + \frac{\mu}{2}, \forall\, (\tilde{\mathbf{w}};\tilde{\mathbf{h}}) \in \tilde{\mathcal{W}} \times \tilde{\mathcal{F}} \tag{20}$$

where $\tilde{\mathcal{W}}, \tilde{\mathcal{F}}$ denote the representation of the sets $\mathcal{W}, \mathcal{F}$ in terms of real variables, respectively. Note that (20) implies that our approximation is tight up to an additive factor in μ. Thus, the quality of approximation can be improved by reducing μ.

In general, it is hard to theoretically quantify the quality of approximation obtained by using the sigmoid and smoothed-PWM functions as surrogates for the indicator function. To aid intuition regarding the strengths/weaknesses these choices of surrogate functions offer, we provide an illustrative plot in Fig. 1. Note that using the sigmoid function results in a overall

[6] This idea is very similar to the idea of collaborative filtering in recommender systems [47], where the preferences of the user of interest is inferred from the preferences of peer users.

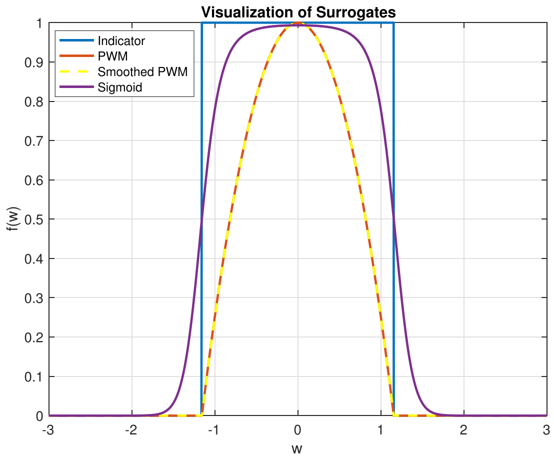

Fig. 1. Illustrative plot of surrogates for the indicator function $f(w) = 1_{\{|w\,h|<\sqrt{\gamma}\}}$ for $w \in \mathbb{R}$, $h \sim \mathcal{N}(1,1)$, $\gamma = 5$, $\mu = 1e^{-3}$.

better approximation compared to the smoothed-PWM function. However, it is precisely because of this fact that makes the sigmoid function comparatively harder to minimize using tools from continuous optimization (as it more closely resembles a non-smooth function).

D. Problem Formulation

Using the surrogates defined in the prior section, we propose to employ the following smooth approximations of the original SO formulation (5)

$$\min_{\tilde{\mathbf{w}}\in\tilde{\mathcal{W}}}\left\{U(\tilde{\mathbf{w}}) := \mathbb{E}_{\tilde{\mathbf{h}}}[u(\tilde{\mathbf{w}};\tilde{\mathbf{h}})]\right\} \tag{21a}$$

$$\min_{\tilde{\mathbf{w}}\in\tilde{\mathcal{W}}}\left\{V^{(\mu)}(\tilde{\mathbf{w}}) := \mathbb{E}_{\tilde{\mathbf{h}}}[v^{(\mu)}(\tilde{\mathbf{w}};\tilde{\mathbf{h}})]\right\} \tag{21b}$$

corresponding to the sigmoid and smoothed PWM approximations respectively. Note that both formulations are non-convex, and hence in general, it is not possible to solve them optimally in polynomial-time. Here, we are interested in computing high-quality sub-optimal solutions using simple algorithms in both offline and online settings.

1) *Offline Setting:* Given $\mathcal{H}_T$, we obtain the following finite-sample approximations of (21a) and (21b)

$$\min_{\tilde{\mathbf{w}}\in\tilde{\mathcal{W}}}\left\{\hat{U}(\tilde{\mathbf{w}};\mathcal{H}_T) := \frac{1}{T}\sum_{t=1}^{T} u(\tilde{\mathbf{w}};\tilde{\mathbf{h}}_t)\right\} \tag{22a}$$

$$\min_{\tilde{\mathbf{w}}\in\tilde{\mathcal{W}}}\left\{\hat{V}^{(\mu)}(\tilde{\mathbf{w}};\mathcal{H}_T) := \frac{1}{T}\sum_{t=1}^{T} v^{(\mu)}(\tilde{\mathbf{w}};\tilde{\mathbf{h}}_t)\right\} \tag{22b}$$

Remark 2 applies here for (22a) and (22b) as well. Note that these problems correspond to minimizing finite sums of smooth functions. Consequently, they are well suited for application of FOMs.

2) *Online Setting:* In this case, channel estimates $\{\tilde{\mathbf{h}}_t\}_{t=0}^{\infty}$ arrive at the BS in a streaming fashion. While our goal is to minimize $U(\tilde{\mathbf{w}})$ and $V^{(\mu)}(\tilde{\mathbf{w}})$, for each random vector $\tilde{\mathbf{h}}_t$, we only have access to the random cost functions $u(\tilde{\mathbf{w}};\tilde{\mathbf{h}}_t)$ and $v^{(\mu)}(\tilde{\mathbf{w}};\tilde{\mathbf{h}}_t)$. We will utilize online FOMs to compute efficient incremental updates of $\tilde{\mathbf{w}}$ based on instantaneous gradient estimates of the ensemble gradients $\nabla_{\tilde{\mathbf{w}}}U(\tilde{\mathbf{w}})$ and $\nabla_{\tilde{\mathbf{w}}}V^{(\mu)}(\tilde{\mathbf{w}})$.

The exact algorithms we employ are outlined in the next section.

III. First-order Methods for Minimum Outage

In this section, we provide a brief overview of the various FOMs utilized for both offline and online settings.

A. Offline Setting

Note that (22a) and (22b) can be abstracted via the following representative problem, where given a batch of d-dimensional data samples $\{\boldsymbol{\xi}_t\}_{t=1}^{T}$ drawn from an unknown probability distribution with support set $\boldsymbol{\Xi} \subset \mathbb{R}^d$, we aim to minimize the finite-sample surrogate

$$\min_{\mathbf{x}\in\mathcal{X}}\left\{F(\mathbf{x}) := \frac{1}{T}\sum_{t=1}^{T} f_t(\mathbf{x})\right\} \tag{23}$$

where we define $f_t(\mathbf{x}) := f(\mathbf{x};\boldsymbol{\xi}_t)$ for ease of notation. It is assumed that $\mathcal{X} \subset \mathbb{R}^d$ is a convex, compact set and each $f_t : \mathcal{X} \times \boldsymbol{\Xi} \to \mathbb{R}$ is a twice differentiable, non-convex function $\forall\, t \in [T]$ with $L-$ Lipschitz continuous gradients:

$$\|\nabla f_t(\mathbf{x}_1) - \nabla f_t(\mathbf{x}_2)\|_2 \le L\|\mathbf{x}_1 - \mathbf{x}_2\|_2, \forall\, \mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}. \tag{24}$$

This implies that the function $F(\mathbf{x})$ is also $L-$ smooth, since smoothness is preserved under convex combinations. A standard method to determine an approximate solution for (23) is gradient descent (GD), which can be described by the following update rule

$$\mathbf{y}^{(k+1)} = \mathbf{x}^{(k)} - \frac{\alpha_k}{T}\sum_{t=1}^{T}\nabla f_t(\mathbf{x}^k) \tag{25a}$$

$$\mathbf{x}^{(k+1)} = \Pi_{\mathcal{X}}(\mathbf{y}^{(k+1)}), \forall\, k \in \mathbb{N} \tag{25b}$$

where $\Pi_{\mathcal{X}}(.)$ represents the Euclidean projection operator onto $\mathcal{X}$ and $\alpha_k > 0$ is the step-size in the k-th iteration. Note that at each step, GD requires evaluation of T gradients, which is expensive for large T. A popular modification which reduces complexity is stochastic gradient descent (SGD). In this algorithm, at each iteration k, an index t_k is drawn uniformly at random from the index set $[T]$, resulting in the update

$$\mathbf{y}^{(k+1)} = \mathbf{x}^{(k)} - \alpha_k \nabla f_{t_k}(\mathbf{x}^k) \tag{26a}$$

$$\mathbf{x}^{(k+1)} = \Pi_{\mathcal{X}}(\mathbf{y}^{(k+1)}), \forall\, k \in \mathbb{N} \tag{26b}$$

Note that the expectation $\mathbb{E}_{t_k}[\mathbf{y}^{(k+1)}|\mathbf{x}^{(k)}]$ in (26a) is identical to (25a), which implies that the updates of SGD are equal to that of GD in expectation. As each iteration of SGD only requires computing single derivative $\nabla f_{t_k}(\mathbf{x}^{(k)})$, this results in a $O(T)$-factor improvement in per-iteration complexity compared to GD.

Aiming to combine the best of both GD and SGD, a hybrid algorithm termed stochastic variance reduced gradient (SVRG)

Algorithm 1: SVRG.

```
Initialization: Initialize $\mathbf{y}_1 \in \mathcal{X}$, set number of stages $S$,
update frequency $K$, and the step-size sequence
$\{\alpha_s^{(k)}\}_{s\in[S],k\in[K]}$
Iterate: for $s = 1, \cdots, S$
    • Compute $\mathbf{g}_s := \nabla F(\mathbf{y}_s)$
    • Set $\mathbf{x}_s^{(1)} = \mathbf{y}_s$
    • Iterate: for $k = 1, \cdots, K$ Choose $t_k \in [T]$ uniformly
      at random and update
        $\mathbf{v}_s^{(k+1)} = \mathbf{x}_s^{(k)} - \alpha_s^{(k)}[\nabla f_{t_k}(\mathbf{x}_s^{(k)}) - \nabla f_{t_k}(\mathbf{y}_s) + \mathbf{g}_s]$
        $\mathbf{x}_s^{(k+1)} = \Pi_{\mathcal{X}}(\mathbf{v}_s^{(k+1)})$
    • End
    • Set $\mathbf{y}_{s+1} = \mathbf{x}_s^{(K+1)}$
End
Return: $\mathbf{y}_{S+1}$
```

Algorithm 2: OVRG.

```
Initialization: Initialize $\mathbf{y}_1 \in \mathcal{X}$, set number of stages $S$,
batch sizes $\{k_s\}_{s=1}^S$, update frequency $K$, and the step-size
sequence $\{\alpha_s^{(t)}\}_{s\in[S],t\in[K]}$
Iterate: for $s = 1, \cdots, S$
    • Obtain samples $(\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \cdots, \boldsymbol{\xi}_{k_s})$
    • Compute $\hat{\mathbf{g}}_s := \frac{1}{k_s}\sum_{i\in[k_s]} \nabla f_i(\mathbf{y}_s)$
    • Set $\mathbf{x}_s^{(1)} = \mathbf{y}_s$
    • Iterate: for $t = 1, \cdots, K$ and do Obtain sample $\boldsymbol{\xi}_t$
      and update
        $\mathbf{v}_s^{(t+1)} = \mathbf{x}_s^{(t)} - \alpha_s^{(t)}(\nabla f_t(\mathbf{x}_s^{(t)}) - \nabla f_t(\mathbf{y}_s) + \hat{\mathbf{g}}_s)$
        $\mathbf{x}_s^{(t+1)} = \Pi_{\mathcal{X}}(\mathbf{v}_s^{(t+1)})$
    • End
    • Set $\mathbf{y}_{s+1} = \mathbf{x}_s^{(K+1)}$
End
Return: $\mathbf{y}_{S+1}$
```

has been recently proposed in [49], [50]. SVRG proceeds in multiple stages, where at the start of each stage s, a "centering variable" $\mathbf{y}_s$ is defined from the output of the past stage. Then, the full gradient $\nabla F(\mathbf{y}_s)$ is computed once, for the purpose of performing modified SGD iterations inside an inner loop with K iterations. In each such inner iteration $k \in [K]$, we sample an index t_k uniformly at random from $[T]$ and perform the following update

$$\mathbf{v}_s^{(k+1)} = \mathbf{x}_s^{(k)} - \alpha_s^{(k)}(\nabla f_{t_k}(\mathbf{x}_s^{(k)}) - \nabla f_{t_k}(\mathbf{y}_s) + \nabla F(\mathbf{y}_s)) \tag{27a}$$

$$\mathbf{x}_s^{(k+1)} = \Pi_{\mathcal{X}}(\mathbf{v}_s^{(k+1)}), \forall\, k \in [K] \tag{27b}$$

where the superscript k denotes the inner SGD iteration counter for stage s and we set $\mathbf{x}_s^{(1)} = \mathbf{y}_s$. Again, the expectation $\mathbb{E}_{t_k}(\mathbf{v}_s^{(k+1)} | \mathbf{x}_s^{(k)})$ equals (25a). However, compared to the sampled gradient $\nabla f_{t_k}(\mathbf{x}_s^{(k)})$ used in SGD, SVRG uses a different unbiased gradient estimator $\nabla f_{t_k}(\mathbf{x}_s^{(k)}) - \nabla f_{t_k}(\mathbf{y}_s) + \nabla F(\mathbf{y}_s)$ with potentially smaller variance, provided the step-size sequence and the stage length K are chosen appropriately (to be specified later). The overall algorithm is summarized in Algorithm 1.

B. Online Setting

In this case, our problem can be viewed as the task of minimizing

$$\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\boldsymbol{\xi}}[f(\mathbf{x};\boldsymbol{\xi})] \tag{28}$$

based on sequentially processing the stream of observed samples $\{\boldsymbol{\xi}_t\}_{t=0}^{\infty}$. Given a sample $\boldsymbol{\xi}_t$ (here, t denotes the streaming index), we define the instantaneous function $f_t(\mathbf{x}) := f(\mathbf{x};\boldsymbol{\xi}_t)$. Using the instantaneous gradient $\nabla f_t(\mathbf{x}^{(t)})$ (which equals the ensemble gradient in expectation), we perform online gradient descent (OGD) updates of the same form as (26), except that t_k now denotes the streaming index t.

Meanwhile, an online variant of SVRG (termed as OVRG here) was proposed in [51]. The algorithm is similar to SVRG, except that we cannot compute the gradient $\mathbb{E}_{\boldsymbol{\xi}}[\nabla f(\mathbf{y}_s;\boldsymbol{\xi})]$ of the centering variable $\mathbf{y}_s$. To overcome this obstacle, at the beginning of each stage s, we use k_s samples to form the surrogate gradient

$$\hat{\mathbf{g}}_s := \frac{1}{k_s}\sum_{i\in[k_s]} \nabla f_i(\mathbf{y}_s) \tag{29}$$

which is then used to compute inner SGD iterations similar to SVRG over a fixed loop of K iterations. Pseudocode for OVRG is presented in Algorithm 2.

Note that if $k_s = \infty$, then OVRG coincides with SVRG.

C. Discussion Regarding Convergence

In this section, we discuss the convergence of the aforementioned FOMs. Considering the offline scenario first, for problems of the form (23), an SGD algorithm with a randomized stopping criterion was proposed in [52], which features convergence to a stationary point of (23) (in expectation) using a constant $1/2L$ step-size. However, a randomized stopping criterion is not desirable in practical implementation. For SVRG, a non-asymptotic convergence result is proved in [53]. In this case, the result assumes perfect knowledge of the Lipschitz constant L of $F(\mathbf{x})$ in order to appropriately set the SVRG parameters (the update frequency K and the step-size sequence). However, L cannot be exactly determined in our case; we can only estimate an upper bound on L. The sensitivity of the convergence results to the use of such estimates to determine the SVRG parameters is unknown[7].

Meanwhile, for the online case, a modified variant of OGD was presented in [54, Section 4] with almost-sure convergence of the iterates to the set of stationary points of (28) using a diminishing step-size rule. Again, the result requires that the functions $f(\mathbf{x},\boldsymbol{\xi})$ are uniformly L-smooth for all $(\mathbf{x},\boldsymbol{\xi}) \in \mathcal{X} \times \Xi$, and this value of L is explicitly used in designing the step-size sequence. Since our aim is to devise algorithms for cases when the channel

[7] The same argument applies for SGD as well, but is more pronounced in the case of SVRG.

distribution is unknown, it is clear that we cannot determine L *a priori* [8]. Finally, we are currently only aware of a convergence analysis of OVRG for the case where the cost function of (28) is strongly convex and L-smooth [51]. Extending this result to the non-convex setting is an open problem.

In our implementation of these algorithms, we first tried using conservative estimates of the Lipschitz constants. However, our experiments indicate that this does not translate to good performance in practice, i.e., the progress made (in terms of minimizing the outage probability) over a reasonable number of iterations was very minimal. Hence, we opted for more aggressive choices of step-sizes, which work well empirically. While we cannot make any theoretical claims regarding convergence for such step-sizes, the improvement in performance we observe in practice is far too significant to overlook.

IV. Markov Approximation Based Methods

In this section, we discuss an alternative class of algorithms for approximating (3) based on Markov's inequality. The basic idea underpinning this approach was proposed in [30], and is reproduced here for completeness.

We begin by equivalently expressing (3) as

$$\min_{\mathbf{w}\in\mathcal{W}} \Pr[|\mathbf{w}^H\mathbf{h}|^2 < \gamma] \Longleftrightarrow \max_{\mathbf{w}\in\mathcal{W}} \Pr[|\mathbf{w}^H\mathbf{h}|^2 \geq \gamma] \tag{30}$$

Markov's inequality states that $Pr[x \geq t] \ \leq t^{-1}\mathbb{E}[x]$ for any nonnegative random variable. It thus follows that

$$\Pr[|\mathbf{w}^H\mathbf{h}|^2 \geq \gamma] \leq \gamma^{-1}\mathbf{w}^H\mathbf{R}\mathbf{w}, \forall\ \mathbf{w}\in\mathcal{W} \tag{31}$$

where $\mathbf{R} := \mathbb{E}[\mathbf{h}\mathbf{h}^H]$ is the channel covariance matrix. Hence, the problem

$$\max_{\mathbf{w}\in\mathcal{W}} \mathbf{w}^H\mathbf{R}\mathbf{w} \tag{32}$$

corresponds to maximizing an upper bound on (30). While it would be preferable to maximize a lower bound on the objective of (30), determining such a suitable lower bound is non-trivial (see [30] for details).

Unlike [30] which assumes that $\mathbf{R}$ is perfectly known, here we are interested in cases where only (possibly stale) representative channel realizations are given. We therefore use stochastic approximation to compute an empirical estimate in both offline and online settings to serve as a baseline for comparison against our FOMs.

A. Offline Setting

Using the sample set $\mathcal{H}_T$, we construct the empirical covariance matrix

$$\hat{\mathbf{R}}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t\mathbf{h}_t^H \tag{33}$$

using which we obtain the following surrogate of (32)

$$\max_{\mathbf{w}\in\mathcal{W}} \mathbf{w}^H\hat{\mathbf{R}}_T\mathbf{w} \tag{34}$$

When the set $\mathcal{W}$ corresponds to sum-power constraints (SPCs), the problem admits a closed form solution given by the principal eigen-vector of $\hat{\mathbf{R}}_T$ scaled by $\sqrt{P}$ (the maximum power budget). We term this method as MM-App (Modified Markov approximation).

B. Online Setting

Here, we describe an incremental algorithm to compute solutions of (32) subject to per-antenna constraints (PACs). For the case of SPCs, we can employ the well known Oja's algorithm [55], [56], which features guaranteed convergence to the principal eigenvector of $\mathbf{R}$. On the other hand, with PACs, the non-convexity of (32) makes it considerably more challenging to compute high-quality solutions. We now describe our approach.

Given a channel sample $\mathbf{h}_t$, we have access to the random function

$$q_t(\tilde{\mathbf{w}}) := q_t(\tilde{\mathbf{w}};\tilde{\mathbf{h}}_t) = \tilde{\mathbf{w}}^T\tilde{\mathbf{R}}_t\tilde{\mathbf{w}} \tag{35}$$

where $\tilde{\mathbf{R}}_t := \tilde{\mathbf{H}}_t\tilde{\mathbf{H}}_t^T$. Since $q_t(\tilde{\mathbf{w}})$ is a convex quadratic, note that the linear surrogate function

$$q_t(\tilde{\mathbf{w}},\tilde{\mathbf{w}}') := q_t(\tilde{\mathbf{w}}') + \nabla q_t(\tilde{\mathbf{w}}')^T(\tilde{\mathbf{w}}-\tilde{\mathbf{w}}') \tag{36}$$

is a global lower bound for $q_t(\tilde{\mathbf{w}})$ for all $\tilde{\mathbf{w}}' \in \tilde{\mathcal{W}}$ (with equality at $\tilde{\mathbf{w}} = \tilde{\mathbf{w}}'$). This implies that if we subtract a proximal term from $q_t(\tilde{\mathbf{w}},\tilde{\mathbf{w}}')$, the resulting strongly concave function

$$q_t^{(\alpha)}(\tilde{\mathbf{w}},\tilde{\mathbf{w}}') := q_t(\tilde{\mathbf{w}}') + \nabla q_t(\tilde{\mathbf{w}}')^T(\tilde{\mathbf{w}}-\tilde{\mathbf{w}}') - \frac{\alpha}{2}\|\tilde{\mathbf{w}}-\tilde{\mathbf{w}}'\|_2^2 \tag{37}$$

is also a global lower bound for $q_t(\tilde{\mathbf{w}})$ for all $\tilde{\mathbf{w}}' \in \tilde{\mathcal{W}}$ and $\alpha > 0$. Using the above defined quadratic surrogate function, we propose to use the following algorithm, which we designate as OM-App (Online Markov approximation)

$$\begin{aligned}\tilde{\mathbf{y}}^{(t+1)} &= \arg\max_{\tilde{\mathbf{w}}} q_t^{(\alpha)}(\tilde{\mathbf{w}},\tilde{\mathbf{w}}^{(t)}) \\ &= \tilde{\mathbf{w}}^{(t)} + \frac{1}{\alpha}\nabla q_t(\tilde{\mathbf{w}}^{(t)}) \qquad (38a)\end{aligned}$$

$$\tilde{\mathbf{z}}^{(t+1)} = \frac{t}{t+1}\tilde{\mathbf{z}}^{(t)} + \frac{1}{t+1}\tilde{\mathbf{y}}^{(t+1)} \tag{38b}$$

$$\tilde{\mathbf{w}}^{(t+1)} = \mathbf{\Pi}_{\tilde{\mathcal{W}}}(\tilde{\mathbf{z}}^{(t+1)}), \forall\ t\in\mathbb{N} \tag{38c}$$

Note that for $t = 0$, the above algorithm corresponds to OGD. For subsequent iterations, a convex combination of the OGD iterate (38a) and the auxiliary iterate $\tilde{\mathbf{z}}^{(t)}$ is performed according to (38b). It can be readily verified OM-App is a special instance of the Stochastic Successive Upper-bound Minimization (SSUM) algorithm proposed in [54, p. 536]. Consequently, under the assumption that the channel samples $\mathbf{h}_1, \mathbf{h}_2, \cdots$ correspond to i.i.d. realizations of $\mathbf{h}$, by virtue of [54, Theorem 1], the sequence of iterates $\{\tilde{\mathbf{w}}^{(t)}\}_{t=0}^{\infty}$ generated by OM-App is guaranteed to globally converge to the set of stationary points of (32), almost surely.

V. Simulation Results

In this section, we evaluate the outage probability performance of the proposed methods in a point-to-point beamforming

[8] Knowledge of the support set Ξ is required in this case.

scenario. With the information of the historical channel realizations, we evaluate the performance of the stochastic approximation solutions, i.e., sigmoid and PWM. All experiments are carried out on a Windows desktop with 4 Intel i7 cores and 8 GB of RAM. It is assumed that the channel vector $\mathbf{h}$ is drawn from a GMM with distribution $p(\mathbf{h})$ given by

$$p(\mathbf{h}) = \sum_{j=1}^{J} \pi_j \mathcal{CN}(\mathbf{h}; \mathbf{m}_j, \sigma_j^2 \mathbf{I}). \tag{39}$$

Here, $\mathcal{CN}((\cdot), \mathbf{m}, \mathbf{R})$ denotes a multivariate Gaussian complex distribution of mean vector $\mathbf{m}$ and covariance matrix $\mathbf{R}$. J is the number of kernels in the GMM and $\pi_j, j \in [J]$ are the corresponding mixture probabilities of each kernel with $\pi_j \geq 0$, $\sum_{j=1}^{J} \pi_j = 1$. The mean $\mathbf{m}_j$ of each Gaussian kernel is modeled as a Vandermonde steering vector $\mathbf{m}_j = [1, e^{i\theta_j} \cdots, e^{i(N-1)\theta_j}]^T$, where $\theta_j \in [0, 2\pi)$ denotes the angle of the j^{th} specular path. Unless stated othwerwise, we set the number of kernels to $J = 4$, the angles $\{\theta_j\}_{j=1}^{J}$ of the specular paths to $\{-\frac{2}{3}\pi, \frac{1}{6}\pi, \frac{1}{2}\pi, \frac{3}{4}\pi\}$, the variances to $\sigma_j^2 = 1$, and the mixture probabilities to $\pi_j = \frac{1}{J}, \forall j \in [J]$ throughout our experiments. Note that given enough kernels J, the GMM can approximately fit any distribution in practice. We test the performance of the proposed sigmoid SGD, sigmoid SVRG, PWM SGD and PWM SVRG with the number of historical channel samples $T = 500$ and the outage threshold $\gamma = 4$. For PWM SGD and PWM SVRG, the smoothing parameter is set to $\mu = 10^{-3}$. Regarding step sizes, a diminishing step-size rule $\alpha_k = c/k$ is used in the SGD based methods. Herein, $c = 3$ and k represents the number of iterations. For sigmoid SVRG and PWM SVRG, the update frequency is set to $K = 2T$ and the step size is fixed at $\alpha = 0.01$. Since the SGD and SVRG based methods require a different number of gradient evaluations per iteration, for fair comparison, we set a fixed number of total gradient evaluations for each method and evaluate the outage probability after every T gradient evaluations. The number of gradient evaluations is set to 10^4 here. Note that this implies that the number of iterations for each method is different, depending on the number of gradients evaluated per iteration. In all examples, the outage probability results are averaged over 250 Monte Carlo simulations; i.e., over 250 different collections of channel realizations.

In our first set of experiments, we consider a traditional beamforming scenario where the number of antennas is set to 16. Since the number of antennas is relatively small, it is reasonable to use the SPCs and here the total transmitted power is set to $P = 4$.

In Fig. 2, we plot both the evolution of the outage cost function and the sample averaged surrogate functions (22a) and (22b) as a function of the number of gradients utilized by the stochastic gradient methods (a single unit on the x-axis represents T gradients processed). As expected, the smoothed PWM function, while being relatively easier to minimize compared to the sigmoid function, overall yields a worse approximation to the outage function. It is evident that all methods are successful in converging to a small outage probability. Initially, sigmoid SGD and PWM SGD converge faster than their SVRG counterparts. However, the sigmoid SGD and PWM SGD eventually attain somewhat higher outage probabilities compared to the sigmoid SVRG and PWM SVRG, respectively, which are slower to converge in this experiment.

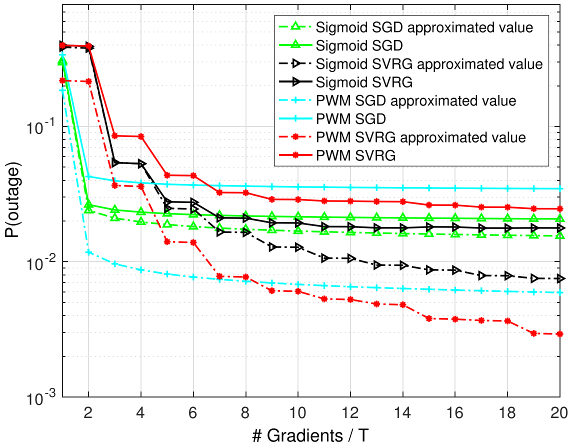

Fig. 2. Outage Probability as a function of iterations with $\|\mathbf{m}_j\|_2^2 = N = 16$, $\sigma_j^2 = 1$, and $\pi_j = \frac{1}{J}, \forall j \in [J]$.

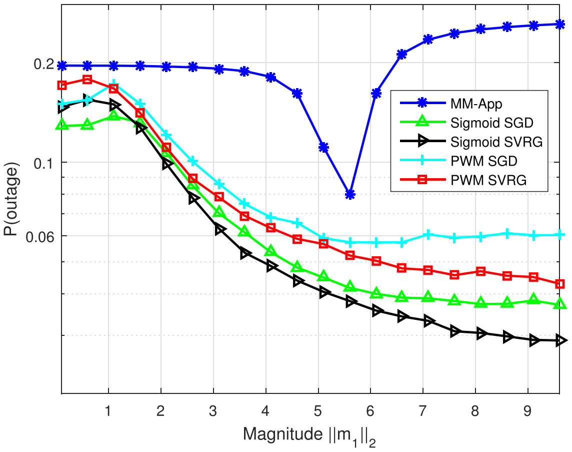

Fig. 3. Outage Probability as a function of magnitude with $\sigma_j^2 = 1$, $\pi_j = \frac{1}{J}$, $\forall j \in [J]$, and $\|\mathbf{m}_j\|_2^2 = N = 16, j = 2, \cdots, J$.

In Fig. 3, we examine the outage probabilities for the case that $\|\mathbf{m}_1\|_2$ is varied from $[0.1, 9.6]$, while the means of the other kernels are fixed. It is observed that MM-App fails to work efficiently especially when there is near-far imbalance. The outage probabilities of the proposed sigmoid and PWM based methods decrease quickly as the magnitude of $\|\mathbf{m}_1\|_2$ increases from 0.1 to 5.1 and then decreases slightly later on. From Fig. 3, we can see that sigmoid SVRG and PWM SVRG yield smaller outage probabilities compared to the sigmoid SGD and PWM SGD schemes, respectively.

Fig. 4 shows the outage probabilities as a function of the mixture probability of the first kernel. It is observed that the sigmoid SGD and sigmoid SVRG arrive at the same steady-state while the PWM SGD and PWM SVRG have the same steady-state behavior. However, the sigmoid based approaches can yield smaller outage probabilities especially when π_1 is

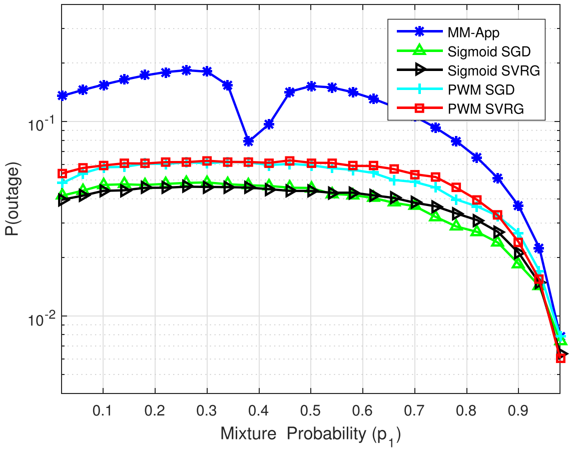

Fig. 4. Outage Probability as a function of the mixture probability with $\|\mathbf{m}_j\|_2^2 = N = 16$, $\pi_j = \frac{1}{J}, \forall j \in [J]$, and $\sigma_j^2 = 1, j = 2, \cdots, J$.

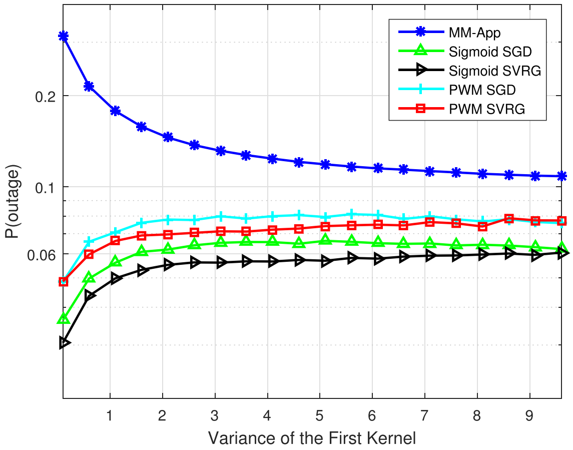

Fig. 5. Outage Probability as a function of the variance with $\|\mathbf{m}_j\|_2^2 = N = 16$, $\sigma_j^2 = 1, \forall j \in [J]$, and $\pi_j = \frac{1}{J}, j = 2, \cdots, J$.

less than 0.8. Furthermore, we can see that the MM-App yields the largest outage probability among all the methods. As π_1 approaches 1, all algorithms achieve their best performances and provide almost the same outage probabilities. That's due to the fact that only the first kernel plays a major role when $\pi_1 \approx 1$ and the GMM is similar to Gaussian distribution model.

The outage probabilities of the MM-App, sigmoid SGD, sigmoid SVRG, PWM SGD and PWM SVRG for different values of σ_1^2 are depicted in Fig. 5. Compared to MM-App, the FOMs can provide much smaller outage probabilities especially when σ_1^2 is less than 1. In terms of approximation, the sigmoid solution is superior to the PWM scheme. In particular, the sigmoid SGD has a slight performance improvement compared to PWM SGD while the sigmoid SVRG yields a smaller outage probability than the PWM SVRG. Moreover, we can see that the sigmoid SGD and PWM SGD are slightly inferior to sigmoid SVRG and PWM SVRG, respectively.

In Fig. 6, we examine the performance of the methods as the size of the channel sample set is increased $T = 10$ to 510. In this case, the total number of gradient evaluations is fixed to $100T$ for each value of T (equivalent to 100 total passes through the sample set). It is observed that the OGD and OVRG methods significantly outperform MM-App overall. The outage probabilities of the FOMs gradually decrease as the size of the sample set increases, with very satisfactory performance attained with only 100 samples. Furthermore, we can see that the sigmoid based methods yield smaller outage probabilities than the PWM based schemes.

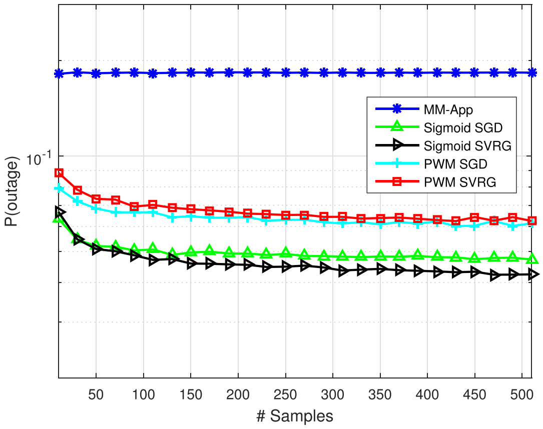

Fig. 6. Outage Probability as a function of the number of channel samples with $\|\mathbf{m}_j\|_2^2 = N = 16$, $\sigma_j^2 = 1$, and $\pi_j = \frac{1}{J}, \forall j \in [J]$.

We now move onto our second set of experiments, where we test our proposed streaming algorithms in large-scale antenna systems, with SPCs replaced by PACs. Unless stated otherwise, the threshold γ is set to 4 and the parameter settings of the GMM are $\|\mathbf{m}_j\|_2^2 = N$, $\sigma_j^2 = 1$, $\pi_j = \frac{1}{J}, \forall j \in [J]$. We allocate a maximum per-antenna power budget of $P_n = 0.25, \forall\, n \in [N]$. All results are averaged over 250 Monte Carlo realizations. We first consider a scenario with $N = 100$ antennas, where we apply OGD and OVRG. Four stochastic algorithms, i.e., sigmoid OGD, PWM OGD, sigmoid OVRG and PWM OVRG, together with OM-App, are tested. All the methods are initialized from the same random vector satisfying the PACs. Again, for fair comparison, we use a maximum budget of 32960 gradient evaluations for each algorithm. Since the OGD and OVRG based methods require different number of gradient evaluations per iteration, the number of iterations executed for a fixed number of channel realizations is different. For OM-App, we set $\alpha = 10^{-3}$, while for sigmoid OGD and PWM OGD, we use a diminishing step size $\frac{1}{\sqrt{t}}$. For sigmoid OVRG and PWM OVRG, we use a constant step size 0.0225. For the OVRG based methods, we set the batch sizes $\{k_s\}_{s=1}^{S}$ as $k_1 = 80$,

$$k_s = \begin{cases} 2k_{s-1}, & k_s < 640 \\ 640, & \text{otherwise} \end{cases} \tag{40}$$

with $s = 2, \cdots, S$ and the inner loop sample size $K = 1000$. The smoothing parameter for the PWM based methods is set to $\mu = 10^{-3}$.

We plot the outage probability results in every $K_s = 200$ gradients in Fig. 7. We can see that OM-App fails to work

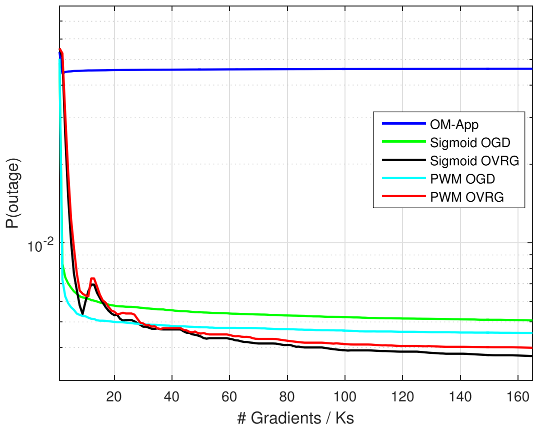

Fig. 7. Outage Probability as a function of iterations with $\|\mathbf{m}_j\|_2^2 = N = 100$, $\sigma_j^2 = 1$, and $\pi_j = \frac{1}{J}, \forall j \in [J]$.

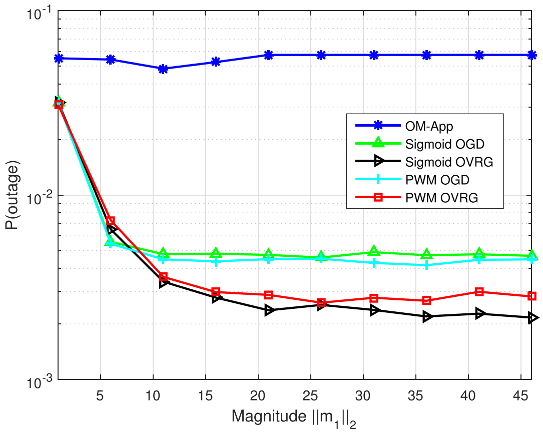

Fig. 8. Outage Probability as a function of magnitude with $\sigma_j^2 = 1$, $\pi_j = \frac{1}{J}$, $\forall j \in [J]$, and $\|\mathbf{m}_j\|_2^2 = N = 100, j = 2, \cdots, J$.

effectively. It is also observed that sigmoid OGD and PWM OGD methods initially converge faster than their OVRG based counterparts. However, we can see that sigmoid OVRG and PWM OVRG yield smaller outage probabilities compared to sigmoid OGD and PWM OGD, respectively.

Fig. 8 depicts the outage probability results for the case where we vary $\|\mathbf{m}_1\|_2$ while the mean vectors of the other kernels is fixed to $N = 100$. It is observed that the performances of sigmoid OGD, sigmoid OVRG, PWM OGD and PWM OVRG are superior to that of OM-App. For sigmoid OGD, the outage probability declines rapidly from 0.03 to 0.004 when $\|\mathbf{m}_1\|_2$ increases from 1 to 10, and keeps at around 0.0035 later on. The PWM OGD has approximately the same performance as sigmoid OGD. Furthermore, we can see that the OVRG based methods can provide smaller outage probabilities especially for $\|\mathbf{m}_1\|_2 \geq 10$. Also, PWM OVRG is slightly inferior to sigmoid OVRG especially when $\|\mathbf{m}_1\|_2 \geq 15$.

The outage probability results as a function of π_1 are depicted in Fig. 9. It is observed that sigmoid OGD, sigmoid OVRG, PWM OGD and PWM OVRG significantly outperform the OM-App over the whole range. For OM-App, the value of π_1 has little influence on the outage probability performance.

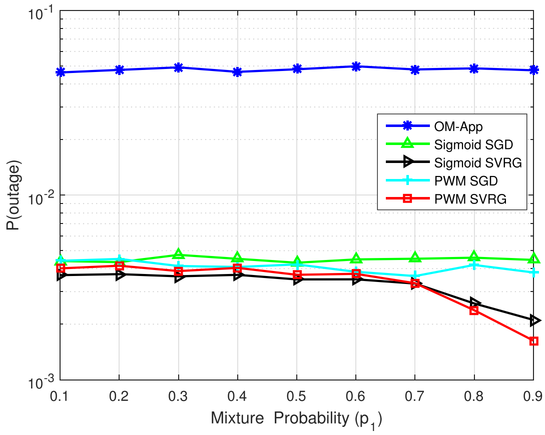

Fig. 9. Outage Probability as a function of the mixture probability. With $\|\mathbf{m}_j\|_2^2 = N = 100$, $\sigma_j^2 = 1, \forall j \in [J]$, and $\pi_j = \frac{1}{J}, j = 2, \cdots, J$.

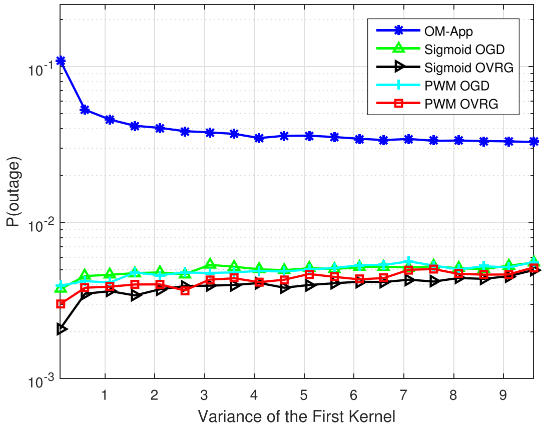

Fig. 10. Outage Probability as a function of the variance with $\|\mathbf{m}_j\|_2^2 = N = 100$, $\pi_j = \frac{1}{J}, \forall j \in [J]$, and $\sigma_j^2 = 1, j = 2, \cdots, J$.

The outage probabilities of the OVRG based methods decrease as π_1 increases from 0.1 to 0.9 while the OGD based schemes yield approximately the same outage probabilities in the whole range. Furthermore, the sigmoid OVRG and PWM OVRG can obtain smaller outage probabilities than their OGD based counterparts. Between the OVRG based methods, PWM OVRG is slightly inferior to sigmoid OVRG when $\pi_1 \leq 0.6$ while it can provide the smallest outage probability for the case of $\pi_1 > 0.7$.

The outage probability results as a function of the variance of the first kernel are depicted in Fig. 10. It is observed that the OGD and OVRG based methods can provide much smaller outage probabilities than OM-App. Furthermore, we can see that the sigmoid OVRG and PWM OVRG are slightly superior than the sigmoid OGD and PWM OGD, respectively.

The outage probability results as a function of the square root of the threshold γ are depicted in Fig. 11. As expected, the outage probabilities of all methods increase with the increase of $\gamma^{1/2}$. We point out that OM-App has the largest outage probability amongst all the methods. As for the PWM based methods, it is observed that PWM OGD is always slightly inferior to the PWM OVRG. However, for the sigmoid approximation based methods, sigmoid OGD can provide larger outage probability

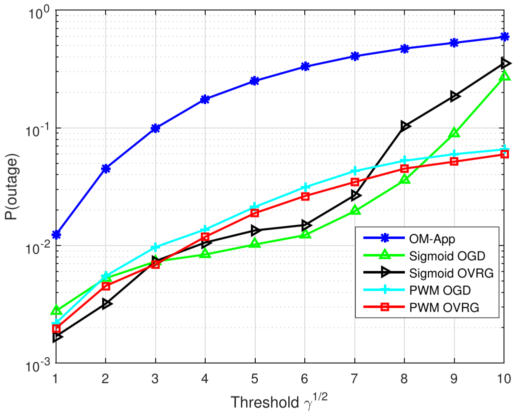

Fig. 11. Outage Probability as a function of the threshold γ with $\|\mathbf{m}_j\|_2^2 = N = 100$, $\sigma_j^2 = 1$, and $\pi_j = \frac{1}{J}, \forall j \in [J]$.

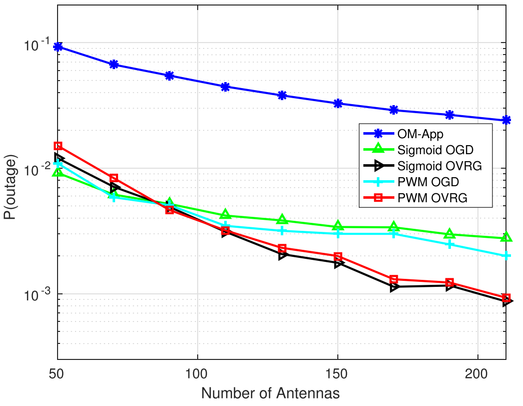

Fig. 12. Outage Probability as a function of the number of antennas with $\|\mathbf{m}_j\|_2^2 = N$, $\sigma_j^2 = 1$, and $\pi_j = \frac{1}{J}, \forall j \in [J]$.

than the sigmoid OVRG for the case of $\gamma^{1/2} \leq 3$ while the former slightly outperforms the latter when $\gamma^{1/2}$ is larger than 3. Furthermore, we can see that the PWM OGD and PWM OVRG yield large performance improvements compared with the sigmoid based schemes when $\gamma^{1/2}$ is larger than 9.

In Fig. 12, we plot the outage probability results as the number of antennas N is increased from $N = 50$ to 210. It is observed that the OM-App solution yields the largest outage probability in the whole range. As for the FOMs, the OVRG based methods are slightly inferior to the OGD based schemes when N is less than 90 while they can provide smaller outage probabilities for the case of $N \geq 90$. Furthermore, we can see that sigmoid OVRG yields approximately the same outage probability as PWM OVRG, while sigmoid OGD has the same performance as PWM OGD.

VI. Conclusion

In this paper, we proposed a novel framework for minimum outage transmit beamforming. The new framework differs from earlier outage-based beamforming approaches which assume knowledge of the channel distribution. Instead, our approach relies on stale, intermittent, historical, or even peer feedback of channel vectors drawn from the target distribution. Our formulation fits two basic wireless communication modalities: point-to-point transmit beamforming, and single-group multicast beamforming. Our designs are based on minimizing outage probability under sum or per-antenna power constraints.

Considering that we have no prior knowledge of the channel distribution, the outage probability criterion is first expressed as the expectation of the indicator function of the outage set. In this form, the criterion is amenable to sample average approximation, using a sum of indicator functions. Since the indicator function is non-convex and discontinuous, two different smooth approximations (sigmoid and pairwise maximum) are employed as smooth optimization surrogates. Two application scenarios are considered: batch and streaming optimization, applicable when one has access to stored historical data, or only a single channel instance at a time, respectively. In the batch setting, four different algorithms (sigmoid SGD, sigmoid SVRG, PWM SGD and PWM SVRG) are proposed for optimization under sum power constraints. In the streaming setting, four algorithms (sigmoid OGD, sigmoid OVRG, PWM OGD and PWM OVRG) are proposed for online optimization under per-antenna power constraints. Given the unfavorable attributes of the indicator function (non-convexity and discontinuity), it is hard to theoretically assess the performance of the proposed stochastic optimization algorithms. Yet, judiciously designed experiments indicate that they are remarkably effective for this non-convex and NP–hard class of problems. Considering also their implementation simplicity, the proposed algorithms appear to be serious candidates for practical implementation. After all, it is hard to argue against something simple that works.

Finally, extending our outage based approach to more general downlink multi-user MISO and MIMO beamforming scenarios is also of interest. However, the problem formulation in these settings is far more complicated than the one considered here, which makes it difficult to develop a straightforward extension of the algorithms presented here. Developing a new algorithmic approach to tackle these problems is deferred for follow-up work. Another line of future work could be to extend our approach for designing transmission schemes with higher-rank covariance matrices.

Appendix A
Outage Minimization With Gaussian Models

Consider the outage minimization problem (3) subject to sum-power constraints where the channel distribution follows the model $\mathbf{h} \sim \mathcal{CN}(\mathbf{m}, \mathbf{C})$. This corresponds to a special case of the GMM with only a single kernel. It has already been established in [30, Claim 2] that this problem can be reduced to a 1-D line search. We summarize and expand upon this result here, as it applies in our context.

In this case, the distribution of the received signal is given by $y \sim \mathcal{CN}(\mathbf{w}^H\mathbf{m}, \mathbf{w}^H\mathbf{C}\mathbf{w})$, which allows us to explicitly express the cost function of (3) as

$$F(\mathbf{w}) = \int\int_{\mathbf{A}_\gamma} \mathcal{CN}(\mathbf{w}^H\mathbf{m}, \mathbf{w}^H\mathbf{C}\mathbf{w}) \tag{41}$$

where $\mathbf{A}_\gamma$ denotes the disc of radius γ in the complex plane. Let $\mathbf{p} \in \mathbf{C}^N$ denote the unit-norm principal component of $(\mathbf{I} - \frac{\mathbf{m}\mathbf{m}^H}{\|\mathbf{m}\|_2^2})\mathbf{C}(\mathbf{I} - \frac{\mathbf{m}\mathbf{m}^H}{\|\mathbf{m}\|_2^2})$. According to the result of [30, Appendix

B], the optimal beamforming vector $\mathbf{w}$ which minimizes $F(\mathbf{w})$ over the norm ball $\|\mathbf{w}\|_2^2 \leq P$ lies in the subspace spanned by $\mathbf{m}$ and $\mathbf{p}$, and can be computed by solving the following one-dimensional line search problem

$$\min_{0 \leq c \leq P\|\mathbf{m}\|_2^2} F\left(\sqrt{c}\frac{\mathbf{m}}{\|\mathbf{m}\|_2^2} + \sqrt{P - \frac{c}{\|\mathbf{m}\|_2^2}}\mathbf{p}\right) \tag{42}$$

Hence, provided that the first and second-order channel statistics can be reliably estimated at the BS prior to downlink transmission, the outage problem can be solved optimally. Otherwise, if the mean (i.e., the "nominal channel direction") is known, but not $\mathbf{C}$, then we can compute $\mathbf{p}$ in an online fashion from a stream of instantaneous estimates $(\mathbf{I} - \frac{\mathbf{m}\mathbf{m}^H}{\|\mathbf{m}\|_2^2})(\mathbf{h}_t\mathbf{h}_t^H - \mathbf{m}\mathbf{m}^H)(\mathbf{I} - \frac{\mathbf{m}\mathbf{m}^H}{\|\mathbf{m}\|_2^2})$ using Oja's algorithm. The streaming estimate $\hat{\mathbf{p}}_t$ of $\mathbf{p}$ can then be used in (42) for obtaining an online solution for the problem. Furthermore, if the channel distribution can be better described by a GMM which happens to change states slowly and the BS can accurately track $\mathbf{m}$ and $\mathbf{C}$, then we can opt to minimize outage on a per-state basis, which again reduces to solving a problem of the form (42) for each state.

References

[1] Y. Shi, A. Konar, N. D. Sidiropoulos, X.-P. Mao, and Y.-T. Liu, "Transmit beamforming for minimum outage via stochastic approximation," presented at the IEEE 7th Int. Workshop Comp. Adv. Multi-Sensor Adaptive Process., Dutch Antilles, pp. 1–5, Dec. 2017.

[2] F. Rashid-Farrokhi, L. Tassiulas, and K. J. R. Liu, "Joint optimal power control and beamforming in wireless networks using antenna arrays," *IEEE Trans. Commun.*, vol. 46, no. 10, pp. 1313–1324, Oct. 1998.

[3] M. Bengtsson and B. Ottersten, "Optimal and suboptimal transmit beamforming," in *Handbook of Antennas in Wireless Communications*, L. C. Godara, Ed. Boca Raton, FL, USA: CRC Press, Aug. 2001, ch. 18.

[4] M. J. Lopez, "Multiplexing, scheduling, and multicasting strategies for antenna arrays in wireless networks," Ph.D. dissertation, Elect. Eng. and Comp. Sci. Dept., Massachusetts Inst. Technol., Cambridge, MA, USA, 2002.

[5] N. D. Sidiropoulos, T. Davidson, and Z.-Q. Luo, "Transmit beamforming for physical-layer multicasting," *IEEE Trans. Signal Process.*, vol. 54, no. 6, pp. 2239–2251, Jun. 2006.

[6] N. Jindal and Z.-Q. Luo, "Capacity limits of multiple antenna multicast," in *Proc. IEEE Int. Symp. Inf. Theory*, Seattle, WA, USA, Jul. 2006, pp. 1841–1845.

[7] E. Karipidis, N. D. Sidiropoulos, and Z.-Q. Luo, "Quality of service and max-min-fair transmit beamforming to multiple co-channel multicast groups," *IEEE Trans. Signal Process.*, vol. 56, no. 3, pp. 1268–1279, Mar. 2008.

[8] D. J. Love, R. W. Heath, V. K. Lau, D. Gesbert, B. D. Rao, and M. Andrews, "An overview of limited feedback in wireless communication systems," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 8, pp. 1341–1365, Oct. 2008.

[9] W. U. Bajwa, J. Haupt, A. M. Sayeed, and R. Nowak, "Compressed channel sensing: A new approach to estimating sparse multipath channels," *Proc. IEEE*, vol. 98, no. 6, pp. 1058–1076, Jun. 2010.

[10] H. Yin, D. Gesbert, M. Filippou, and Y. Liu, "A coordinated approach to channel estimation in large-scale multiple-antenna systems," *IEEE J. Sel. Areas Commun.*, vol. 31, no. 2, pp. 264–273, Feb. 2013.

[11] J. Choi, Z. Chance, D. J. Love, and U. Madhow, "Noncoherent trellis coded quantization: A practical limited feedback technique for massive MIMO systems," *IEEE Trans. Commun.*, vol. 61, no. 12, pp. 5016–5029, Dec. 2013.

[12] A. Adhikary *et al.*, "Joint spatial division and multiplexing for mm-wave channels," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1239–1255, Jun. 2014.

[13] O. Mehanna and N. D. Sidiropoulos, "Channel tracking and transmit beamforming with frugal feedback," *IEEE Trans. Signal Process.*, vol. 62, no. 24, pp. 6402–6413, Dec. 2014.

[14] B. Gopalakrishnan and N. D. Sidiropoulos, "Cognitive transmit beamforming from binary CSIT," *IEEE Trans. Wireless Commun.*, vol. 14, no. 2, pp. 895–906, Feb. 2015.

[15] Z. Jiang, A. F. Molisch, G. Caire, and Z. Niu, "Achievable rates of FDD massive MIMO systems with spatial channel correlation," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2868–2882, May 2015.

[16] B. Lee, J. Choi, J.-Y. Seol, D. J. Love, and B. Shim, "Antenna grouping based feedback compression for FDD-based massive MIMO systems," *IEEE Trans. Commun.*, vol. 63, no. 9, pp. 3261–3274, Sep. 2015.

[17] M. B. Shenouda and T. N. Davidson, "Convex conic formulations of robust downlink precoder designs with quality of service constraints," *IEEE J. Sel. Topics Signal Process.*, vol. 1, no. 4, pp. 714–724, Dec. 2007.

[18] G. Zheng, K.-K. Wong, and T.-S. Ng, "Robust linear MIMO in the downlink: A worst-case optimization with ellipsoidal uncertainty regions," *EURASIP J. Adv. Signal Process.*, vol. 2008, pp. 1–15, Jun. 2008.

[19] N. Vucic and H. Boche, "Robust QoS-constrained optimization of downlink multiuser MISO systems," *IEEE Trans. Signal Process.*, vol. 57, no. 2, pp. 714–725, Feb. 2009.

[20] G. Zheng, K.-K. Wong, and B. Ottersten, "Robust cognitive beamforming with bounded channel uncertainties," *IEEE Trans. Signal Process.*, vol. 57, no. 12, pp. 4871–4881, Dec. 2009.

[21] A. Tajer, N. Prasad, and X. Wang, "Robust linear precoder design for multi-cell downlink transmission," *IEEE Trans. Signal Process.*, vol. 59, no. 1, pp. 235–251, Jan. 2011.

[22] E. Song, Q. Shi, M. Sanjabi, R.-Y. Sun, and Z.-Q. Luo, "Robust SINR constrained MISO downlink beamforming: When is semidefinite programming relaxation tight?" *EURASIP J. Wireless Commun. Netw.*, vol. 1, no. 1, pp. 1–11, Dec. 2012.

[23] Y. Huang, D. P. Palomar, and S. Zhang, "Lorentz-positive maps and quadratic matrix inequalities with applications to robust MISO transmit beamforming," *IEEE Trans. Signal Process.*, vol. 61, no. 5, pp. 1121–1130, Mar. 2013.

[24] W.-K. Ma, J. Pan, A. M.-C. So, and T.-H. Chang, "Unraveling the rank-one solution mystery of robust MISO downlink transmit optimization: A verifiable sufficient condition via a new duality result," *IEEE Trans. Signal Process.*, vol. 65, no. 7, pp. 1909–1924, Apr. 2017.

[25] Y. Xie, C. N. Georghiades, and A. Arapostathis, "Minimum outage probability transmission with imperfect feedback for MISO fading channels," *IEEE Trans. Wireless Commun.*, vol. 4, no. 3, pp. 1084–1091, May 2005.

[26] S. A. Vorobyov, Y. Rong, and A. B. Gershman, "Robust adaptive beamforming using probability-constrained optimization," in *Proc. IEEE 13th Statist. Signal Process. Workshop*, Jul. 2005, pp. 934–939.

[27] B. K. Chalise, S. Shahbazpanahi, A. Czylwik, and A. B. Gershman, "Robust downlink beamforming based on outage probability specifications," *IEEE Trans. Wireless Commun.*, vol. 6, no. 10, pp. 3498–3503, Oct. 2007.

[28] M. B. Shenouda and T. N. Davidson, "Probabilistically-constrained approaches to the design of the multiple antenna downlink," in *Proc. 42nd Asilomar Conf.*, Pacific Grove, USA, Oct. 2008, pp. 1120–1124.

[29] S. A. Vorobyov, H. Chen, and A. B. Gershman, "On the relationship between robust minimum variance beamformmers with probabilistic and worst-case distortionless response constraints," *IEEE Trans. Signal Process.*, vol. 56, no. 11, pp. 5719–5724, Nov. 2008.

[30] V. Ntranos, N. D. Sidiropoulos, and L. Tassiulas, "On multicast beamforming for minimum outage," *IEEE Trans. Wireless Commun.*, vol. 8, no. 6, pp. 3172–3181, Jun. 2009.

[31] M. B. Shenouda, T. N. Davidson, and L. Lampe, "Outage-based design of robust Tomlinson–Harashima transceivers for the MISO downlink with QoS requirements," *Signal Process.*, vol. 93, no. 12, pp. 3341–3352, Dec. 2013.

[32] Q. Li, A. M.-C. So, and W.-K. Ma, "Distributionally robust chance constrained transmit beamforming for multiuser MISO downlink," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Florence, Italy, May 2014, pp. 3479–3483.

[33] K.-Y. Wang, A. M.-C. So, T.-H. Chang, W.-K. Ma, and C.-Y. Chi, "Outage constrained robust transmit optimization for multiuser MISO downlinks: Tractable approximations by conic optimization," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5690–5705, Sep. 2014.

[34] X. He and Y.-C. Wu, "Tight probabilistic SINR constrained beamforming under channel uncertainties," *IEEE Trans. Signal Process.*, vol. 63, no. 13, pp. 3490–3505, Jul. 2015.

[35] F. Sohrabi and T. N. Davidson, "Coordinate update algorithms for robust power loading for the MU-MISO downlink with outage constraints," *IEEE Trans. Signal Process.*, vol. 64, no. 11, pp. 2761–2773, Jun. 2016.

[36] Y.-F. Liu, M. Hong, and E. Song, “Sample approximation-based deflation approaches for chance SINR-constrained joint power and admission control,” *IEEE Trans. Wireless Commun.*, vol. 15, no. 7, pp. 4535–4547, Jul. 2016.

[37] A. Ribeiro, “Ergodic stochastic optimization algorithms for wireless communication and networking,” *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 6369–6386, Dec. 2010.

[38] W. I. Li, X. Huang, and H. E. Leung, “Performance evaluation of digital beamforming strategies for satellite communications,” *IEEE Trans. Aerosp. Electron. Syst.*, vol. 40, no. 1, pp. 12–26, Jan. 2004.

[39] J. Li and P. Stoica, *MIMO Radar Signal Processing*, Hoboken, NJ, USA: Wiley, 2009.

[40] P. Popovski *et al.*, “Ultra-reliable low-latency communication (URLLC): Principles and and building blocks,” *IEEE Netw.*, vol. 32, no. 2, pp. 16–23, Mar./Apr. 2018.

[41] H. Robbins and S. Monro, “A stochastic approximation method,” *Ann. Math. Statist.*, vol. 22, no. 3, pp. 400–407, Sep. 1951.

[42] H. J. Kushner and G. Yin, *Stochastic Approximation and Recursive Algorithms and Applications*, vol. 35, 2nd ed. New York, NY, USA: Springer-Verlag, 2003.

[43] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro, “Robust stochastic approximation approach to stochastic programming,” *SIAM J. Optim.*, vol. 19, no. 4, pp. 1574–1609, Jan. 2009.

[44] R. Gray and L. Davidson, *Random Processes: A Mathematical Approach for Engineers*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1986.

[45] B. K. Pagnoncelli, S. Ahmed, and A. Shapiro, “Sample average approximation method for chance constrained programming: Theory and applications,” *J. Optim. Theory Appl.*, vol. 142, no. 2, pp. 399–416, Aug. 2009.

[46] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory*. Philadelphia, PA, USA: SIAM, 2009.

[47] C. C. Aggarwal, *Recommender Systems*. New York, NY, USA: Springer-Verlag, 2016.

[48] Y. Nesterov, “Smooth minimization of non-smooth functions,” *Math. Program.*, vol. 103, no. 1, pp 127–152, May 2005.

[49] R. Johnson and T. Zhang, “Accelerating stochastic gradient descent using predictive variance reduction,” in *Proc. 26th Int. Conf. Neural Info. Process. Syst.*, 2013, pp. 315–323.

[50] L. Xiao and T. Zhang, “A proximal stochastic gradient method with progressive variance reduction,” *SIAM J. Optim.*, vol. 24, no. 4, pp. 2057–2075, 2014.

[51] R. Frostig, R. Ge, S. M. Kakade, and A. Sidford, “Competing with the empirical risk minimizer in a single pass,” in *Proc. Conf. Learn. Theory*, Paris, France, Jul. 2015, pp. 728–763.

[52] S. Ghadimi, G. Lan, and H. Zhang, “Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization,” *Math. Program*, vol. 155, no. 1/2, pp. 267–305, 2016.

[53] S. J. Reddi, S. Sra, B. Poczos, and A. Smola, “Fast stochastic methods for nonsmooth nonconvex optimization,” 2016, arXiv:1605.06900v1.

[54] M. Razaviyayn, M. Sanjabi, and Z.-Q. Luo, “A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks,” *Math. Prog.*, vol. 157, no. 2, pp. 515–545, Jun. 2016.

[55] E. Oja, “Simplified neuron model as a principal component analyzer,” *J. Math. Biol.*, vol. 15, no. 3, pp. 263–273, Nov. 1982.

[56] P. Jain, C. Jin, S. M. Kakade, P. Netrapalli, and A. Sidford, “Streaming PCA: Matching matrix Bernstein and near-optimal finite sample guarantees for Oja’s algorithm,” in *Proc. Conf. Learn. Theory*, New York, NY, USA, Jun. 2016, pp. 1147–1164.

Yunmei Shi was born in Shandong, China. She received the B.Sc. degree in electronic and information engineering from the Harbin Institute of Technology, Weihai, China, in 2012, the M.Sc. degree in electronic and communication engineering from the Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China, and she is currently working towards the Ph.D. degree in communication and information engineering with the Harbin Institute of Technology, Harbin, China. From 2016 to 2017, she was a research scholar with the University of Minnesota in the Department of Electrical and Computer Engineering. Her research interests include array signal processing and optimization.

Aritra Konar (M’17) received the B.Tech. degree in electronics and communications engineering from the West Bengal University of Technology, West Bengal, India, and the M.S. and Ph.D. degrees in electrical engineering from the University of Minnesota, Minneapolis, MN, USA, in 2011, 2014, and 2017 respectively. He is currently a Postdoctoral Associate with the Department of Electrical and Computer Engineering, University of Virginia, Charlottesville, VA, USA. His research interests include statistical signal processing, wireless communications, nonlinear optimization, and data analytics.

Nicholas D. Sidiropoulos (F’09) received the Diploma in electrical engineering from the Aristotle University of Thessaloniki, Thessaloniki, Greece, and the M.S. and Ph.D. degrees in electrical engineering from the University of Maryland, College Park, College Park, MD USA, in 1988, 1990, and 1992, respectively. He has served on the faculty of the University of Virginia, University of Minnesota, and the Technical University of Crete, Greece, prior to his current appointment as Louis T. Rader Professor and the Chair of ECE, University of Virginia. From 2015 to 2017 he was an ADC Chair Professor with the University of Minnesota. His research interests include signal processing, communications, optimization, tensor decomposition, and factor analysis, with applications in machine learning and communications. He was a recipient of the NSF/CAREER award in 1998, the IEEE Signal Processing Society (SPS) Best Paper Award in 2001, 2007, and 2011, served as an IEEE SPS Distinguished Lecturer from 2008 to 2009, and currently serves as Vice President–Membership of the IEEE SPS. He was a recipient of the 2010 IEEE Signal Processing Society Meritorious Service Award, and the 2013 Distinguished Alumni Award from the University of Maryland, Department of Electrical and Computer Engineering. He is a Fellow of EURASIP (2014).

Xing-Peng Mao (M’06) was born in Liaoning, P.R. China, in 1972. He received the B.S. degree in radio electronics from the Northeast Normal University, Changchun, China, in 1993, and the M.Eng. and Ph.D. degrees from the Harbin Institute of Technology, Harbin, China, in 1999 and 2004, respectively, all in electrical engineering. He joined the Harbin Institute of Technology in 1993 as a Lecturer, working on the design and development of radar and other electronical system. From 2005 to 2006, he was a Postdoctoral Fellow with the University of Waterloo, Waterloo, ON, Canada. He is now a Professor with the Department of Electronic and Information Engineering, Harbin Institute of Technology. His current interests include signal processing in wireless communication and radar.He is a senior member of the Chinese Institute of Electronics.

Yong-Tan Liu (SM’91) was born in Nanjing, Jiangsu, China, in 1936. He received the B.S. degree from the radio engineering department of Tsinghua University, Beijing, China, in 1958. After graduation, he joined the Department of Radio Engineering, Harbin Institute of Technology, Harbin, China. He has been working on the radar system and signal processing of high-frequency ground-wave OTH radar, continuous wave radar, and microwave imaging radar. Mr. Liu is a member of the Chinese Academy of Sciences and the Academy of Engineering.