



HAL
open science

An Approximate ML Detector for MIMO Channels Corrupted by Phase Noise

Richard Combes, Sheng Yang

► **To cite this version:**

Richard Combes, Sheng Yang. An Approximate ML Detector for MIMO Channels Corrupted by Phase Noise. *IEEE Transactions on Communications*, 2018, 66 (3), pp.1176 -1189. 10.1109/tcomm.2017.2774803 . hal-01575803

HAL Id: hal-01575803

<https://hal.science/hal-01575803>

Submitted on 9 Apr 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

An Approximate ML Detector for MIMO Channels Corrupted by Phase Noise

Richard Combes and Sheng Yang

Abstract

We consider the multiple-input multiple-output (MIMO) communication channel impaired by phase noises at both the transmitter and receiver. We focus on the maximum likelihood (ML) detection problem for uncoded single-carrier transmission. We derive an approximation of the likelihood function, based on which we propose an efficient detection algorithm. The proposed algorithm, named *self-interference whitening* (SIW), consists in 1) estimating the self-interference caused by the phase noise perturbation, then 2) whitening the said interference, and finally 3) detecting the transmitted vector. While the exact ML solution is computationally intractable, we construct a simulation-based lower bound on the error probability of ML detection. Leveraging this lower bound, we perform extensive numerical experiments demonstrating that SIW is, in most cases of interest, very close to optimal with moderate phase noise. More importantly and perhaps surprisingly, such near-ML performance can be achieved by applying only twice the nearest neighbor detection algorithm. In this sense, our results reveal a striking fact: near-ML detection of phase noise corrupted MIMO channels can be done as efficiently as for conventional MIMO channels without phase noise.

Index Terms

MIMO systems, phase noise, maximum likelihood detection, probability of error.

I. INTRODUCTION

We consider the signal detection problem for the following discrete-time multiple-input multiple-output (MIMO) channel

$$\mathbf{y} = \text{diag}(e^{j\theta_{r,1}}, \dots, e^{j\theta_{r,n_r}}) \mathbf{H} \text{diag}(e^{j\theta_{t,1}}, \dots, e^{j\theta_{t,n_t}}) \mathbf{x} + \mathbf{z}, \quad (1)$$

The authors are with the Laboratory of Signals and Systems (L2S), CentraleSupélec, 3 rue Joliot-Curie, 91190 Gif-sur-Yvette, France. (e-mail: richard.combes, sheng.yang@centralesupelec.fr)

where $\mathbf{H} \in \mathbb{C}^{n_r \times n_t}$ is the channel matrix known to the receiver; $\mathbf{z} \in \mathbb{C}^{n_r \times 1}$ represents a realization of the additive noise whereas $\theta_{t,l}$ and $\theta_{r,k}$ are the phase noises at the l th transmit antenna and the k th receive antenna, respectively; the input vector $\mathbf{x} \in \mathbb{C}^{n_t \times 1}$ is assumed to be carved from a quadratic amplitude modulation (QAM). The goal is to estimate \mathbf{x} from the observation $\mathbf{y} \in \mathbb{C}^{n_r \times 1}$, with only statistical knowledge on the additive noise and the phase noises.

In the case where the phase noise is absent, the problem is well understood, and the maximum likelihood (ML) solution can be found using any nearest neighbor detection (NND) algorithm (see [1] and the references therein). In particular, the sphere decoder [2] has been shown to be very efficient [3], so that its expected complexity (averaged over channel realizations) is polynomial in the problem dimension n_t . Furthermore, there exist approximate NND algorithms (e.g., based on lattice reduction) that achieve near-ML performance when applied for MIMO detection [4].

The presence of phase noise in (1) is both a practical and long-standing problem in communication. In their seminal paper [5] back in the 70's, Foschini *et al.* used this model to capture the residual phase jitter at the phase-locked loop of the receiver side, and investigated both the performance of decoding algorithms as well as constellation design in the scalar case ($n_t = n_r = 1$). As a matter of fact, in most wireless communication systems, phase noise is present due to the phase and frequency instabilities in the radio frequency oscillators used at both the transmitter and the receiver [6]. The channel (1) can be seen as a valid mathematical model when the phase noise varies slowly as compared to the symbol duration.¹ While phase noise can be practically ignored in conventional MIMO systems, its impact becomes prominent at higher carrier frequencies since it can be shown that phase noise power increases *quadratically* with carrier frequency [6], [8]. The performance degradation due to phase noise becomes even more severe with the use of higher order modulations for which the angular separation between constellation points can be small. At medium to high SNR, phase noise dominates additive noise, becoming the capacity bottleneck [9], [10]. As for signal detection, finding the ML solution for the MIMO phase noise channel (1) is hard in general. Indeed, unlike for conventional MIMO channels, the likelihood function of the transmitted signal cannot be obtained in closed form.

¹As pointed out in [7] and the references therein, an effective discrete-time channel is usually obtained from a waveform phase noise channel after filtering. When the continuous-time phase noise varies rapidly during the symbol period, the filtered output also suffers from amplitude perturbation. More discussion is provided in Section VI-D.

Our Contribution. In this work, we propose an efficient MIMO detection algorithm which finds an approximate ML solution in the presence of phase noise. The main contributions of this work are summarized as follows.

- (i) We derive a tractable approximation of the likelihood function of the transmitted signal. While the exact likelihood does not have a close-form expression, the proposed approximation has a simple form and turns out to be accurate for weak to medium phase noises.
- (ii) Since maximizing the approximate likelihood function over a discrete signal set is still hard, we propose a heuristic method that finds an approximate solution by applying twice the nearest neighbor detection algorithm. The proposed algorithm, called *self-interference whitening* (SIW), has a simple geometric interpretation. Intuitively, the phase noise perturbation generates self-interference that depends on the transmitted signal through the covariance matrix. The main idea is to first estimate the covariance of the self-interference with a potentially inaccurate initial signal solution, then perform the whitening with the estimated covariance, followed by a second detection. From the optimization point of view, our algorithm can be seen as a (well-chosen) concave approximation to a non-concave objective function.
- (iii) We assess the performance of SIW and competing algorithms in different communication scenarios. Since the error probability of ML decoding is unknown, we propose a simulation-based lower bound which we use as a benchmark. Simulation results show that SIW achieves near ML performance in most scenarios. In this sense, our work reveals that near optimal MIMO detection with phase noise can be done as efficiently as without phase noise. Although the likelihood approximation is derived using the assumption that the phase noise has a Gaussian distribution, our numerical experiments show that SIW works well even with non-Gaussian phase noises.

Related Work. Receiver design with phase noise mitigation has been extensively investigated in the past years (see [11], [12] and references therein). More complex channel models, including multi-carrier systems (e.g., OFDM) and time-correlated phase noises (e.g., the Wiener process) have also been considered. In particular, joint data detection and phase noise estimation algorithms have been proposed in [13], [11]. A phase noise estimation based scheme to improve the system performance for smaller alphabets has been proposed in [12]. However, the challenging problem of signal detection in MIMO phase noise channels using higher order modulation, where

performance is extremely sensitive to phase noise, has not been addressed adequately before.

The remainder of the paper is organized as follows. We start with a formal description of the problem in the next section. The approximation of the likelihood function is derived in Section III-A, followed by the proposed algorithm described in Section III-B. The hardness of finding the exact ML solution is investigated in Section IV. We present the numerical experiments in Section V. Further discussion on the proposed algorithm and relevance of the considered channel model is provided in Section VI. Section VII concludes the paper.

II. ASSUMPTIONS AND PROBLEM FORMULATION

Notation

Throughout the paper, we use the following notation. For random quantities, we use upper case letters, e.g., X , for scalars, upper case letters with bold and non-italic fonts, e.g., \mathbf{V} , for vectors, and upper case letter with bold and sans serif fonts, e.g., \mathbf{M} , for matrices. Deterministic quantities are denoted in a rather conventional way with italic letters, e.g., a scalar x , a vector \mathbf{v} , and a matrix \mathbf{M} . The Euclidean norm of a vector \mathbf{v} is denoted by $\|\mathbf{v}\|$, respectively. The transpose and conjugated transpose of \mathbf{M} are \mathbf{M}^T and \mathbf{M}^H , respectively.

A. System model

In the following, we describe formally the channel model (1) presented in the previous section. We assume a MIMO channel with n_t transmit and n_r receive antennas. Let \mathbf{H} denote the channel matrix, where the (k, l) -th element of \mathbf{H} , denoted as $h_{k,l}$, represents the channel gain between the l th transmit antenna and k th receive antenna. The transmitted vector is denoted by $\mathbf{x} = [x_1, \dots, x_{n_t}]^T$, where $x_l \in \mathcal{X}, l = 1, \dots, n_t$, \mathcal{X} being typically a QAM constellation with normalized average energy, i.e., $\frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} |x|^2 = 1$. For a given transmitted vector \mathbf{x} , the received vector in base-band can be written as the following random vector

$$\mathbf{Y} = \mathbf{\Lambda}_R \mathbf{H} \mathbf{\Lambda}_T \mathbf{x} + \mathbf{Z}, \quad (2)$$

where the diagonal matrices $\mathbf{\Lambda}_R := \text{diag}(e^{j\Theta_{r,1}}, \dots, e^{j\Theta_{r,n_r}})$ and $\mathbf{\Lambda}_T := \text{diag}(e^{j\Theta_{t,1}}, \dots, e^{j\Theta_{t,n_t}})$ capture the phase perturbation at the receiver and transmitter, respectively; \mathbf{Z} is the additive white Gaussian noise (AWGN) vector with $\mathbf{Z} \sim \mathcal{CN}(0, \gamma^{-1} \mathbf{I})$, where γ is the nominal signal-to-noise ratio (SNR). We assume that the phase noise $\mathbf{\Theta} := [\Theta_{t,1} \ \dots \ \Theta_{t,n_t} \ \Theta_{r,1} \ \dots \ \Theta_{r,n_r}]^T$ is jointly Gaussian with $\mathbf{\Theta} \sim \mathcal{N}(0, \mathbf{Q}_\theta)$ where the covariance matrix \mathbf{Q}_θ can be arbitrary. Note

that this model includes as a special case the uplink channel in which n_t is the number of single-antenna users. In such a case, the transmit phase noises are independent. For simplicity, we consider uncoded transmission in which each symbol x_l can take any value from \mathcal{X} with equal probability.

Further, we assume that the channel matrix can be random but is perfectly known at the receiver, whereas such knowledge at the transmitter side is irrelevant in uncoded transmission. We also define $\mathbf{H}_\Theta := \mathbf{\Lambda}_R \mathbf{H} \mathbf{\Lambda}_T$ and accordingly \mathbf{H}_θ for some realization of $\Theta = \theta$. By definition, we have $\mathbf{H}_0 = \mathbf{H}$. Finally, we ignore the temporal correlation of the phase noise process and the channel process, and focus on the spatial aspect of the signal detection problem.

B. Problem formulation

With AWGN, we have the following conditional probability density function (pdf)

$$p(\mathbf{y} | \mathbf{x}, \theta, \mathbf{H}) = \frac{\gamma^{n_r}}{\pi^{n_r}} e^{-\gamma \|\mathbf{y} - \mathbf{H}_\theta \mathbf{x}\|^2}, \quad (3)$$

from which we obtain the likelihood function by integrating over Θ

$$p(\mathbf{y} | \mathbf{x}, \mathbf{H}) = \mathbb{E}_\Theta [p(\mathbf{y} | \mathbf{x}, \Theta, \mathbf{H})] \quad (4)$$

$$= \ln \left(\mathbb{E}_\Theta \left[e^{-\gamma \|\mathbf{y} - \mathbf{H}_\Theta \mathbf{x}\|^2} \right] \right) + \ln \frac{\gamma^{n_r}}{\pi^{n_r}}. \quad (5)$$

The ML detector finds an input vector from the alphabet \mathcal{X}^{n_t} such that the likelihood function is maximized. In practice, it is often more convenient to use the log-likelihood function as the objective function, i.e., after removing a constant term

$$f(\mathbf{x}, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta) := \ln \left(\mathbb{E}_\Theta \left[e^{-\gamma \|\mathbf{y} - \mathbf{H}_\Theta \mathbf{x}\|^2} \right] \right), \quad (6)$$

where the arguments γ and \mathbf{Q}_θ can be omitted whenever confusion is not likely. Thus,

$$\hat{\mathbf{x}}_{\text{ML}}(\mathbf{y}, \mathbf{H}) := \arg \max_{\mathbf{x} \in \mathcal{X}^{n_t}} f(\mathbf{x}, \mathbf{y}, \mathbf{H}). \quad (7)$$

From (7) we see two main challenges to compute the optimal solution:

- 1) The expectation in (7) cannot be obtained in closed form. A numerical implementation is equivalent to finding the numerical integral in $n_t + n_r$ dimensions. This can be extremely hard in high dimensions.
- 2) The size of the optimization space, $|\mathcal{X}|^{n_t}$, can be prohibitively large when the modulation size $|\mathcal{X}|$ and the input dimension n_t become large.

In Section IV, we examine in more details why both of these issues are indeed challenging.

In a conventional MIMO channel without phase noise, finding the ML solution is reduced to solving the following problem

$$\hat{\mathbf{x}}_{\text{ML}}^0(\mathbf{y}, \mathbf{H}) := \arg \min_{\mathbf{x} \in \mathcal{X}^{n_t}} \|\mathbf{y} - \mathbf{H}_0 \mathbf{x}\|^2, \quad (8)$$

which is also called the minimum Euclidean distance detection or nearest neighbor detection (NND). Although the search space in (8) remains large, the expectation is gone. Furthermore, since the objective function is the Euclidean distance, efficient algorithms (e.g., sphere decoder [2] or lattice decoder [1]) exploiting the geometric structure of the problem can be applied without searching over the whole space \mathcal{X}^{n_t} . It is shown in [3] that the sphere decoder has a polynomial average complexity with respect to the input dimension n_t when the channel matrix is drawn i.i.d. from a Rayleigh distribution.

In practice, one may simply ignore the existence of phase noise and still apply (8) to obtain $\hat{\mathbf{x}}_{\text{ML}}^0$ which we refer to as the *naive ML* solution in our work. While this can work relatively well when the phase noise is close to 0, it becomes highly suboptimal with stronger phase noise which is usually the case in high frequency bands with imperfect oscillators. In this paper, we provide a near ML solution by circumventing the two challenges mentioned earlier. We first propose an approximation of the likelihood function. Then we propose an algorithm to solve approximately the optimization problem (7).

III. PROPOSED SCHEME

A. Proposed Approximation of the Likelihood Function

In this section, we propose to approximate the likelihood function with the implicit assumption that the phase noise is not large. Indeed, in practice, the standard deviation of the phase noise is typically smaller than 10 degrees ≈ 0.174 rad. For stronger phase noises, it is no longer reasonable to use QAM and the problem should be addressed differently.

The likelihood function (3) depends on the Euclidean norm $\|\mathbf{y} - \Lambda_{\text{R}} \mathbf{H} \Lambda_{\text{T}} \mathbf{x}\| = \|\Lambda_{\text{R}}^H \mathbf{y} - \mathbf{H} \Lambda_{\text{T}} \mathbf{x}\|$, in which the difference vector $\Lambda_{\text{R}}^H \mathbf{y} - \mathbf{H} \Lambda_{\text{T}} \mathbf{x}$ can be rewritten and approximated as follows

$$\Lambda_{\text{R}}^H \mathbf{y} - \mathbf{H} \Lambda_{\text{T}} \mathbf{x} = \begin{bmatrix} -\mathbf{H} \mathbf{D}_x & \mathbf{D}_y \end{bmatrix} \begin{bmatrix} e^{j\theta_t} \\ e^{-j\theta_r} \end{bmatrix} \quad (9)$$

$$\approx (\mathbf{y} - \mathbf{H} \mathbf{x}) - j[\mathbf{H} \mathbf{D}_x \quad \mathbf{D}_y] \boldsymbol{\theta}, \quad (10)$$

where we define $\mathbf{D}_x := \text{diag}(x_1, \dots, x_{n_t})$, $\mathbf{D}_y := \text{diag}(y_1, \dots, y_{n_r})$, and recall that $\boldsymbol{\theta} := [\boldsymbol{\theta}_t^T \ \boldsymbol{\theta}_r^T]^T$; (10) is from the linear approximation² $e^{j\boldsymbol{\theta}} = 1 + j\boldsymbol{\theta} + o(\boldsymbol{\theta})$. Thus the Euclidean norm has the corresponding real approximation:

$$\|\mathbf{y} - \boldsymbol{\Lambda}_R \mathbf{H} \boldsymbol{\Lambda}_T \mathbf{x}\|^2 \approx \|\mathbf{A}\boldsymbol{\theta} + \mathbf{b}\|^2, \quad (11)$$

where $\mathbf{A} \in \mathbb{R}^{2n_r \times (n_t + n_r)}$ and $\mathbf{b} \in \mathbb{R}^{2n_r \times 1}$ are defined as

$$\mathbf{A} := \begin{bmatrix} \text{Im}(\mathbf{H}\mathbf{D}_x) & \text{Im}(\mathbf{D}_y) \\ -\text{Re}(\mathbf{H}\mathbf{D}_x) & -\text{Re}(\mathbf{D}_y) \end{bmatrix}, \quad \mathbf{b} := \begin{bmatrix} \text{Re}(\mathbf{y} - \mathbf{H}\mathbf{x}) \\ \text{Im}(\mathbf{y} - \mathbf{H}\mathbf{x}) \end{bmatrix}. \quad (12)$$

With the above approximation, we can derive the approximation of the log-likelihood function.

Proposition 1. *Let \mathbf{A} and \mathbf{b} be defined as in (12). Then we have the following approximation of the log-likelihood function $\ln\left(\mathbb{E}_{\boldsymbol{\Theta}}\left[e^{-\gamma\|\mathbf{y}-\mathbf{H}\boldsymbol{\theta}\mathbf{x}\|^2}\right]\right) \approx \hat{f}(\mathbf{x}, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_{\boldsymbol{\theta}})$ with*

$$\hat{f}(\mathbf{x}, \mathbf{y}, \mathbf{H}) := -\gamma \mathbf{b}^T \mathbf{W}_x^{-1} \mathbf{b} - \frac{1}{2} \ln \det(\mathbf{W}_x), \quad (13)$$

where \mathbf{W}_x is defined as

$$\mathbf{W}_x := \mathbf{I} + 2\gamma \mathbf{A} \mathbf{Q}_{\boldsymbol{\theta}} \mathbf{A}^T. \quad (14)$$

Hence, the proposed approximate ML (aML) solution is

$$\hat{\mathbf{x}}_{\text{aML}}(\mathbf{y}, \mathbf{H}) := \arg \min_{\mathbf{x} \in \mathcal{X}^{n_t}} \left\{ \gamma \mathbf{b}^T \mathbf{W}_x^{-1} \mathbf{b} + \frac{1}{2} \ln \det(\mathbf{W}_x) \right\}. \quad (15)$$

Proof. Since we assume that $\boldsymbol{\Theta} \sim \mathcal{N}(0, \mathbf{Q}_{\boldsymbol{\theta}})$, we have

$$\mathbb{E}_{\boldsymbol{\Theta}} \left[e^{-\gamma \|\mathbf{A}\boldsymbol{\Theta} + \mathbf{b}\|^2} \right] = \frac{1}{\sqrt{\det(2\pi\mathbf{Q}_{\boldsymbol{\theta}})}} \int_{\boldsymbol{\theta}} d\boldsymbol{\theta} \exp\left(-\boldsymbol{\theta}^T \underbrace{\left(\gamma \mathbf{A}^T \mathbf{A} + \frac{1}{2} \mathbf{Q}_{\boldsymbol{\theta}}^{-1}\right)}_{\frac{1}{2} \mathbf{Q}^{-1}} \boldsymbol{\theta} - 2\gamma \mathbf{b}^T \mathbf{A} \boldsymbol{\theta} - \gamma \|\mathbf{b}\|^2\right) \quad (16)$$

$$\begin{aligned} &= \sqrt{\frac{\det(\mathbf{Q})}{\det(\mathbf{Q}_{\boldsymbol{\theta}})}} \exp(-\gamma \|\mathbf{b}\|^2 + \gamma^2 \mathbf{b}^T \mathbf{A} (2\mathbf{Q}) \mathbf{A}^T \mathbf{b}) \\ &\quad \cdot \int_{\boldsymbol{\theta}} d\boldsymbol{\theta} \frac{1}{\sqrt{\det(2\pi\mathbf{Q})}} \exp\left(-(\boldsymbol{\theta} + \gamma(2\mathbf{Q})\mathbf{A}^T \mathbf{b})^T \frac{1}{2} \mathbf{Q}^{-1} (\boldsymbol{\theta} + \gamma(2\mathbf{Q})\mathbf{A}^T \mathbf{b})\right) \quad (17) \end{aligned}$$

$$= \sqrt{\frac{\det(\mathbf{Q})}{\det(\mathbf{Q}_{\boldsymbol{\theta}})}} \exp(-\gamma \|\mathbf{b}\|^2 + \gamma^2 \mathbf{b}^T \mathbf{A} (2\mathbf{Q}) \mathbf{A}^T \mathbf{b}) \quad (18)$$

$$= \sqrt{\frac{\det(\mathbf{Q})}{\det(\mathbf{Q}_{\boldsymbol{\theta}})}} \exp(-\gamma \mathbf{b}^T (\mathbf{I} - \gamma \mathbf{A} ((2\mathbf{Q}_{\boldsymbol{\theta}})^{-1} + \gamma \mathbf{A}^T \mathbf{A})^{-1} \mathbf{A}^T) \mathbf{b}) \quad (19)$$

²Here we use, with a slight abuse of notation, $e^{j\boldsymbol{\theta}}$ to denote the vector obtained from the element-wise complex exponential operation. Similarly, the little- o Landau notation $o(\boldsymbol{\theta})$ is element-wise.

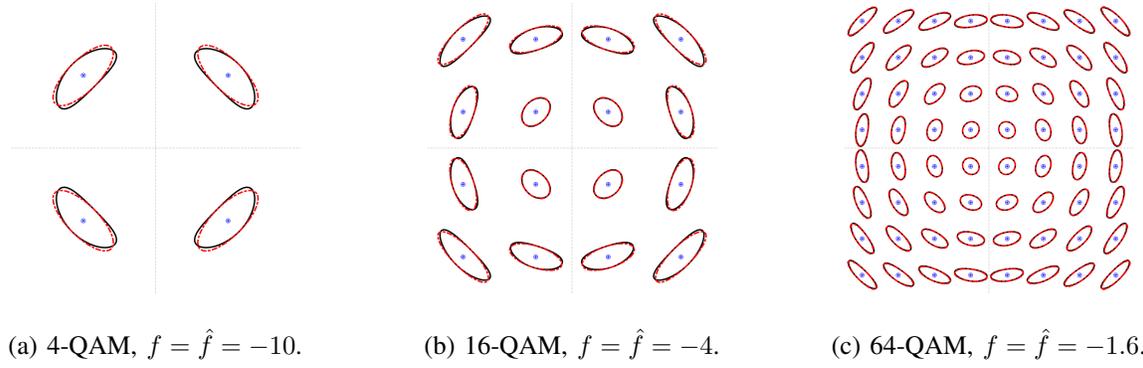


Fig. 1: The proposed approximation of the likelihood function in the scalar case. Solid line is the actual likelihood level set, dashed line is the approximation. Here $\gamma = 30\text{dB}$ and phase noise has standard deviation 3° at the transmitter and at the receiver.

$$= \sqrt{\frac{\det(\mathbf{Q})}{\det(\mathbf{Q}_\theta)}} \exp(-\gamma \mathbf{b}^\top (\mathbf{I} + \gamma \mathbf{A} (2\mathbf{Q}_\theta) \mathbf{A}^\top)^{-1} \mathbf{b}) \quad (20)$$

$$= \frac{1}{\sqrt{\det(\mathbf{I} + 2\gamma \mathbf{A} \mathbf{Q}_\theta \mathbf{A}^\top)}} \exp(-\gamma \mathbf{b}^\top (\mathbf{I} + 2\gamma \mathbf{A} \mathbf{Q}_\theta \mathbf{A}^\top)^{-1} \mathbf{b}), \quad (21)$$

where (18) holds since the integrand in (17) is a pdf with respect to $\boldsymbol{\theta}$; (20) is from the Woodbury matrix identity $(\mathbf{I} + \mathbf{UCV})^{-1} = \mathbf{I} - \mathbf{U}(\mathbf{C}^{-1} + \mathbf{VU})^{-1}\mathbf{V}$. Taking the logarithm and we obtain the approximated log-likelihood function (13). \square

In Figure 1, we illustrate the proposed approximation for 4-, 16-, and 64-QAM. In all three cases, we plot for each constellation point a level set of the likelihood function with respect to “ \mathbf{y} ” in solid line. The level sets of the approximated likelihood function are plotted similarly in dashed line. While the likelihood function is evaluated using numerical integration, the approximation is in closed form given by (13). In this figure, we observe that the approximation is quite accurate, especially for signal points with smaller amplitude. Further, the resemblance of the level sets for the approximate likelihood to ellipsoids suggests that the main contribution in the right hand side of (13) comes from the first term $-\gamma \mathbf{b}^\top \mathbf{W}_x^{-1} \mathbf{b}$. We shall exploit this feature later on to construct the proposed algorithm.

Remark 1. We can check that when $\gamma \mathbf{Q}_\theta \rightarrow 0$, i.e., when the phase noise vanishes faster than the AWGN does, the above solution minimizes $\|\mathbf{b}\|^2 = \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2$, which corresponds to the optimal NND solution in the conventional MIMO case. Nevertheless, when the phase noise does

not vanish, the approximate log-likelihood function (13) depends on \mathbf{x} in a rather complex way, due to the presence of the matrix $\mathbf{W}_{\mathbf{x}}$. Focusing on the term $\mathbf{b}^T \mathbf{W}_{\mathbf{x}}^{-1} \mathbf{b}$, we can think of $\mathbf{W}_{\mathbf{x}}$ as the covariance matrix of some equivalent noise. Indeed, if we approximate the multiplicative phase noise as an additive perturbation, then the perturbation is a self-interference that depends on the input vector \mathbf{x} . This perturbation is not isotropic nor circularly symmetric, and can be captured by the covariance matrix $\mathbf{W}_{\mathbf{x}}$. Some more discussions in this regard will be given in the following subsection.

While the proposed approximation simplifies significantly the objective function, the optimization problem (15) remains hard when the search space is large. For instance, with 64-QAM and 4×4 MIMO, the number of points in \mathcal{X}_t^n is more than 10^7 ! Therefore, we need further simplification by exploiting the structure of the problem.

B. The Self-Interference Whitening Algorithm

As mentioned above, the difficulty of the optimization (15) is mainly due to the presence of the matrix $\mathbf{W}_{\mathbf{x}}$ that depends on \mathbf{x} . Let us first assume that the $\mathbf{W}_{\mathbf{x}}$ corresponding to the optimal solution $\hat{\mathbf{x}}_{\text{aML}}$ were somehow known, and is denoted by $\mathbf{W}_{\hat{\mathbf{x}}}$. Then the optimization problem (15) would be equivalent to

$$\hat{\mathbf{x}}_{\text{aML}}(\mathbf{y}, \mathbf{H}) = \arg \min_{\mathbf{x} \in \mathcal{X}^{n_t}} \left\{ \gamma \mathbf{b}^T \mathbf{W}_{\hat{\mathbf{x}}}^{-1} \mathbf{b} + \frac{1}{2} \ln \det(\mathbf{W}_{\hat{\mathbf{x}}}) \right\} \quad (22)$$

$$= \arg \min_{\mathbf{x} \in \mathcal{X}^{n_t}} \mathbf{b}^T \mathbf{W}_{\hat{\mathbf{x}}}^{-1} \mathbf{b} \quad (23)$$

$$= \arg \min_{\mathbf{x} \in \mathcal{X}^{n_t}} \|\mathbf{W}_{\hat{\mathbf{x}}}^{-\frac{1}{2}} \tilde{\mathbf{y}} - \mathbf{W}_{\hat{\mathbf{x}}}^{-\frac{1}{2}} \tilde{\mathbf{H}} \tilde{\mathbf{x}}\|^2, \quad (24)$$

where $\mathbf{W}_{\hat{\mathbf{x}}}^{-\frac{1}{2}}$ is any matrix such that $(\mathbf{W}_{\hat{\mathbf{x}}}^{-\frac{1}{2}})^H \mathbf{W}_{\hat{\mathbf{x}}}^{-\frac{1}{2}} = \mathbf{W}_{\hat{\mathbf{x}}}^{-1}$; $\tilde{\mathbf{x}}$, $\tilde{\mathbf{y}}$, and $\tilde{\mathbf{H}}$ are defined by

$$\tilde{\mathbf{x}} := \begin{bmatrix} \text{Re}(\mathbf{x}) \\ \text{Im}(\mathbf{x}) \end{bmatrix}, \quad \tilde{\mathbf{y}} := \begin{bmatrix} \text{Re}(\mathbf{y}) \\ \text{Im}(\mathbf{y}) \end{bmatrix}, \quad \tilde{\mathbf{H}} := \begin{bmatrix} \text{Re}(\mathbf{H}) & -\text{Im}(\mathbf{H}) \\ \text{Im}(\mathbf{H}) & \text{Re}(\mathbf{H}) \end{bmatrix}. \quad (25)$$

Note that for a given $\mathbf{W}_{\hat{\mathbf{x}}}$, (24) can be solved efficiently with any NND algorithm. Unfortunately, without knowing the optimal solution $\hat{\mathbf{x}}_{\text{aML}}$, the exact $\mathbf{W}_{\hat{\mathbf{x}}}$ cannot be found. Therefore, the idea is to first estimate the matrix $\mathbf{W}_{\hat{\mathbf{x}}}$ with some suboptimal solution $\hat{\mathbf{x}}$, and then solve the optimization problem (24) with a NND. We call this two-step procedure *self-interference whitening* (SIW).

For instance, we can use the naive ML solution $\hat{\mathbf{x}}_{\text{ML}}^0$ as the initial estimate to obtain $\mathbf{W}_{\hat{\mathbf{x}}}$, and have

$$\hat{\mathbf{x}}'_{\text{aML}}(\mathbf{y}, \mathbf{H}) = \arg \min_{\mathbf{x} \in \mathcal{X}^{n_t}} \|\mathbf{W}_{\hat{\mathbf{x}}_{\text{ML}}^0}^{-\frac{1}{2}} \tilde{\mathbf{y}} - \mathbf{W}_{\hat{\mathbf{x}}_{\text{ML}}^0}^{-\frac{1}{2}} \tilde{\mathbf{H}} \tilde{\mathbf{x}}\|^2. \quad (26)$$

Remark 2. *The intuition behind the SIW scheme is as follows. From the definition of $\mathbf{W}_{\mathbf{x}}$ in (14) and \mathbf{A} in (12), we see that $\mathbf{W}_{\mathbf{x}}$ depends on \mathbf{x} only through $\mathbf{H}\mathbf{D}_{\mathbf{x}}$. First, the column space of $\mathbf{H}\mathbf{D}_{\mathbf{x}}$ does not vary with \mathbf{x} since $\mathbf{D}_{\mathbf{x}}$ is diagonal. Second, a small perturbation of \mathbf{x} does not perturb $\mathbf{W}_{\mathbf{x}}$ too much in the Euclidean space. Since the naive ML point $\hat{\mathbf{x}}_{\text{ML}}^0$ is close to the actual point \mathbf{x} in the column space of \mathbf{H} , it provides an accurate estimate of $\mathbf{W}_{\mathbf{x}}$. This can also be observed on Figure 1c, where we see that the ellipsoid-like dashed lines have similar sizes and orientations for constellation points that are close to each other.*

Remark 3. *Another possible initial estimate is the naive linear minimum mean square error (LMMSE) solution. As the naive ML, the naive LMMSE ignores the phase noise and returns*

$$\hat{\mathbf{x}}_{\text{LMMSE}}^0(\mathbf{y}, \mathbf{H}) := \arg \min_{\mathbf{x} \in \mathcal{X}^{n_t}} \|\mathbf{H}^H(\gamma^{-1}\mathbf{I} + \mathbf{H}\mathbf{H}^H)^{-1}\mathbf{y} - \mathbf{x}\|^2. \quad (27)$$

It is worth mentioning that in the presence of phase noise the naive LMMSE is not necessarily dominated by the naive ML solution, as will shown in the numerical experiments of Section V.

The main algorithm of this work is described in Algorithm 1. In the algorithm, the complex function $\text{NND}(\mathbf{y}, \mathbf{H}, \mathcal{X})$ finds among the points from the alphabet \mathcal{X} the closest one to \mathbf{y} in the column space of \mathbf{H} ; the function $\text{realNND}(\tilde{\mathbf{y}}, \tilde{\mathbf{H}}, \tilde{\mathcal{X}})$ is the real counterpart of NND. The function “ $\text{complex}(\tilde{\mathbf{x}}')$ ” embeds the real vector $\tilde{\mathbf{x}}'$ to the complex space by taking the upper half as the real part and the lower half as the imaginary part. It is worth noting that the newly obtained point is accepted only when it has a higher approximate likelihood value than the naive ML point does. An example of the scalar case is provided in Figure 2 where 256-QAM is used. The transmitted point is x and the received point is y . The solid line is the level set of the likelihood function. If the likelihood function was computed for each point in the constellation, then one would recover x from y successfully. But this would be hard computationally. With the Euclidean detection, \hat{x} that is closer to y than x is would be found instead, which would cause an erroneous detection. The SIW algorithm can “correct” the error as follows. First, to estimate the unknown matrix $\mathbf{W}_{\mathbf{x}}$, we compute the matrix $\mathbf{W}_{\hat{\mathbf{x}}}$ which is represented by the red dashed ellipse around $\hat{\mathbf{x}}$. We can see that the estimate $\mathbf{W}_{\hat{\mathbf{x}}}$ is very close to the correct value $\mathbf{W}_{\mathbf{x}}$,

Algorithm 1 Self-interference whitening

Input: $\mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta$
 Find $\hat{\mathbf{x}}_{\text{LMMSE}}^0$ from (27)
 Find $\hat{\mathbf{x}}_{\text{ML}}^0 \leftarrow \text{NND}(\mathbf{y}, \mathbf{H}, \mathcal{X})$
if $\hat{f}(\hat{\mathbf{x}}_{\text{LMMSE}}^0, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta) > \hat{f}(\hat{\mathbf{x}}_{\text{ML}}^0, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta)$ **then**
 $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}}_{\text{LMMSE}}^0$
else
 $\hat{\mathbf{x}} \leftarrow \hat{\mathbf{x}}_{\text{ML}}^0$
end if
 Generate $\mathbf{W}_{\hat{\mathbf{x}}}$ from $\hat{\mathbf{x}}$ using (12) and (14)
 Find $\mathbf{W}_{\hat{\mathbf{x}}}^{\frac{1}{2}}$ using the Cholesky decomposition
 Generate $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{H}}$ according to (25)
 $\tilde{\mathbf{x}}' \leftarrow \text{realNND}(\mathbf{W}_{\hat{\mathbf{x}}}^{-\frac{1}{2}}\tilde{\mathbf{y}}, \mathbf{W}_{\hat{\mathbf{x}}}^{-\frac{1}{2}}\tilde{\mathbf{H}}\tilde{\mathbf{x}}, \tilde{\mathcal{X}})$
 $\hat{\mathbf{x}}' \leftarrow \text{complex}(\tilde{\mathbf{x}}')$
if $\hat{f}(\hat{\mathbf{x}}', \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta) > \hat{f}(\hat{\mathbf{x}}, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta)$ **then**
 $\hat{\mathbf{x}}'_{\text{aML}} \leftarrow \hat{\mathbf{x}}'$
else
 $\hat{\mathbf{x}}'_{\text{aML}} \leftarrow \hat{\mathbf{x}}$
end if
 Output: $\hat{\mathbf{x}}'_{\text{aML}}^{(2)}$

given by the actual x (blue dashed line). Then, we generate the coordinate system with $\mathbf{W}_{\hat{\mathbf{x}}}$ and search for the closest constellation point to \mathbf{y} in this coordinate system. In this example, x can be recovered successfully. More importantly, computationally efficient NND algorithms can be used to perform the search.

Remark 4. *The complexity of the SIW algorithm is essentially twice that of the NND algorithm used, since the other operations including the LMMSE detection have at most cubic complexity with respect to the dimension of the channel. The complexity of the NND algorithm depends directly on the conditioning of the given matrix. If the columns are close to orthogonal, then channel inversion is almost optimal. However, in the worse case, when the matrix is ill-conditioned, the NND algorithm can be slow and its complexity is exponential in the problem dimension. As*

alternative is to use the Monte-Carlo method, using the following estimate:

$$\hat{f}(\mathbf{x}, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta) \approx \hat{F}_s := \ln \left(\frac{1}{s} \sum_{t=1}^s e^{-\gamma \|\mathbf{y} - \mathbf{H}_{\Theta^{(t)}} \mathbf{x}\|^2} \right), \quad (28)$$

where $\Theta^{(1)}, \dots, \Theta^{(s)}$ are s i.i.d. copies of Θ . Using the delta method [14] and the central limit theorem, the asymptotic variance of this estimator is

$$\text{Var}(\hat{F}_s) \sim \frac{1}{s} \frac{\text{Var}(e^{-\gamma \|\mathbf{y} - \mathbf{H}_{\Theta} \mathbf{x}\|^2})}{\left(\mathbb{E}[e^{-\gamma \|\mathbf{y} - \mathbf{H}_{\Theta} \mathbf{x}\|^2}] \right)^2}, \quad s \rightarrow \infty. \quad (29)$$

Let us compute this quantity in the (arguably easiest) case where \mathbf{H} is the identity matrix with $n_t = n_r = n$:

$$\|\mathbf{y} - \mathbf{H}_{\Theta} \mathbf{x}\|^2 = \sum_{k=1}^n |y_k - x_k e^{j(\Theta_{r,k} + \Theta_{t,k})}|^2. \quad (30)$$

Define the one-dimensional likelihood

$$g(x, y, \gamma, \sigma) := \mathbb{E} \left[e^{-\gamma |y - x e^{j(\Theta_r + \Theta_t)}|^2} \right], \quad (31)$$

and using independence we calculate the moments:

$$\mathbb{E} \left[e^{-\gamma \|\mathbf{y} - \mathbf{H}_{\Theta} \mathbf{x}\|^2} \right] = \prod_{k=1}^n g(x_k, y_k, \gamma, \sigma), \quad (32)$$

$$\text{Var}(e^{-\gamma \|\mathbf{y} - \mathbf{H}_{\Theta} \mathbf{x}\|^2}) = \prod_{k=1}^n g(x_k, y_k, 2\gamma, \sigma) - \prod_{k=1}^n g(x_k, y_k, \gamma, \sigma)^2. \quad (33)$$

From (29), the asymptotic variance is hence:

$$\text{Var}(\hat{F}_s) = \frac{1}{s} \left(\prod_{k=1}^n v(x_k, y_k, \gamma, \sigma) - 1 \right). \quad (34)$$

where $v(x, y, \gamma, \sigma) := \frac{g(x, y, 2\gamma, \sigma)}{g(x, y, \gamma, \sigma)^2}$. It is noted that, unless $\sigma_r^2 + \sigma_t^2 = 0$ or $x = 0$, or $y = 0$, the random variable $e^{-\gamma |y - x e^{j(\Theta_r + \Theta_t)}|^2}$ cannot be a constant. Thus, we have $v(x_k, y_k, \gamma, \sigma) > 1$ for $k = 1, \dots, n$. As a result, the asymptotic error (34) grows exponentially with the dimension n , so that the Monte-Carlo method is infeasible in high dimensions, since the number of samples must also scale exponentially with n to maintain a constant error.

For instance, consider $x_k = y_k = 1$ for all k , $\gamma = 40\text{dB}$, $n_r = n_t = 20$, and $\sigma_r = \sigma_t = 3$ degrees. Then $v \approx 7$, so that, to obtain an error smaller than 0.1, one would require $s \approx 10^{18}$ samples, which is clearly not feasible in practice.

B. Hardness of maximum likelihood search

Assume that one is able to estimate the value of the likelihood with high accuracy (as seen above this is typically hard), and denote by \bar{f} this value. We may then consider the following algorithm. Given the received symbol \mathbf{y} , and a radius ρ , one first computes the set of points $\mathcal{S} := \{\mathbf{x} \in \mathcal{X}^{n_t} : \|\mathbf{y} - \mathbf{H}\mathbf{x}\|^2 \leq \rho^2\}$, then one computes the value of \bar{f} for each of those points and returns the point maximizing \bar{f} . In the large system limit, the following concentration phenomenon occurs.

Proposition 2. *Assume that \mathbf{H} has i.i.d. entries with distribution $\mathcal{CN}(0, 1)$ and \mathbf{X} is chosen uniformly at random from \mathcal{X}^{n_t} . Define the radius $R^2 := \|\mathbf{Y} - \mathbf{H}\mathbf{X}\|^2$. Then we have:*

$$\mathbb{E}[R^2] = 2n_t n_r \left(1 - e^{-\frac{\sigma_r^2 + \sigma_t^2}{2}}\right) + \gamma^{-1} n_r, \quad (35)$$

and for any $\eta > 0$ we have:

$$\mathbb{P}\left\{(1 - \eta)\mathbb{E}[R^2] \leq R^2 \leq (1 + \eta)\mathbb{E}[R^2]\right\} \rightarrow 1, \quad n_t, n_r \rightarrow \infty. \quad (36)$$

Note that the above result is general and does not impose that the number of antennas n_r, n_t scale at the same speed. We draw two conclusions from this result: (i) Any such algorithm applied with radius $(1 + \eta)\sqrt{\mathbb{E}[R^2]}$ for any $\eta > 0$ is guaranteed to inspect the optimal point with high probability. (ii) Any such algorithm which has a large success probability needs to inspect every point in a sphere of radius $O\left(\sqrt{n_t n_r} \sqrt{1 - \exp\left(-\frac{\sigma_r^2 + \sigma_t^2}{2}\right)}\right)$, and therefore typically has a very high complexity.

From the above analysis we see that the second difficulty of the decoding problem for phase noise channels, even when the likelihood can be computed, lies in the number of points to be inspected which is exponentially large in n_r, n_t . The problem is that computing the likelihood at any given point does not give us any information about the structure of f , and does not help in maximizing f efficiently.

C. Non-concavity in the high SNR regime

Indeed maximizing f is difficult, and it seems that even performing zero-forcing, i.e., maximizing f over $x \in \mathbb{C}^{n_t}$ rather than over $\mathbf{x} \in \mathcal{X}^{n_t}$, is difficult since f is non-concave, at least in the high SNR regime. We first show that, in the high SNR regime, the log-likelihood can be approximated by a function of the minimal value of $\|\mathbf{y} - \mathbf{H}_\theta \mathbf{x}\|^2$, where the minimum is taken over all possible phases θ .

Lemma 1. For all $\mathbf{x}, \mathbf{y}, \mathbf{H}, \mathbf{Q}_\theta$, we have the following high SNR behavior

$$\lim_{\gamma \rightarrow \infty} -\frac{1}{\gamma} f(\mathbf{x}, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta) = m(\mathbf{x}, \mathbf{y}, \mathbf{H}) := \min_{\boldsymbol{\theta} \in \mathbb{R}^{n_t+n_r}} \|\mathbf{y} - \mathbf{H}_\boldsymbol{\theta} \mathbf{x}\|^2. \quad (37)$$

Proof. Consider $\epsilon > 0$, we have that

$$\mathbb{1} \{ \|\mathbf{y} - \mathbf{H}_\boldsymbol{\theta} \mathbf{x}\|^2 \leq m(\mathbf{x}, \mathbf{y}, \mathbf{H}) + \epsilon \} e^{-\gamma(m(\mathbf{x}, \mathbf{y}, \mathbf{H}) + \epsilon)} \leq e^{-\gamma \|\mathbf{y} - \mathbf{H}_\boldsymbol{\theta} \mathbf{x}\|^2} \leq e^{-\gamma m(\mathbf{x}, \mathbf{y}, \mathbf{H})}. \quad (38)$$

Taking expectations and then logarithms:

$$-(m(\mathbf{x}, \mathbf{y}, \mathbf{H}) + \epsilon) + \frac{1}{\gamma} \ln \mathbb{P} \{ \|\mathbf{y} - \mathbf{H}_\boldsymbol{\theta} \mathbf{x}\|^2 \leq m(\mathbf{x}, \mathbf{y}, \mathbf{H}) + \epsilon \} \leq \frac{1}{\gamma} f(\mathbf{x}, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta) \leq -m(\mathbf{x}, \mathbf{y}, \mathbf{H}).$$

Since the mapping $\boldsymbol{\theta} \mapsto \|\mathbf{y} - \mathbf{H}_\boldsymbol{\theta} \mathbf{x}\|^2$ is continuous, for any given $\epsilon > 0$, the probability

$\mathbb{P} \{ \|\mathbf{y} - \mathbf{H}_\boldsymbol{\theta} \mathbf{x}\|^2 \leq m(\mathbf{x}, \mathbf{y}, \mathbf{H}) + \epsilon \}$ is strictly positive. Letting $\gamma \rightarrow \infty$, we have

$$-(m(\mathbf{x}, \mathbf{y}, \mathbf{H}) + \epsilon) \leq \liminf_{\gamma \rightarrow \infty} \frac{1}{\gamma} f(\mathbf{x}, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta) \leq \limsup_{\gamma \rightarrow \infty} \frac{1}{\gamma} f(\mathbf{x}, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta) \leq -m(\mathbf{x}, \mathbf{y}, \mathbf{H}).$$

Since the above holds for all $\epsilon > 0$, we have $\lim_{\gamma \rightarrow \infty} -\frac{1}{\gamma} f(\mathbf{x}, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta) = m(\mathbf{x}, \mathbf{y}, \mathbf{H})$. \square

We can now show that in general the log-likelihood is not concave, hence maximizing it is not straightforward.

Proposition 3. For γ large enough and $\mathbf{H} \neq \mathbf{0}$, there exists \mathbf{y} such that $\mathbf{x} \mapsto f(\mathbf{x}, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta)$ is a non-concave function.

Proof. Assume that f is concave, then for all \mathbf{x} we must have:

$$\frac{1}{2} (f(\mathbf{x}, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta) + f(\mathbf{x}^*, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta)) \leq f\left(\frac{\mathbf{x} + \mathbf{x}^*}{2}, \mathbf{y}, \mathbf{H}, \gamma, \mathbf{Q}_\theta\right). \quad (39)$$

From Lemma 1, the above inequality implies that

$$\frac{1}{2} (m(\mathbf{x}, \mathbf{y}, \mathbf{H}) + m(\mathbf{x}^*, \mathbf{y}, \mathbf{H})) \geq m\left(\frac{\mathbf{x} + \mathbf{x}^*}{2}, \mathbf{y}, \mathbf{H}\right). \quad (40)$$

We shall construct an example to show that the above does not hold in general. Consider any \mathbf{z} such that $\mathbf{H}\mathbf{z} \neq \mathbf{0}$, and let $\mathbf{y} = \mathbf{H}\mathbf{z}$ and $\mathbf{x} = j(|z_1|, \dots, |z_{n_t}|)$, so that $\mathbf{x}^* = -\mathbf{x}$. Then \mathbf{x} and \mathbf{z} are equal up to a phase transformation, so $m(\mathbf{x}, \mathbf{y}, \mathbf{H}) = m(\mathbf{z}, \mathbf{y}, \mathbf{H}) = 0$. Similarly $m(\mathbf{x}^*, \mathbf{y}, \mathbf{H}) = 0$. By definition $m(\frac{\mathbf{x} + \mathbf{x}^*}{2}, \mathbf{y}, \mathbf{H}) = m(\mathbf{0}, \mathbf{y}, \mathbf{H}) = \|\mathbf{y}\|^2$. In this example, (40) would imply $0 \geq \|\mathbf{y}\|^2$, which is clearly a contradiction since $\mathbf{y} \neq \mathbf{0}$. Hence f cannot be concave for γ large enough. \square

This fact that the log-likelihood is in general non-concave gives another important insight: SIW can in fact be seen as a (well chosen) concave approximation of a non-concave function. To circumvent the problem of non-concavity of the log-likelihood SIW approximates it by a

function which, when $W_{\mathbf{x}}$ is fixed, is concave. As non-concavity appears mainly for high SNRs, the discrepancy between the performance of ML and SIW (if any) should be more visible in the high SNR regime, and this will be confirmed by our numerical experiments in the next section.

V. NUMERICAL EXPERIMENTS

In this section, we look at different communication scenarios in which we compare the proposed scheme to some baseline schemes, including two schemes that ignore the phase noise:

- The naive LMMSE solution given by (27)
- The naive ML solution (8)

and a scheme that takes the phase noise into account:

- Selection between naive LMMSE and ML: the receiver first finds the naive LMMSE and naive ML solutions, then computes the proposed approximate likelihood function and selects the one with higher value.

Note that we focus on the vector detection error rate³ as our performance metric.

A. Simulation-based lower bounds

In order to appreciate the performance of the proposed algorithm, we need to compare it not only with the existing schemes, but also to the fundamental limit given by ML detection, which is optimal. Let us recall that the proposed SIW algorithm may suffer from two levels of suboptimality. First, the approximate likelihood function (13) may be inaccurate in some cases. Second, even if (13) is accurate, the SIW algorithm is not guaranteed to find the optimal solution (15). Therefore, in order to identify the source of the potential suboptimality, it would be useful to compare the performance of the SIW scheme with the performance given by (13) and with that given by (7).

Unfortunately, as pointed out earlier, finding (13) requires an exhaustive search with complexity growing as $|\mathcal{X}|^{n_t}$. Finding (7) is even harder because of the numerical multi-dimensional integration. As such, we resort to lower bounds on the detection error for (13) and (7), respectively, which are enough for our purpose of benchmarking. To that end, we write

$$P_e^{\text{aML}} \geq \mathbb{P} \left\{ \hat{f}(\mathbf{X}, \mathbf{Y}, \mathbf{H}) < \max_{\mathbf{x} \in \mathcal{X}^{n_t}} \hat{f}(\mathbf{x}, \mathbf{Y}, \mathbf{H}) \right\} \quad (41)$$

³The detection is considered successful only when all the symbols in \mathbf{x} are recovered correctly. Otherwise an error is declared.

$$\geq \mathbb{P} \left\{ \hat{f}(\mathbf{X}, \mathbf{Y}, \mathbf{H}) < \max_{\mathbf{x} \in \mathcal{L}} \hat{f}(\mathbf{x}, \mathbf{Y}, \mathbf{H}) \right\}, \quad \forall \mathcal{L} \subseteq \mathcal{X}^m, \quad (42)$$

where the first lower bound is from the definition of the detection criterion, namely, error occurs if there exists at least one input vector that has a strictly higher approximate likelihood value. Note that although the second inequality is valid for all \mathcal{L} , it becomes equality if \mathcal{L} contains *all* the points in \mathcal{X}^m that have a higher approximate likelihood value than \mathbf{X} does. In this work, we only take a large set around \mathbf{X} to obtain the lower bound (42), without any theoretical guarantee of tightness of (42). Similarly, for ML detection, we have

$$P_e^{\text{ML}} \geq \mathbb{P} \left\{ f(\mathbf{X}, \mathbf{Y}, \mathbf{H}) < \max_{\mathbf{x} \in \mathcal{X}^m} f(\mathbf{x}, \mathbf{Y}, \mathbf{H}) \right\} \quad (43)$$

$$\geq \mathbb{P} \{ f(\mathbf{X}, \mathbf{Y}, \mathbf{H}) < f(\mathbf{X}', \mathbf{Y}, \mathbf{H}) \} \quad (44)$$

$$= \mathbb{P} \{ \mathbf{X} \neq \mathbf{X}', f(\mathbf{X}, \mathbf{Y}, \mathbf{H}) < f(\mathbf{X}', \mathbf{Y}, \mathbf{H}) \}, \quad \forall \mathbf{X}' \in \mathcal{X}^m, \quad (45)$$

where the first lower bound is from the definition of the ML detection criterion, namely, error occurs if there exists at least one input vector that has a strictly higher likelihood value; the equality (45) holds since $\mathbf{X} \neq \mathbf{X}'$ is a consequence of $f(\mathbf{X}, \mathbf{Y}, \mathbf{H}) < f(\mathbf{X}', \mathbf{Y}, \mathbf{H})$; Note that the second lower bound (44) holds for any vector \mathbf{X}' from the alphabet \mathcal{X}^m and with equality when \mathbf{X}' is the exact ML solution. Since the ML solution is unknown, we can use any suboptimal solution instead and still obtain a valid lower bound. Now we can see that (45) is much easier to evaluate than (43) is since there is no need to perform the maximization over \mathcal{X}^m . Intuitively, if \mathbf{X}' is a near ML solution, then the lower bound should be tight enough. We shall have some more discussions on this with the upcoming numerical examples. The lower bound (45) can be obtained by simulation:

- 1) For a given observation \mathbf{y} and channel \mathbf{H} , find a suboptimal solution \mathbf{x}' .
- 2) If $\mathbf{x}' \neq \mathbf{x}$, compute $f(\mathbf{x}, \mathbf{y}, \mathbf{H})$ and $f(\mathbf{x}', \mathbf{y}, \mathbf{H})$, count an ML error only when $f(\mathbf{x}, \mathbf{y}, \mathbf{H}) < f(\mathbf{x}', \mathbf{y}, \mathbf{H})$; otherwise the counter remains unchanged.

With the proposed method, we need to perform twice the numerical integration (e.g., Monte-Carlo integration) only when $\mathbf{x}' \neq \mathbf{x}$. If the latter event happens with small probability, then the average complexity to evaluate (45) is low. In other words, using a \mathbf{x}' from a better reference scheme not only makes the lower bound tighter but also makes it easier to evaluate.

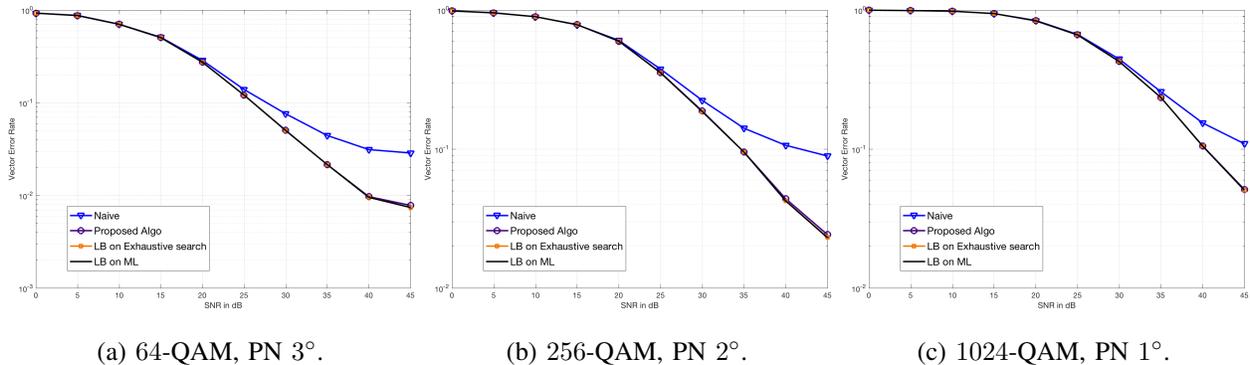


Fig. 3: SISO Rayleigh fading with i.i.d. phase noise.

B. Scenario 1: Point-to-point SISO channel

The first scenario focuses on the point-to-point Rayleigh fading single-antenna channels, also known as single-input single-output (SISO) channels. We consider three different modulation orders (64, 256, and 1024) with correspondingly three values of phase noise strength (3° , 2° , and 1°) in terms of the standard deviation at both the transmitter and receiver sides. The idea is to assess the performance of the proposed algorithm in different phase noise limited regimes. In the SISO case, we compare the proposed scheme with the naive ML scheme which consists in a simple threshold detection for the real and imaginary parts. Several remarks are in order. First, we see that ignoring the existence of phase noise incurs a significant performance loss. Second, if exhaustive search is done with the proposed likelihood approximation, then it achieves the ML performance. This can be seen from the fact that the proposed simulation-based lower bound overlaps with the curve with exhaustive search. This confirms the accuracy of the closed-form approximation (13) at least in the SISO case. It is worth mentioning that the exact likelihood in this case can also be derived via the Tikhonov distribution as shown in [5], [15] where the analytic expression involving the Bessel function has been provided. Finally, more remarkably, the SIW algorithm almost achieves the ML performance without exhaustive search.

C. Scenario 2: Point-to-point LoS-MIMO channel

The second scenario is the point-to-point line-of-sight (LoS) MIMO system commonly deployed as microwave backhaul links [16], [9], [17]. We assume that the channel is constant over time but each antenna is driven by its own oscillator. This is the worst-case assumption but also often motivated by the fact that the communication distance is large and thus the distance

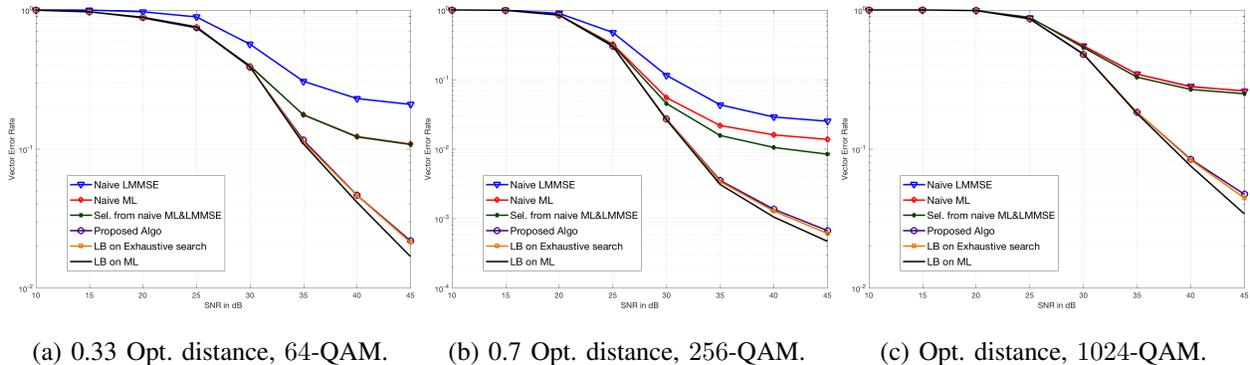


Fig. 4: 4×4 LoS MIMO. Each antenna has i.i.d. phase noise of standard deviation 1° .

between antenna elements is increased accordingly to make sure that the channel matrix is well conditioned [16], [17]. Here, we adopt the model with two transmit and two receive antennas each one with dual polarizations. This is effectively a 4×4 MIMO channel. The optimal distance between the antenna elements at each side can be derived as a function of the communication distance [16]. However, it may not always be possible to install the antennas with the optimal spacing due to practical constraints. The condition number of the channel matrix increases when the antenna spacing decreases away from the optimal distance. In Figure 4, we consider three configurations with distances 0.33, 0.7, and 1 of the optimal value, generated using the model from [17]. Accordingly, we use 64-, 256-, and 1024-QAM. For simplicity, we do not consider any precoding although it may further improve the performance as shown in [17]. We assume that the phase noises are i.i.d. with standard deviation of 1° . We make the following observations. First, as in the SISO case, phase noise mitigation substantially improves performance. Also, the proposed likelihood approximation remains accurate as shown by the comparison between the exhaustive search (15) and the lower bound on ML detection. Further, the proposed SIW algorithm follows closely the exhaustive search curve and hence achieves near ML performance. Finally, although in the considered scenario the naive LMMSE is outperformed by the naive ML scheme, the selection between them can provide a non-negligible gain as shown in Figure 4b.

D. Scenario 3: Uplink SIMO channel with centralized receiver oscillator

The third scenario is the uplink cellular communication channel with four single-antenna users and one multi-antenna base station receiver. It is assumed that the phase noises at the users' side are i.i.d., whereas there is no phase noise at the receiver side. This is a reasonable

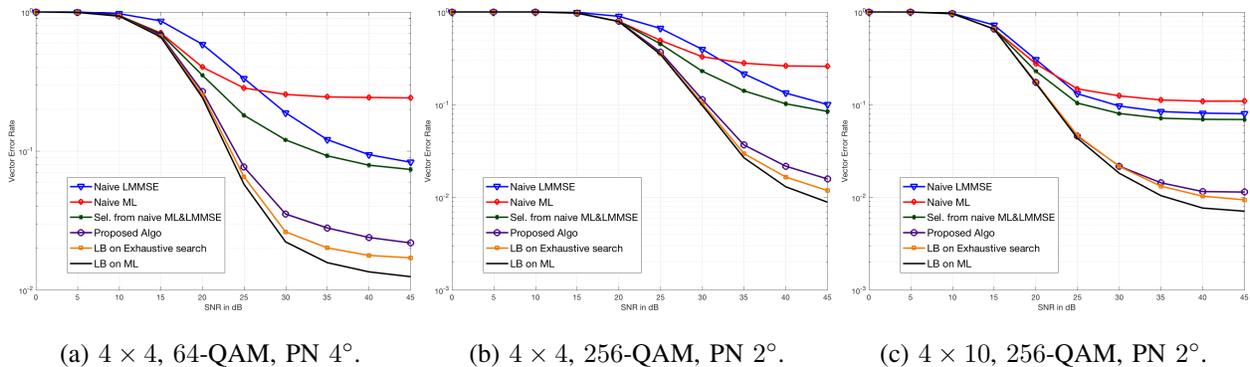


Fig. 5: Uplink SIMO channel. Four single-antenna users with i.i.d. phase noise, multi-antenna receiver without phase noise.

assumption since the oscillators at the base station are usually of higher quality than those used by mobile devices. We assume i.i.d. Rayleigh fading in this scenario where three configurations are considered, as shown in Figure 5. Unlike in the previous scenarios, the naive ML is dominated by the naive LMMSE at high SNR. This somewhat counter-intuitive observation can be explained as follows. Without receiver phase noise, the channel can be inverted and we obtain a spatial parallel channel. Although channel inversion incurs some power loss when the channel is not orthogonal, each of the resulting parallel subchannels sees an independent phase noise. Therefore, the demodulation only suffers from a scalar self-interference. On the other hand, with naive ML the receiver tries to find the closest vector in the image space of \mathbf{H} to the received vector \mathbf{y} . Since the linear map \mathbf{H} mixes the perturbation of the different transmit phase noises, the naive ML detection, ignoring the presence of phase noise, suffers from the aggregated perturbation from all the phase noises. That is why the naive LMMSE can be better than the naive ML scheme in the high SNR regime where phase noise dominates the additive noise. In the case when both the transmitter and the receiver have comparable phase noises, such a phenomenon is rarely observed since the channel inversion also increases the perturbation with the presence of receiver phase noises.

From Figure 5, we remark that as before the proposed SIW scheme is superior to the other schemes. With a relatively strong phase noise of 4° , the error rate of SIW is 3 to 4 times lower than that of the naive schemes and is at most twice that of the ML lower bound. With a smaller phase noise of 2° , the SIW scheme can support 256-QAM with four receive antennas, achieving an error rate 5 times lower than that of the naive schemes. In Figure 5c, we increase the number

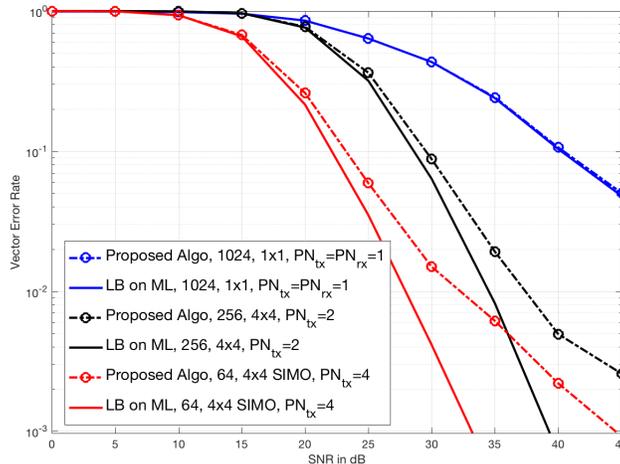


Fig. 6: Uniform phase noise at both transmitter and receiver, Rayleigh fading.

of receive antennas to 10, we see that the gap between the naive schemes is decreased due to the increased orthogonality of the channel. Nevertheless, the gap between the naive schemes and the proposed scheme does not decrease since the orthogonality between the users does not reduce the impact of the phase noise from the transmitter side. We could expect the same observation even with massive MIMO. Nevertheless, with massive MIMO uplink, the receiver phase noise can be mitigated substantially due to the asymptotic orthogonality [18].

VI. FURTHER DISCUSSION AND EXPERIMENTS

A. Robustness to the phase noise distribution

One of the main assumptions of our work is that the phase noise follows a Gaussian distribution. Indeed, our derivation of the closed-form approximation (13) depends on this assumption. We have seen that this approximation is very accurate in various practical scenarios with Gaussian phase noise. In practice, however, phase noises may not be Gaussian, which leads to the following natural question on the robustness: *Does the proposed algorithm still work when the phase noise is not Gaussian?* The answer turns out to be positive when we let the phase noise be uniformly distributed. In Figure 6, we consider three previous configurations (shown in Figure 3c, 5a, and 5b, respectively) but with uniform phase noises. From the results, we see that the proposed algorithm works well as in the Gaussian phase noise, especially when the phase noise is small. In fact, we believe that the phase noise distribution with certain regularity (e.g., continuous and

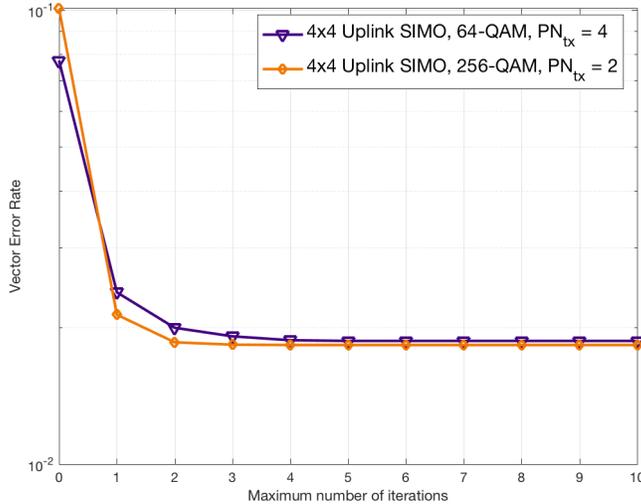


Fig. 7: Impact of iterations.

bounded density) should not have great impact on the performance of the proposed algorithm when the phase noise is not too strong.

B. Further improvement with more iterations

Although the proposed scheme achieves near ML performance in many cases, there is still room for improvement in some situations. As shown in Figure 5 for the 4×4 channel with 64-QAM, the error floor of the proposed scheme is three times higher than the one of the exhaustive search. In order to reduce the error floor, we propose to introduce a list of candidate solutions by iteration. We recall that in the SIW scheme we start from the best solution between naive ZF and naive ML, and we use this solution as a starting point to estimate the matrix \mathbf{W}_x followed by a nearest neighbor detection. The SIW algorithm replaces the starting point with the newly found one if the latter has a higher approximated likelihood. We can extend the SIW algorithm with more iterations. Specifically, we can fix a maximum number of iterations. As long as the newly found point is not inside the list, it should be added to the list and we continue to iterate. The procedure stops either when we hit the maximum number of iterations or we find a point already inside the list. At the end, we select the point with the highest likelihood value from the list.

C. Potential issues in the very high SNR regime

At very high SNR, the proposed approximation (13) may become less accurate and the performance loss increases with SNR. It can be seen from (6) that the approximation error of $\|\mathbf{y} - \mathbf{H}_\Theta \mathbf{x}\|$ has greater impact as γ grows. Numerical experiments show that such performance loss may be apparent for SNR higher than 40dB depending on the constellation size. A straightforward workaround is to impose a ceiling value of the SNR γ_{\max} for the decoder. In other words, for SNR higher than the threshold γ_{\max} , we let the decoder work as if it were at γ_{\max} . Intuitively, the probability of error cannot be larger than the one at γ_{\max} , since we are feeding less noisy observations to the decoder than what they are supposed to be. With the decoding function, the observation space can be partitioned into $|\mathcal{X}|^m$ regions, each one of which corresponds to a vector in \mathcal{X}^m . The probability of decoding error is the probability that the observation is outside of the decoding region corresponding to the actual input. With a smaller variance of AWGN, the observation has a higher probability to be inside the region, hence a lower probability of error. Nevertheless, the formal proof of this argument is not trivial, and is outside of the scope of the current paper.⁴

Another issue at very high SNR is that the likelihood may be too small as compared to the finite numerical precision. Therefore, it becomes impossible to obtain the simulation-based lower bound in a reliable way. Furthermore, the number of Monte-Carlo samples required to reach any given accuracy grows with the SNR.

D. On the practical validity of the adopted discrete-time model

The discrete-time channel model (1) that we adopt in this work is a simplification of the waveform phase noise channel. Indeed, the discrete-time output sequence is obtained from filtering followed by sampling in most communication systems. Filtering a waveform corrupted by phase noise results in not only phase perturbation but also amplitude variation [19]. In the following, we shall show that the amplitude variation is negligible in the practical regime of interest. For simplicity, we focus on the single-antenna case with a rectangular filter (i.e., an integrator). We adhere to the commonly accepted Wiener model for the continuous-time phase

⁴We can always add an artificial noise to the observation in order to reduce the SNR to γ_{\max} before the detection. In this way we have an error floor for SNR beyond γ_{\max} .

noise process $\{\Theta(t)\}$. In particular, $\Theta(t) \sim \mathcal{N}(0, \beta t)$. The equivalent filtered channel gain for the k th symbol interval of duration T is

$$\frac{1}{T} \int_{kT}^{(k+1)T} e^{j\Theta(t)} dt \stackrel{d}{=} e^{j\Theta(kT)} \frac{1}{T} \int_0^T e^{j\Theta(t)} dt \quad (46)$$

from the property of a Wiener process where $\stackrel{d}{=}$ means equality in distribution. Assuming that we can somehow track the *past state* $\Theta(kT)$ perfectly, we now focus on the following random variable due to the residual phase noise corresponding to the *innovation part*:

$$B(\beta, T) := \frac{1}{T} \int_0^T e^{j\Theta(t)} dt \quad (47)$$

$$\stackrel{d}{=} \frac{1}{\beta T} \int_0^{\beta T} e^{j\tilde{\Theta}(t)} dt \quad (48)$$

$$\stackrel{d}{=} B(1, \beta T) \quad (49)$$

where in (48) follows from $\Theta(t) \stackrel{d}{=} \tilde{\Theta}(\beta t)$ for some normalized Wiener process $\{\tilde{\Theta}(t)\}$ with $\tilde{\Theta}(t) \sim \mathcal{N}(0, t)$. We notice that the random variable $B(\beta, T)$ depends on the parameters β and T only through the product βT . The distribution of $B(\beta, T)$ has been characterized both approximately [19] and exactly [20]. In particular, an approximate moment-generating function has been derived in [19] for small phase noise. We can use the results therein to obtain the following characterization.

Proposition 4. *Let us consider the polar representation $B(\beta, T) = Ge^{j\Phi}$ with $G \geq 0$ and $\Phi \in [-\pi, \pi)$. When $S := \beta T$ is small, we have*

$$\text{Var}(\Phi) \approx \frac{S}{3}, \quad \text{Var}(G) \approx \frac{S^2}{180}, \quad (50)$$

and thus,

$$\text{Var}(G) \approx \frac{1}{20} [\text{Var}(\Phi)]^2. \quad (51)$$

In Figure 8, we compare the approximation given by (51) to the correct value obtained numerically. The approximation (51) is surprisingly accurate even for a standard deviation of 20° for the phase. More importantly, such results show that for a small phase perturbation up to 5 degrees – this is the regime of interest in the present work – the amplitude variation is less than -55 dB. Therefore, the interference caused by the amplitude variation is dominated by the AWGN and can be treated as noise without any performance loss. In other words, the discrete-time model adopted in this work – ignoring the amplitude variation – is indeed valid for phase noise with moderate variance.

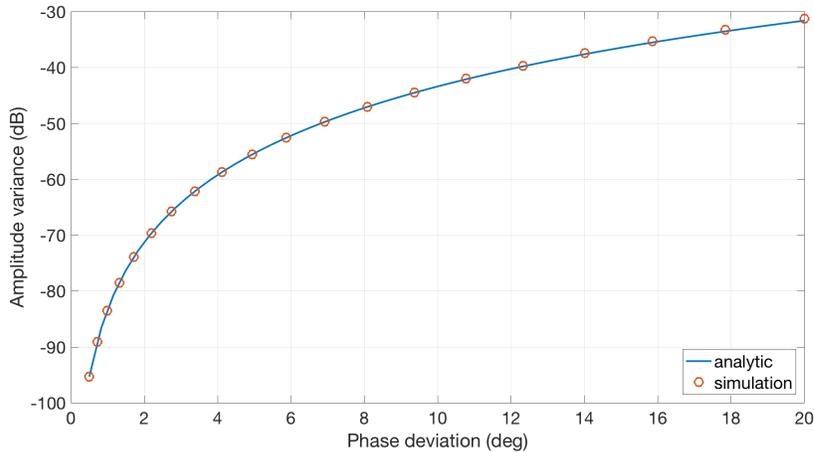


Fig. 8: Amplitude vs phase perturbation for filtered channel gain.

E. Constellation design

The purpose of the current work is to design a detector that takes into account the phase perturbation for existing systems in which typical QAM signaling is used. As we observe from the numerical experiments, in some of the scenarios, even the lower bound on the ML detection error exhibits an error floor. The error floor is however not a fundamental limit of either the detector or the channel. With a carefully designed signaling scheme, the probability of ML detection error can be arbitrarily small with an increasing SNR. For instance, if we use amplitude modulation, then when the SNR grows, the amplitude ambiguity decreases and the detection error vanishes. The cost of such an extreme scheme is a reduced spectral efficiency. Algorithms to design a constellation based on the statistics of both the additive and phase noises have been proposed in the literature (see [5], [21] and the references therein). To the best of the authors' knowledge, only SISO has been considered so far. In the MIMO case, such constellations can still be used and should provide improvements over QAM. The difficulty with non-QAM constellation lies in the MIMO detection part, since efficient NND algorithms cannot be applied directly. The constellation design problem for MIMO phase noise channels is a challenging and interesting problem in its own right, which is however out of the scope of the current work.

VII. CONCLUSIONS

In this work, we have studied the ML detection problem for uncoded MIMO phase noise channels. We have proposed an approximation of the likelihood function that has been shown

to be accurate in the regimes of practical interest. More importantly, thanks to the geometric interpretation of the approximate likelihood function, we have designed a simple algorithm that can solve approximately the optimization problem with only two nearest neighbor detections. Numerical experiments show that the proposed algorithm can greatly mitigate the impact of phase noises in different communication scenarios.

REFERENCES

- [1] E. Agrell, T. Eriksson, A. Vardy, and K. Zeger, "Closest point search in lattices," *IEEE Trans. Inf. Theory*, vol. 48, no. 8, pp. 2201–2214, Aug. 2002.
- [2] E. Viterbo and J. Boutros, "A universal lattice code decoder for fading channel," *IEEE Trans. Inf. Theory*, vol. 45, no. 10, pp. 1639–1642, Jul. 1999.
- [3] B. Hassibi and H. Vikalo, "On the sphere-decoding algorithm I. Expected complexity," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2806–2818, Aug. 2005.
- [4] J. Jaldén and P. Elia, "DMT optimality of LR-aided linear decoders for a general class of channels, lattice designs, and system models," *IEEE Trans. Inf. Theory*, vol. 56, no. 10, pp. 4765–4780, Oct. 2010.
- [5] G. J. Foschini, R. D. Gitlin, and S. B. Weinstein, "On the selection of a two-dimensional signal constellation in the presence of phase jitter and gaussian noise," *Bell Labs Technical Journal*, vol. 52, no. 6, pp. 927–965, 1973.
- [6] A. Hajimiri and T. H. Lee, "A general theory of phase noise in electrical oscillators," *IEEE Journal of Solid-State Circuits*, vol. 33, no. 2, pp. 179–194, Feb. 1998.
- [7] H. Ghozlan and G. Kramer, "Models and information rates for wiener phase noise channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2376–2393, Apr. 2017.
- [8] A. Demir, A. Mehrotra, and J. Roychowdhury, "Phase noise in oscillators: A unifying theory and numerical methods for characterization," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 47, no. 5, pp. 655–674, May 2000.
- [9] G. Durisi, A. Tarable, C. Camarda, R. Devassy, and G. Montorsi, "Capacity bounds for MIMO microwave backhaul links affected by phase noise," *IEEE Trans. Commun.*, vol. 62, no. 3, pp. 920–929, Mar. 2014.
- [10] S. Yang and S. Shamai (Shitz), "On the multiplexing gain of discrete-time MIMO phase noise channels," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2394–2408, Apr. 2017.
- [11] R. Krishnan, G. Colavolpe, A. G. i Amat, and T. Eriksson, "Algorithms for joint phase estimation and decoding for MIMO systems in the presence of phase noise and quasi-static fading channels," *IEEE Trans. Signal Process.*, vol. 63, no. 13, pp. 3360–3375, July 2015.
- [12] T. Datta and S. Yang, "Improving MIMO detection performance in presence of phase noise using norm difference criterion," in *53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, Sep. 2015, pp. 286–292.
- [13] H. Mehrpouyan, A. A. Nasir, S. D. Blostein, T. Eriksson, G. K. Karagiannidis, and T. Svensson, "Joint estimation of channel and oscillator phase noise in MIMO systems," *IEEE Trans. Signal Process.*, vol. 60, no. 9, pp. 4790–4807, Sep. 2012.
- [14] G. W. Oehlert, "A note on the delta method," *The American Statistician*, vol. 46, no. 1, pp. 27–29, 1992.
- [15] G. Colavolpe, A. Barbieri, and G. Caire, "Algorithms for iterative decoding in the presence of strong phase noise," *IEEE J. Select. Areas Commun.*, vol. 23, no. 9, pp. 1748–1757, Sep. 2005.

- [16] F. Bøghagen, P. Orten, and G. E. Øien, “Design of optimal high-rank line-of-sight MIMO channels,” *IEEE Trans. Wireless Commun.*, vol. 6, no. 4, pp. 1420–1425, Apr. 2007.
- [17] P. Ferrand and S. Yang, “Blind precoding in line-of-sight MIMO channels,” in *IEEE SPAWC*, Edinburgh, UK, 2016.
- [18] E. Björnson, J. Hoydis, M. Kountouris, and M. Debbah, “Massive MIMO systems with non-ideal hardware: Energy efficiency, estimation, and capacity limits,” *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7112–7139, Nov. 2014.
- [19] G. J. Foschini and G. Vannucci, “Characterizing filtered light waves corrupted by phase noise,” *IEEE Trans. Inf. Theory*, vol. 34, no. 6, pp. 1437–1448, Jun. 1988.
- [20] Y. Wang, Y. Zhou, D. K. Maslen, and G. S. Chirikjian, “Solving phase-noise Fokker-Planck equations using the motion-group Fourier transform,” *IEEE Trans. Commun.*, vol. 54, no. 5, pp. 868–877, May 2006.
- [21] R. Krishnan, A. G. i Amat, T. Eriksson, and G. Colavolpe, “Constellation optimization in the presence of strong phase noise,” *IEEE Trans. Commun.*, vol. 61, no. 12, pp. 5056–5066, Dec. 2013.

APPENDIX

A. Proof of Proposition 2

Define the vector $\mathbf{V} := (\mathbf{\Lambda}_R \mathbf{H} \mathbf{\Lambda}_T - \mathbf{H}) \mathbf{X}$ so that $R^2 = \|\mathbf{V} + \mathbf{Z}\|^2$. We have:

$$R^2 = \|\mathbf{V}\|^2 + \|\mathbf{Z}\|^2 + \mathbf{Z}^H \mathbf{V} + \mathbf{V}^H \mathbf{Z} \quad (52)$$

$$R^4 = (\|\mathbf{V}\|^2 + \|\mathbf{Z}\|^2)^2 + (\mathbf{Z}^H \mathbf{V} + \mathbf{V}^H \mathbf{Z})^2 + 2(\|\mathbf{V}\|^2 + \|\mathbf{Z}\|^2)^2 (\mathbf{Z}^H \mathbf{V} + \mathbf{V}^H \mathbf{Z}). \quad (53)$$

Using the fact that \mathbf{V} and \mathbf{Z} are independent and \mathbf{Z} has i.i.d. $\mathcal{CN}(0, \gamma^{-1})$ entries, we verify that

$$\mathbb{E}[R^2] = \mathbb{E}[\|\mathbf{V}\|^2] + \mathbb{E}[\|\mathbf{Z}\|^2], \quad (54)$$

$$\mathbb{E}[R^4] = \mathbb{E}[\|\mathbf{V}\|^4] + \mathbb{E}[\|\mathbf{Z}\|^4] + 2\mathbb{E}[\|\mathbf{Z}\|^2] \mathbb{E}[\|\mathbf{V}\|^2] + 2\gamma^{-1} \|\mathbf{V}\|^2, \quad (55)$$

$$\text{Var}(R^2) = \text{Var}(\|\mathbf{V}\|^2) + \text{Var}(\|\mathbf{Z}\|^2) + 2\gamma^{-1} \mathbb{E}[\|\mathbf{V}\|^2]. \quad (56)$$

Since $\mathbb{E}[\|\mathbf{Z}\|^2] = n_r \gamma^{-1}$ and $\text{Var}(\|\mathbf{Z}\|^2) = 2n_r \gamma^{-2}$, it follows that

$$\mathbb{E}[R^2] = \mathbb{E}[\|\mathbf{V}\|^2] + \gamma^{-1} n_r, \quad (57)$$

$$\text{Var}(R^4) = \text{Var}(\|\mathbf{V}\|^2) + 2\gamma^{-2} n_r + 2\gamma^{-1} \mathbb{E}[\|\mathbf{V}\|^2]. \quad (58)$$

We now calculate the moments of $\|\mathbf{V}\|^2$.

Expectation over \mathbf{H} : Let us define $A_{k,l} := X_l(e^{j(\Theta_{t,l} + \Theta_{r,k})} - 1)$ and write, conditional on \mathbf{A} ,

$$\|\mathbf{V}\|^2 = \sum_{k=1}^{n_r} \left| \sum_{l=1}^{n_t} \mathbf{H}_{k,l} A_{k,l} \right|^2 \stackrel{d}{=} \sum_{k=1}^{n_r} E_k, \quad (59)$$

where $E_k \sim \text{Exp}(1/\|\mathbf{A}_k\|^2)$ with $\mathbf{A}_k := [A_{k,l}]_{l=1,\dots,n_t}$, $k = 1, \dots, n_r$; we used the fact that \mathbf{H} has i.i.d. $\mathcal{CN}(0, 1)$ entries. Then we can calculate the first and second moments of $\|\mathbf{V}\|^2$ for a given $\mathbf{A} = [A_{j,k}]_{j,k}$:

$$\mathbb{E}_{\mathbf{H}|\mathbf{A}}[\|\mathbf{V}\|^2] = \sum_{k=1}^{n_r} \mathbb{E}_{\mathbf{H}|\mathbf{A}}[E_k] = \sum_{k=1}^{n_r} \sum_{l=1}^{n_t} |A_{k,l}|^2, \quad (60)$$

$$\mathbb{E}_{\mathbf{H}|\mathbf{A}}[\|\mathbf{V}\|^4] = \mathbb{E}_{\mathbf{H}|\mathbf{A}}\left[\sum_{k=1}^{n_r} E_k^2\right] + \mathbb{E}_{\mathbf{H}|\mathbf{A}}\left[\sum_{\substack{k'=1 \\ k' \neq k}}^{n_t} \sum_{k=1}^{n_t} E_k E_{k'}\right] \quad (61)$$

$$= \sum_{k=1}^{n_r} 2\|\mathbf{A}_k\|^4 + \left[\sum_{\substack{k'=1 \\ k' \neq k}}^{n_t} \sum_{k=1}^{n_t} \|\mathbf{A}_k\|^2 \|\mathbf{A}_{k'}\|^2\right] \quad (62)$$

$$= \sum_{k'=1}^{n_r} \sum_{k=1}^{n_r} (1 + \mathbb{1}\{k = k'\}) \|\mathbf{A}_k\|^2 \|\mathbf{A}_{k'}\|^2. \quad (63)$$

Moments of \mathbf{A} : We recall that $|A_{k,l}|^2 = 2|X_l|^2(1 - \cos(\Theta_{r,k} + \Theta_{t,l}))$. We have that $\mathbb{E}[|X_l|^2] = 1$ and define $\mathbb{E}[|X_l|^4] = \bar{P}^2$, for $l = 1, \dots, n_t$. Using the independence and the identity $\mathbb{E}[\cos(\Theta)] = \exp(-\frac{\text{Var}(\Theta)}{2})$ for zero-mean Gaussian Θ , we obtain

$$\mathbb{E}[|A_{k,l}|^2] = 2\mathbb{E}[|X_l|^2] \mathbb{E}[1 - \cos(\Theta_{t,l} + \Theta_{r,k})] = 2\left(1 - e^{-\frac{\sigma_r^2 + \sigma_t^2}{2}}\right). \quad (64)$$

We now calculate the correlation between the entries of \mathbf{A}

$$\mathbb{E}(|A_{k,l}|^2 |A_{k',l'}|^2) = 4\mathbb{E}(|X_l|^2 |X_{l'}|^2) \rho_{k,k',l,l'} \quad (65)$$

with

$$\rho_{k,k',l,l'} := \mathbb{E}((1 - \cos(\Theta_{t,l} + \Theta_{r,k}))(1 - \cos(\Theta_{t,l'} + \Theta_{r,k'}))) \quad (66)$$

$$= 1 - 2e^{-\frac{\sigma_r^2 + \sigma_t^2}{2}} + e^{-\sigma_r^2 - \sigma_t^2} \cosh(\sigma_r^2(\mathbb{1}\{k = k'\} + \sigma_t^2 \mathbb{1}\{l = l'\})), \quad (67)$$

where to obtain the last equality we use the trigonometric identities and again apply the identity $\mathbb{E}[\cos(\Theta)] = \exp(-\frac{\text{Var}(\Theta)}{2})$; we recall that $\cosh(x) = \frac{e^x + e^{-x}}{2}$.

Moments of $\|\mathbf{V}\|^2$: From (60) and (64), we have the first moment

$$\mathbb{E}[\|\mathbf{V}\|^2] = \sum_{k=1}^{n_r} \sum_{l=1}^{n_t} \mathbb{E}[|A_{k,l}|^2] = n_t n_r 2\left(1 - e^{-\frac{\sigma_r^2 + \sigma_t^2}{2}}\right). \quad (68)$$

For the variance, we apply (63) and (68)

$$\text{Var}(\|\mathbf{V}\|^2) = \mathbb{E}[\|\mathbf{V}\|^4] - (\mathbb{E}[\|\mathbf{V}\|^2])^2 \quad (69)$$

$$= \sum_{k'=1}^{n_r} \sum_{k=1}^{n_r} \sum_{l'=1}^{n_t} \sum_{l=1}^{n_t} \left(\mathbb{E}[|A_{k,l}|^2 |A_{k',l'}|^2] (1 + \mathbb{1}\{k = k'\}) - \mathbb{E}[|A_{k,l}|^2] \mathbb{E}[|A_{k',l'}|^2] \right). \quad (70)$$

Noting that $\mathbb{E}[|A_{k,l}|^2 |A_{k',l'}|^2] = \mathbb{E}[|A_{k,l}|^2] \mathbb{E}[|A_{k',l'}|^2]$ if $k \neq k'$ and $l \neq l'$, we obtain the variance

$$\text{Var}(\|\mathbf{V}\|^2) = 4n_t n_r \left(w_1 + w_2(n_t - 1) + w_3(n_r - 1) \right), \quad (71)$$

where $w_1, w_2, w_3 > 0$ do not depend on n_t, n_r and correspond to the cases ($k = k', l = l'$), ($k = k', l \neq l'$), ($k \neq k', l = l'$), respectively, in the summation (70),

$$w_1 := 2\bar{P}^2(1 - 2e^{-\sigma^2} + e^{-2\sigma^2} \cosh(2\sigma^2)) - (1 - e^{-\sigma^2})^2, \quad (72)$$

$$w_2 := 2(1 - 2e^{-\sigma^2} + e^{-2\sigma^2} \cosh(\sigma_r^2)) - (1 - e^{-\sigma^2})^2, \quad (73)$$

$$w_3 := \bar{P}^2(1 - 2e^{-\sigma^2} + e^{-2\sigma^2} \cosh(\sigma_t^2)) - (1 - e^{-\sigma^2})^2, \quad (74)$$

where we define $\sigma^2 := \frac{\sigma_t^2 + \sigma_r^2}{2}$.

Putting it together: From (54) and (68), we have

$$\mathbb{E}[R^2] = \mathbb{E}[\|\mathbf{V}\|^2] + \gamma^{-1} n_r = n_t n_r 2 \left(1 - e^{-\frac{\sigma_r^2 + \sigma_t^2}{2}} \right) + \gamma^{-1} n_r, \quad (75)$$

which yields the first result. Note that $\mathbb{E}[R^2] \geq c_1 n_t n_r$ for some constant $c_1 > 0$ with respect to (n_t, n_r) . From (56) and (71), we have proven that there exists a constant $c_2 > 0$ such that

$$\text{Var}(R^2) = \text{Var}(\|\mathbf{V}\|^2) + 2\gamma^{-2} n_r + 2\gamma^{-1} \mathbb{E}[\|\mathbf{V}\|^2] \leq c_2 n_r n_t (n_t + n_r). \quad (76)$$

Hence

$$\frac{\text{Var}(R^2)}{(\mathbb{E}[R^2])^2} \leq \frac{c_2}{c_1^2} \left(\frac{1}{n_r} + \frac{1}{n_t} \right) \rightarrow 0, \quad n_t, n_r \rightarrow \infty, \quad (77)$$

and applying Chebychev's inequality yields the second result:

$$\mathbb{P}\left\{ (1 - \eta) \mathbb{E}[R^2] \leq R^2 \leq (1 + \eta) \mathbb{E}[R^2] \right\} \rightarrow 1, \quad n_t, n_r \rightarrow \infty. \quad (78)$$

B. Proof of Proposition 4

From [19, eq.6]⁵, we can derive the moment-generating function (MGF) of (G, Φ)

$$M_{G,\Phi}(\xi, \eta) = e^{\xi} \operatorname{sinhc}^{-\frac{1}{2}}\left(\sqrt{S\xi}\right) \cdot \exp\left[\frac{S\eta^2}{8} \left[\operatorname{cothc}\left(\sqrt{\frac{S\xi}{4}}\right) - \left(\frac{S\xi}{4}\right)^{-1} + \operatorname{tanhc}\left(\sqrt{S\xi}\right) \right]\right], \quad (79)$$

where we define $\operatorname{sinhc}(x) := \frac{\sinh(x)}{x}$, $\operatorname{cothc}(x) := \frac{\coth(x)}{x}$, $\operatorname{tanhc}(x) := \frac{\tanh(x)}{x}$, and recall that $S := \beta T$. It follows that the MGF of G and Φ are

$$M_G(\xi) = M_{G,\Phi}(\xi, 0) = e^{\xi} \operatorname{sinhc}^{-\frac{1}{2}}\left(\sqrt{S\xi}\right), \quad (80)$$

$$M_{\Phi}(\eta) = M_{G,\Phi}(0, \eta) = \exp\left(\frac{1}{6}S\eta^2\right), \quad (81)$$

where we used the fact that, when $x \rightarrow 0$, $\operatorname{sinhc}(x) = 1 + O(x^2)$, $\operatorname{tanhc}(x) = 1 + O(x^2)$, and

$$\operatorname{cothc}(x) - \frac{1}{x^2} = \frac{1}{3} + O(x^2). \quad (82)$$

After finding the first and second derivatives of both MGF (80) and (81) with some elementary manipulations, we obtain the desired variances

$$\operatorname{Var}(G) = \mathbb{E}(G^2) - (\mathbb{E}(G))^2 = M_G''(0) - (M_G'(0))^2 = \frac{S^2}{180}, \quad (83)$$

$$\operatorname{Var}(\Phi) = \mathbb{E}(\Phi^2) - (\mathbb{E}(\Phi))^2 = M_{\Phi}''(0) - (M_{\Phi}'(0))^2 = \frac{S}{3}. \quad (84)$$

Note that we use approximate equality in (50) and (51) since the MGF derived in [19] is indeed approximative with the assumption of small S .

⁵Note that the random variable B in (47) differs with the one in [19, eq.1] in a normalization factor T . The MGF has been scaled accordingly.