The Variational Inference Approach to Joint Data Detection and Phase Noise Estimation in OFDM

Darryl Dexu Lin, *Student Member, IEEE*, and Teng Joon Lim, *Senior Member, IEEE*

***Abstract*—This paper studies the mitigation of phase noise (PHN) in orthogonal frequency-division multiplexing (OFDM) data detection. We present a systematic probabilistic framework that leads to both optimal and near-optimal OFDM detection schemes in the presence of unknown PHN. In contrast to the conventional approach that cancels the common (average) PHN, our aim is to jointly estimate the complete PHN sequence and the data symbol sequence. We derive a family of low-complexity OFDM detectors for this purpose. The theoretical foundation on which these detectors are based is called variational inference, an approximate probabilistic inference technique associated with the minimization of variational free energy. In deriving the proposed schemes, we also point out that the expectation-maximization algorithm is a special case of the variational-inference-based joint estimator. Further complexity reduction is obtained using the conjugate gradient (CG) method, and only a few CG iterations are needed to closely approach the ideal joint estimator output.**

***Index Terms*—Conjugate gradient (CG), expectation maximization (EM), orthogonal frequency-division multiplexing (OFDM), phase noise (PHN), variational inference.**

I. Introduction

ORTHOGONAL frequency-division multiplexing (OFDM) is becoming the technology of choice in fourth generation wireless communication systems. Wireless applications that use OFDM include wireless LAN IEEE 802.11 a/g, fixed broadband wireless access IEEE 802.16a, terrestrial broadcast of digital television (DVB-T), HiperLAN2 in Europe, and HiSWANa in Japan. While OFDM is considered a practical scheme to combat frequency selective channel fading and to increase data rate, many practical challenges still face OFDM system designers [1]. In this paper, we consider the phase noise (PHN) problem that arises from the imperfections of a practical voltage-controlled oscillator (VCO). Improved RF circuit design may conceivably alleviate the problem but cannot eliminate it. Therefore, it is necessary to design digital signal processing techniques to combat residual PHN in the high-performance systems envisioned for the future, which are very sensitive to PHN. Despite the efforts of many researchers working in this area, most results are still based on over simplified models and are suboptimal.

The PHN problem appears similar to the channel estimation problem in that data detection must be accomplished in the presence of unknown channel effects. However, a channel impulse response (CIR) changes slowly with respect to the OFDM symbol rate, and, hence, channel estimates obtained at some time in the past remain valid in the present and can be used for data detection. On the other hand, PHN varies rapidly and estimates obtained in one OFDM symbol are not strongly correlated with the PHN process in another symbol interval, and, hence, PHN cannot be easily mitigated using a training symbol approach.

In [2], the effect of PHN on the system performance was studied and it was found that OFDM is much more sensitive to PHN than a single carrier system. Tomba [3] provides a more detailed treatment on the OFDM error probability in the presence of PHN for different modulation schemes. Various methods to estimate and mitigate the effect of PHN have also been presented in the literature [4]–[6], where PHN is commonly decomposed into two components: the *common PHN*, also known as the *common phase rotation* (CPR), which depends on the average value of the PHN over one OFDM symbol and has the same effect on all subcarriers, and the *random PHN* which induces intercarrier interference (ICI). Common PHN can be mitigated with a few pilot tones, as in [4], for example. Nikitopoulos and Polydoros [5] assume that the common PHN evolves slowly over consecutive OFDM symbols so that previous estimates of common PHN may be used in the processing of the current OFDM symbol. In [6], an MMSE equalization technique was used to suppress the common PHN after modelling the ICI caused by the random PHN as extra additive noise.

These papers represent state-of-the-art PHN mitigation techniques, where common PHN is the target of estimation and cancellation while random PHN is treated as unavoidable noise. However, the removal of the complete PHN sequence (common PHN + random PHN) must lead to much improved performance since then the ICI introduced by random PHN can be suppressed, but it has never been rigorously studied due to the difficulty of jointly estimating both the PHN profile and data symbols. In this paper, we will show that joint data detection and PHN estimation is feasible through a relaxation of the form of the likelihood function containing both the data and PHN. We will not distinguish between the two components of PHN, and instead investigate the general PHN issue, by first developing a probabilistic model, and then deriving efficient

Manuscript received October 7, 2005; revised June 25, 2006. This work was presented in part at the IEEE Wireless Communications and Networking Conference, New Orleans, LA, March 2005. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Luc Vandendorpe.

The authors are with The Edward S. Rogers, Sr., Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4 Canada (e-mail: linde@comm.utoronto.ca; limtj@comm.utoronto.ca).

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.2006.890915

algorithms to solve the problem. A significant paradigm change is made, in that rather than generating "hard" estimates of the data and PHN profile, we derive probability distributions for the data and PHN, which provide measures of the reliability of the estimator outputs, or alternatively our uncertainty over those estimates.

The mathematical foundations of our solutions are the approximate probabilistic inference algorithms. One instance of probabilistic inference, the sum-product algorithm in factor graphs [7], has been actively researched in recent years in connection with the decoding of Turbo codes and low-density parity-check (LDPC) codes, as well as LDPC decoding in the presence of PHN [8]. Recently, probabilistic inference has also been successfully used in image processing to perform scene analysis [9]. Motivated by these pioneering works, we apply approximate probabilistic inference algorithms, namely variational inference and its variants, to alleviate the PHN problem in OFDM. This novel approach leads to surprisingly low-complexity solutions.

The rest of the paper will be organized as follows. Section II discusses the statistical (prior) distribution of the PHN process and presents the signal model. Section III summarizes the conventional PHN cancellation algorithm and derives the optimal OFDM symbol detector. Section IV presents a family of algorithms based on the variational inference framework that jointly detect the data symbols and estimate PHN. Section V provides complexity analysis on the proposed algorithms and derives complexity reduction methods based on the conjugate gradient (CG) algorithm. Section VI presents simulation results that test the proposed algorithms in terms of the bit-error rate (BER) of detected OFDM symbols, and Section VII contains the conclusion.

Notation: Upper and lower case bold face letters indicate matrices and column vectors; $(\cdot)^*$, $(\cdot)^T$, and $(\cdot)^H$ denote conjugation, transpose, and Hermitian transpose, respectively; $\mathbf{1}$ represents the all-one column vector; $\mathbf{0}$ represents the all-zeros column vector or matrix; $\mathrm{diag}(\mathbf{x})$ is a diagonal matrix with the vector $\mathbf{x}$ on its diagonal; $\mathrm{diag}(\mathbf{X})$ is a diagonal matrix with the diagonal elements of square matrix $\mathbf{X}$ on its diagonal; $\mathrm{Re}(\cdot)$ and $\mathrm{Im}(\cdot)$ denote the real and imaginary part of a vector or matrix; $\mathrm{E}(\cdot)$ and $\mathrm{V}(\cdot)$ stand for the expected value and variance of a random variable; $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ represent, respectively, real and circularly symmetric complex Gaussian random vectors with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. In particular, for an N-dimensional circularly symmetric complex Gaussian random vector $\mathbf{x}$

$$\mathcal{CN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\pi^N |\boldsymbol{\Sigma}|} \exp\left\{-(\mathbf{x}-\boldsymbol{\mu})^H \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\}. \quad (1)$$

II. System Description

A. Prior Statistics of PHN

Two different models of PHN are available in the literature [10]. The first one models a free-running oscillator and assumes the PHN process to be a Wiener process that is nonstationary, with its variance growing with time. The second one models an oscillator controlled by a phase-locked loop (PLL) and approximates the PHN process as a zero-mean coloured Gaussian process that is wide sense stationary (WSS) and has finite variance. In this paper, our solution covers both scenarios. For simplicity, we will refer to the first one as *Wiener PHN* and the second one as *Gaussian PHN*, even though both assume Gaussian statistics.

In both cases, denoting the phase noise process at the output of the VCO by $\theta(t)$, the samples of $\theta(t)$ within the mth OFDM symbol, $\boldsymbol{\theta}_m$, has a multivariate Gaussian prior distribution: $p(\boldsymbol{\theta}_m) = \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi})$, where the samples are taken at a rate of N/T samples per second, N is the number of OFDM subcarriers, and T is the period of the OFDM symbol. For this model to be useful, however, the covariance matrix, $\boldsymbol{\Phi}$, must be available. [11] explains how $\boldsymbol{\Phi}$ can be determined from the power spectral density (PSD) of the VCO output. The main results are summarized as follows.

1) Wiener PHN: The Wiener PHN process is modelled as $\theta(t) = \int_0^t \phi(\tau) d\tau$, where $\phi(t)$ is zero-mean white Gaussian noise. The discrete-time samples of $\theta(t)$ form a random-walk process $\theta_n = \theta_{n-1} + \phi_n$, $n = 0, \cdots, N-1$, where $p(\phi_n) = \mathcal{N}(0, \alpha_\phi^2)$. Setting $\theta_{-1} = 0$ due to perfect synchronization at the beginning of the OFDM symbol, the Gaussian-distributed PHN vector $\boldsymbol{\theta} = [\theta_0, \cdots, \theta_{N-1}]^T$ has covariance matrix

$$\boldsymbol{\Phi} = \alpha_\phi^2 \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 2 & \cdots & 2 \\ \vdots & \vdots & \ddots & \\ 1 & 2 & & N \end{bmatrix}. \quad (2)$$

2) Gaussian PHN: The Gaussian PHN process $\theta(t)$ is modelled as a stationary random process with autocorrelation function $R_\theta(\tau)$. It can be shown that the ith row and jth column of the PHN covariance matrix $\boldsymbol{\Phi}$ is

$$\boldsymbol{\Phi}_{i,j} = R_\theta\left(|i-j|\frac{T}{N}\right) \quad (3)$$

where T/N is the sampling period.

In (2) and (3), both α_ϕ and $R_\theta(\tau)$ can be determined from the PSD of the VCO output. In the subsequent derivations, since both types of PHN can be sufficiently characterized by the covariance matrix $\boldsymbol{\Phi}$, we shall not distinguish between the two unless specifically stated.

B. Signal Model

We consider a slow fading frequency-selective channel where the CIR is assumed to remain constant during each packet of transmission which consists of multiple OFDM symbols including the initial preambles for synchronization and channel estimation, as well as the variable-length payload that follows. In practice, both the preambles and payload suffer from damaging effects caused by transmitter and receiver oscillator jitter, which translate into carrier frequency offset (CFO) and PHN.

Since frame synchronization is commonly performed using autocorrelation based metrics [12], [13] which are not sensitive to CFO and PHN disturbances, perfect frame synchronization can be safely assumed. Thus, we may concentrate on one OFDM symbol period. The system block diagram, including the transmitter, channel and receiver, is given in Fig. 1, where the CFO

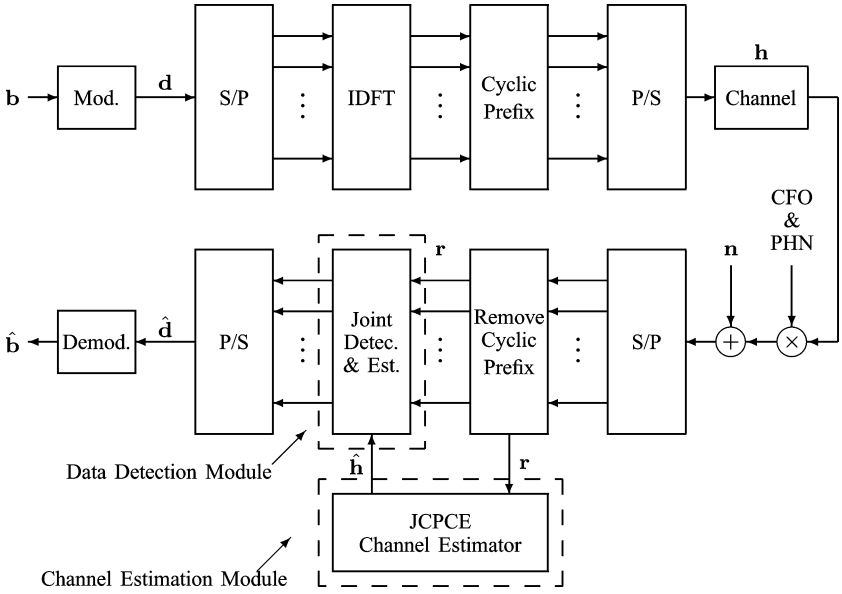

Fig. 1. OFDM transmitter/receiver structure and phase noise channel model. The channel is estimated using the joint CFO/PHN/CIR estimator (JCPCE) [14].

and PHN effects are explicitly shown as multiplicative disturbances to the received signal.

PHN does distort CFO and CIR estimates, however. To resolve this issue, the *channel estimation module* in Fig. 1 has already been studied in [14], where an optimal joint CFO/PHN/CIR estimator (JCPCE) was proposed that performs almost as well as if no CFO or PHN existed. The exact CFO and PHN that distort the channel estimation training symbols are also estimated. It is, therefore, reasonable to assume that the CIR and CFO (which are quasi-static) are known at the data detection stage. However, the same assumption cannot be made about PHN since it is time varying and differs from one OFDM symbol to the next. This is what motivates this work—our goal is to investigate methods for efficient OFDM data detection in the presence of unknown PHN distortion assuming known CIR and removal of CFO (as depicted in the *data detection module* in Fig. 1).

In the subsequent description of the signal model for data detection, we will assume that the CFO is perfectly estimated at the channel estimation stage and removed. The complex baseband received signal of one OFDM data symbol within the payload section sampled at rate N/T can be written as an N point sequence for $n = 0, \cdots, N-1$

$$r_n = \frac{1}{\sqrt{N}} e^{j\theta_n} \sum_{k=0}^{N-1} h_k d_k e^{j2\pi nk/N} + \eta_n \tag{4}$$

where $\{\theta_n\}_{n=0}^{N-1}$ is the discrete-time PHN sequence; $\{h_k\}_{k=0}^{N-1}$ is the channel frequency response at subcarriers 0 to $N-1$; $\{d_k\}_{k=0}^{N-1}$ are the transmitted data symbols belonging to an M-QAM constellation, and $\{\eta_n\}_{n=0}^{N-1}$ is complex white Gaussian noise with variance σ^2 per dimension. (4) may be written in matrix form as

$$\mathbf{r} = \mathbf{P}\mathbf{F}^H\mathbf{H}\mathbf{d} + \mathbf{n} \tag{5}$$

where $\mathbf{F} \in \mathbb{C}^{N\times N}$ is the DFT matrix with the (l,m)th element being $\mathbf{F}_{l,m} = (1/\sqrt{N})e^{-j(2\pi(l-1)(m-1)/N)}$; $\mathbf{d} = [d_0, \cdots, d_{N-1}]^T$ is the data vector; $\mathbf{n} = [\eta_0, \cdots, \eta_{N-1}]^T$ is the noise vector; $\mathbf{P} = \mathrm{diag}([e^{j\theta_0}, \cdots, e^{j\theta_{N-1}}]^T)$ is the PHN matrix; and $\mathbf{H} = \mathrm{diag}(\mathbf{h}) = \mathrm{diag}([h_0, \cdots, h_{N-1}]^T)$ is the channel matrix. Notice that although a full OFDM symbol contains $N_g + N$ time samples, N_g being the length of the cyclic prefix, in this signal model we assume the CIR has a length $L < N_g$ so that the cyclic prefix can be removed and there are only N samples per OFDM symbol.

Note that, in contrast to many papers discussing PHN, we do not assume that the channel is perfectly equalized before PHN estimation and cancellation, which would be difficult, looking at (5).

III. Conventional and Optimal Phase Noise Cancellation

A. Conventional Schemes

In this section, we describe the essence of conventional PHN cancellation schemes [4]–[6]. The discrete Fourier transform (DFT) of the time-domain received signal vector $\mathbf{r} = [r_0, \cdots, r_{N-1}]^T$ expressed in (4) produces a frequency domain sequence $[R_0, \cdots, R_{N-1}]^T$ which can be written as

$$\begin{aligned} R_k &= \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} r_n e^{j2\pi nk/N} \\ &= d_k h_k U_0 + \sum_{l=0, l\neq k}^{N-1} d_l h_l U_{(l-k)_N} + v_k \end{aligned} \tag{6}$$

for $k = 0$ to $N-1$. $[U_0, \cdots, U_{N-1}]^T$ is $1/\sqrt{N}$ times the DFT of the PHN vector $\mathbf{u} = [e^{j\theta_0}, \cdots, e^{j\theta_{N-1}}]^T$, and $\mathbf{v} = [v_0, \cdots, v_{N-1}]^T$ is the DFT of the noise vector $\mathbf{n}$. $(l-k)_N$ stands for $(l-k) \bmod N$. It can be shown that $v_k \sim \mathcal{CN}(0, 2\sigma^2)$, and the ICI term $\sum_{l=0, l\neq k}^{N-1} d_l h_l U_{(l-k)_N}$ is approximated as zero-mean complex Gaussian noise with

variance $2\sigma_{ICI}^2 = E_s \sum_{l=0,l\neq k}^{N-1} |h_l|^2 \mathrm{E}[|U_{(l-k)_N}|^2]$, where $E_s = 2\rho^2$ and ρ^2 is the symbol energy per dimension per subcarrier. Assuming pilot symbols transmitted on carriers with index set $\mathcal{S}_p$, the least-squares (LS) estimate of the common PHN term U_0 is

$$\hat{U}_0 = \frac{\sum_{k\in\mathcal{S}_p} R_k d_k^* h_k^*}{\sum_{k\in\mathcal{S}_p} |d_k h_k|^2}. \tag{7}$$

Given the estimated common PHN, the general phase noise suppression (GPNS) scheme [6] estimates d_k as

$$\hat{d}_k = \frac{E_s \hat{U}_0^* h_k^* R_k}{E_s |\hat{U}_0 h_k|^2 + 2\sigma_{\mathrm{tot}}^2} \tag{8}$$

where $2\sigma_{\mathrm{tot}}^2 = 2\sigma_{ICI}^2 + 2\sigma^2$ is the effective additive noise variance. In [6], pilots are used to estimate both U_0 and $2\sigma_{\mathrm{tot}}^2$. GPNS will be simulated in Section VI as the "conventional scheme" to be compared with our proposed schemes, where no pilots are necessary. Therefore, for comparison purposes, we will simply assume U_0 is known perfectly when simulating GPNS. Furthermore, with channel knowledge and known PHN statistics, $2\sigma_{\mathrm{tot}}^2$ can be calculated exactly in advance. With these assumptions, pilots are not needed in the simulations, even for GPNS.

B. Maximum Likelihood Detector

Here, we give an original derivation of the optimal detector (also called *exact inference* in probabilistic inference literature) given the prior distribution of PHN $\boldsymbol{\theta}$ and show that it has complexity exponential in N.

From (5), the received signal can be expressed alternatively as

$$\begin{aligned} \mathbf{r} &= \mathrm{diag}(\mathbf{F}^H \mathbf{H} \mathbf{d})\mathbf{u} + \mathbf{n} \\ &\approx \mathrm{diag}(\mathbf{F}^H \mathbf{H} \mathbf{d})(\mathbf{1} + j\boldsymbol{\theta}) + \mathbf{n} \end{aligned} \tag{9}$$

where $\mathbf{u} = e^{j\boldsymbol{\theta}} = [e^{j\theta_0}, \cdots, e^{j\theta_{N-1}}]^T$. The approximation is tight since $\boldsymbol{\theta}$ is small. We, thus, have the following prior and conditional pdfs:

$$\begin{aligned} p(\boldsymbol{\theta}) &= \mathcal{N}(\mathbf{0}, \boldsymbol{\Phi}) \\ p(\mathbf{r}|\mathbf{d}, \boldsymbol{\theta}) &= \mathcal{CN}\left(\mathrm{diag}(\mathbf{F}^H \mathbf{H} \mathbf{d})(\mathbf{1} + j\boldsymbol{\theta}), 2\sigma^2 \mathbf{I}\right). \end{aligned} \tag{10}$$

The ML estimate of $\mathbf{d}$ is derived using classical estimation theory by treating $\boldsymbol{\theta}$ as a nuisance parameter and "integrating it out" [15] to obtain $p(\mathbf{r}|\mathbf{d})$. As derived in Appendix I, we find that

$$\begin{aligned} p(\mathbf{r}|\mathbf{d}) = \mathcal{CN}\big(&\mathbf{F}^H \mathbf{H} \mathbf{d}, \mathrm{diag}(\mathbf{F}^H \mathbf{H} \mathbf{d}) \\ &\times \boldsymbol{\Phi} \mathrm{diag}(\mathbf{F}^H \mathbf{H} \mathbf{d})^H + 2\sigma^2 \mathbf{I}\big). \end{aligned} \tag{11}$$

Unfortunately, the maximizer of (11) does not have a closed form and the optimal $\mathbf{d}$ can only be found if each symbol hypothesis is tested, resulting in complexity of $\mathcal{O}(M^N)$, where M is the constellation size.

IV. Joint Estimation via Variational Inference

Our basic problem is the estimation of $\mathbf{d}$, whose optimal solution is the maximization of $p(\mathbf{d}|\mathbf{r}) = \int p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{r}) d\boldsymbol{\theta}$. This is hard to do because $\mathbf{d}$ is drawn from a discrete sample space, and, hence, the problem is NP-complete. The variational inference approach first relaxes the problem constraints by allowing $\mathbf{d}$ to be continuous; then it approximates $p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{r})$ with a function $Q(\mathbf{d}, \boldsymbol{\theta})$ that has convenient properties, such as $Q(\mathbf{d}, \boldsymbol{\theta}) = Q_d(\mathbf{d}) Q_\theta(\boldsymbol{\theta})$. This last assumption is equivalent to assuming that $\mathbf{d}$ and $\boldsymbol{\theta}$ are independent conditioned on $\mathbf{r}$, and immediately leads to the result that the vector $\hat{\mathbf{d}}$ that maximizes $Q(\mathbf{d}, \boldsymbol{\theta})$ over $\mathbf{d}$ and $\boldsymbol{\theta}$ also maximizes $\int Q(\mathbf{d}, \boldsymbol{\theta}) d\boldsymbol{\theta}$. In other words, the joint estimation of $\mathbf{d}$ and $\boldsymbol{\theta}$ directly yields the optimal estimate of $\mathbf{d}$. Finally, $Q_d(\mathbf{d})$ and $Q_\theta(\boldsymbol{\theta})$ must be chosen so that they can be easily manipulated. Details will now be provided.

A. Variational Inference

First applied in statistical physics in the so-called *Ising model* that describes magnetic spin glasses [16], [17], variational inference is a versatile technique for approximating complex distributions. Instead of going into abstract details about its physics applications, we will merely concentrate on the problem at hand and offer a simple interpretation.

Consider the optimization of $p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{r})$ over $\mathbf{d}$ and $\boldsymbol{\theta}$. The heart of the variational technique is that it looks for a parameterized Q-distribution, $Q(\mathbf{d}, \boldsymbol{\theta})$, which closely resembles $p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{r})$, and then finds $\mathbf{d}$ and $\boldsymbol{\theta}$ that maximize $Q(\mathbf{d}, \boldsymbol{\theta})$. The versatility and simplicity of the variational technique lies in the fact that when $Q(\mathbf{d}, \boldsymbol{\theta})$ is properly selected (e.g., as a Gaussian distribution), its maximizers can be easily deduced. It can be shown that the problem has been transformed from the maximization of $p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{r})$ itself to that of its lower-bound [9], yielding enormous computational savings.

To derive the variational inference algorithm, we first introduce a concept called *variational free energy* (also called *Gibbs free energy*) [18] in the context of the PHN problem

$$\mathcal{F}(Q, p) = \int_{\mathbf{d}, \boldsymbol{\theta}} Q(\mathbf{d}, \boldsymbol{\theta}) \log \frac{Q(\mathbf{d}, \boldsymbol{\theta})}{p(\mathbf{d}, \boldsymbol{\theta}, \boldsymbol{r})} d\mathbf{d} d\boldsymbol{\theta}. \tag{12}$$

Here, we use $p(\mathbf{d}, \boldsymbol{\theta}, \mathbf{r})$ instead of $p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{r})$ because they are proportional and, hence, equivalent in the free energy formulation. This expression is exactly the Kullback–Leibler (KL) divergence [19] between $Q(\mathbf{d}, \boldsymbol{\theta})$ and $p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{r})$, denoted $D(Q(\mathbf{d}, \boldsymbol{\theta}) || p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{r}))$, to within an additive constant. By minimizing $\mathcal{F}(Q, p)$ over the parameters of $Q(\mathbf{d}, \boldsymbol{\theta})$, we obtain a Q-distribution $Q(\mathbf{d}, \boldsymbol{\theta})$ that most "resembles" $p(\mathbf{d}, \boldsymbol{\theta}|\mathbf{r})$ in terms of KL divergence.

In cases where there are multiple arguments in the Q-function, an additional simplification can be made by factorizing $Q(\mathbf{d}, \boldsymbol{\theta})$ into a product form (also known as *mean-field* approximation), i.e., $Q(\mathbf{d}, \boldsymbol{\theta}) = Q_d(\mathbf{d}) Q_\theta(\boldsymbol{\theta})$. The mean-field approximation is also important for justifying the use of the $\mathbf{d}$ component of the maximizer of $Q(\mathbf{d}, \boldsymbol{\theta})$ as the optimal estimate of $\mathbf{d}$; without it, we should find $\int Q(\mathbf{d}, \boldsymbol{\theta}) d\boldsymbol{\theta}$, and then maximize the result, which may turn out to be infeasible.

For PHN estimation, we assume that (after dropping the subscripts of the Q-functions for simplicity of notation)

$$\begin{aligned} Q(\mathbf{d}) &= \mathcal{CN}(\mathbf{m}_d, \mathbf{S}_d) \\ Q(\boldsymbol{\theta}) &= \mathcal{N}(\mathbf{m}_\theta, \mathbf{S}_\theta). \end{aligned} \tag{13}$$

TABLE I
UPDATING EQUATIONS FOR VARIATIONAL INFERENCE ALGORITHM

Initialization:	Choose initial values for $\mathbf{S}_d^{(0)}$ and $\mathbf{m}_d^{(0)}$ (e.g. $\mathbf{S}_d^{(0)}=\mathbf{0}$ and $\mathbf{m}_d^{(0)}=\hat{\mathbf{d}}_{init}$).
Iterations:	*For* $t=1$ *to* n
Update for $Q(\boldsymbol{\theta})$	$\mathbf{S}_\theta^{(t)}=\sigma^2\left[\sigma^2\boldsymbol{\Phi}^{-1}+\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{S}_d^{(t-1)}\mathbf{H}^H\mathbf{F})+\mathbf{X}_m^{H(t-1)}\mathbf{X}_m^{(t-1)}\right]^{-1}$ $\mathbf{m}_\theta^{(t)}=\sigma^{-2}\mathbf{S}_\theta^{(t)}\cdot\mathrm{Im}\left[\mathbf{X}_m^{H(t-1)}(\mathbf{r}-\mathbf{F}^H\mathbf{H}\mathbf{m}_d^{(t-1)})\right]$
Update for $Q(\mathbf{d})$	$\mathbf{S}_d^{(t)}=2\sigma^2\left[\sigma^2\rho^{-2}\mathbf{I}+\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta^{(t)})\mathbf{F}^H\mathbf{H}+\mathbf{H}^H\mathbf{H}\right]^{-1}$ $\mathbf{m}_d^{(t)}=(2\sigma^2)^{-1}\mathbf{S}_d^{(t)}\mathbf{H}^H\mathbf{F}\mathbf{P}_m^{H(t)}\mathbf{r}$
	End

It is worth noting that the posteriors of $\mathbf{d}$ and $\boldsymbol{\theta}$ are now parameterized by their means and variances (namely, $\mathbf{m}_d$, $\mathbf{m}_\theta$, $\mathbf{S}_d$, and $\mathbf{S}_\theta$), which then become the targets of optimization instead of the Q-functions themselves.

The *complete likelihood function* $p(\mathbf{d},\boldsymbol{\theta},\mathbf{r})$ in (12) may be written as $p(\mathbf{d},\boldsymbol{\theta},\mathbf{r})=p(\mathbf{r}|\mathbf{d},\boldsymbol{\theta})p(\boldsymbol{\theta})p(\mathbf{d})$, where $p(\mathbf{r}|\mathbf{d},\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ are given in (10). In addition, we let $p(\mathbf{d})=\mathcal{CN}(\mathbf{0},2\rho^2\mathbf{I})$. Note that instead of defining a discrete distribution over the signal constellation, we have made a Gaussian approximation, which leads to a linear detector. Substituting the Q-functions from (13) into (12), we have the closed-form expression of $\mathcal{F}(Q,p)$, now expressed as a function of the parameters of the Q-functions, as derived in Appendix II

$$\begin{aligned}\mathcal{F}(\mathbf{m}_d,\mathbf{S}_d,\mathbf{m}_\theta,\mathbf{S}_\theta)&=\frac{1}{2\rho^2}\left[\mathrm{tr}(\mathbf{S}_d)+\mathbf{m}_d^H\mathbf{m}_d\right]+\frac{1}{2}\mathrm{tr}(\boldsymbol{\Phi}^{-1}\mathbf{S}_\theta)\\&\quad+\frac{1}{2}\mathbf{m}_\theta^T\boldsymbol{\Phi}^{-1}\mathbf{m}_\theta-\frac{1}{2}\log|\mathbf{S}_\theta|-\log|\mathbf{S}_d|\\&\quad+\frac{1}{2\sigma^2}\left\{\mathrm{tr}\left[(j\mathbf{M}_\theta+\mathbf{I})\mathbf{F}^H\mathbf{H}\mathbf{S}_d\mathbf{H}^H\mathbf{F}(j\mathbf{M}_\theta+\mathbf{I})^H\right]\right.\\&\qquad+\mathrm{tr}\left[\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\mathbf{S}_d\right]\\&\qquad+\mathbf{m}_d^H\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\mathbf{m}_d\\&\qquad+\left[(j\mathbf{M}_\theta+\mathbf{I})\mathbf{F}^H\mathbf{H}\mathbf{m}_d-\mathbf{r}\right]^H\\&\qquad\left.\times\left[(j\mathbf{M}_\theta+\mathbf{I})\mathbf{F}^H\mathbf{H}\mathbf{m}_d-\mathbf{r}\right]\right\}.\end{aligned}\tag{14}$$

Obviously, the optimal parameters are hard to obtain analytically in one step. The usual practice is to update each one of them in turn, while holding the others constant, a simple technique termed *coordinate descent* in the optimization literature [20]. The algorithm is guaranteed to converge to a local minimum of the free energy expression [21]. Taking the partial derivative of $\mathcal{F}(\mathbf{m}_d,\mathbf{S}_d,\mathbf{m}_\theta,\mathbf{S}_\theta)$ w.r.t. each parameter and equating to zero, we obtain the following set of equations, with detailed derivations in Appendix III:

$$\mathbf{S}_\theta=\sigma^2\left[\sigma^2\boldsymbol{\Phi}^{-1}+\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{S}_d\mathbf{H}^H\mathbf{F})+\mathbf{X}_m\mathbf{X}_m^H\right]^{-1}\tag{15}$$

$$\mathbf{m}_\theta=\sigma^{-2}\mathbf{S}_\theta\cdot\mathrm{Im}\left[\mathbf{X}_m^H(\mathbf{r}-\mathbf{F}^H\mathbf{H}\mathbf{m}_d)\right]\tag{16}$$

$$\mathbf{S}_d=2\sigma^2\left[\sigma^2\rho^{-2}\mathbf{I}+\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}+\mathbf{H}^H\mathbf{H}\right]^{-1}\tag{17}$$

$$\mathbf{m}_d=(2\sigma^2)^{-1}\mathbf{S}_d\mathbf{H}^H\mathbf{F}\mathbf{P}_m^H\mathbf{r}\tag{18}$$

where $\mathbf{X}_m=\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{m}_d)$ and $\mathbf{P}_m=\mathrm{diag}(e^{j\mathbf{m}_\theta})$.

We summarize the parameter updating procedure in Table I. In each iteration, the parameters are updated in turn to generate new posterior estimates of $\mathbf{d}$ and $\boldsymbol{\theta}$ that decrease $\mathcal{F}(\mathbf{m}_d,\mathbf{S}_d,\mathbf{m}_\theta,\mathbf{S}_\theta)$ monotonically. This particular update order was chosen based on the dependence of $\mathbf{m}_\theta$ and $\mathbf{m}_d$ on $\mathbf{S}_\theta$ and $\mathbf{S}_d$. For initialization, a tentative data decision is made assuming zero PHN and fed back as $\mathbf{m}_d^{(0)}$. $\mathbf{S}_d^{(0)}$ is set to the all-zero matrix. At the last iteration, approximate posterior distributions of $\mathbf{d}$ and $\boldsymbol{\theta}$ are extracted. Since $Q(\mathbf{d})$ is assumed to be Gaussian, it is maximized at $\mathbf{d}=\mathbf{m}_d$, and so hard decisions can be obtained from slicing $\mathbf{m}_d$, or $Q(\mathbf{d})$ can be used as a soft estimate of $\mathbf{d}$ for further error control decoding.

The difference between the variational inference and the ML approach in Section III-B lies in the fact that in variational inference the posterior density $p(\mathbf{d},\boldsymbol{\theta}|\mathbf{r})$ is "forced" to be a product of Gaussian densities $Q(\mathbf{d})$ and $Q(\boldsymbol{\theta})$, making final symbol decisions easy to make since no integration over $\boldsymbol{\theta}$ is necessary, and the mean of $Q(\mathbf{d})$ is its maximizer. These approximations mean that the algorithm in general does not converge to the global maximum of the complete likelihood function, but it has been found to work very well in many applications.

In the sequel, we investigate a few important variants of the variational inference algorithm. Although each is given a specific name, they only differ from the original version in their use of different Q functions.

B. Iterative Conditional Mode (ICM)

Variational inference has made a very complex problem computationally tractable, but a further simplification is possible by assuming the posteriors $Q(\mathbf{d})$ and $Q(\boldsymbol{\theta})$ to be delta functions instead of Gaussian.

In this case, the Q-functions are $\delta(\mathbf{d},\hat{\mathbf{d}})$ and $\delta(\boldsymbol{\theta},\hat{\boldsymbol{\theta}})$, respectively. The notation $\delta(\mathbf{a},\hat{\mathbf{a}})$ denotes a vector Dirac delta function with the following properties: $\int\delta(\mathbf{a},\hat{\mathbf{a}})f(\mathbf{a})d\mathbf{a}=f(\hat{\mathbf{a}})$, and $\int\delta(\mathbf{a},\hat{\mathbf{a}})d\mathbf{a}=1$. The minimization of variational free energy over the parameters $\hat{\mathbf{d}}$ and $\hat{\boldsymbol{\theta}}$ is equivalent to minimizing $\mathcal{L}(\hat{\mathbf{d}},\hat{\boldsymbol{\theta}})=-\log p(\mathbf{r},\hat{\mathbf{d}},\hat{\boldsymbol{\theta}})$ over $\hat{\mathbf{d}}\in\mathbb{C}^N$ and $\hat{\boldsymbol{\theta}}\in\mathbb{R}^N$. An algorithm based on coordinate descent will iteratively perform optimal point estimation for one of the two unknowns while holding the other fixed, hence the name iterative conditional mode.

Since $p(\mathbf{r},\hat{\mathbf{d}},\hat{\boldsymbol{\theta}})=p(\mathbf{r}|\hat{\mathbf{d}},\hat{\boldsymbol{\theta}})p(\hat{\mathbf{d}})p(\hat{\boldsymbol{\theta}})$, $\mathcal{L}(\hat{\mathbf{d}},\hat{\boldsymbol{\theta}})$ is evaluated to be

$$\begin{aligned}\mathcal{L}(\hat{\mathbf{d}},\hat{\boldsymbol{\theta}})&=\frac{1}{2\rho^2}(\hat{\mathbf{d}}^H\hat{\mathbf{d}})+\frac{1}{2}\hat{\boldsymbol{\theta}}^T\boldsymbol{\Phi}^{-1}\hat{\boldsymbol{\theta}}\\&\quad+\frac{1}{2\sigma^2}(\mathbf{r}-\hat{\mathbf{P}}\mathbf{F}^H\mathbf{H}\hat{\mathbf{d}})^H(\mathbf{r}-\hat{\mathbf{P}}\mathbf{F}^H\mathbf{H}\hat{\mathbf{d}}).\end{aligned}\tag{19}$$

TABLE II
UPDATING EQUATIONS FOR ICM ALGORITHM

Initialization:	Choose initial values for $\hat{\mathbf{d}}^{(0)}$ (e.g. $\hat{\mathbf{d}}^{(0)} = \hat{\mathbf{d}}_{init}$).
Iterations:	*For* $t = 1$ *to* n
Update for $Q(\boldsymbol{\theta})$	$\hat{\boldsymbol{\theta}}^{(t)} = \left[\sigma^2\boldsymbol{\Phi}^{-1} + \hat{\mathbf{X}}^{H(t-1)}\hat{\mathbf{X}}^{(t-1)}\right]^{-1} \cdot \mathrm{Im}\left[\hat{\mathbf{X}}^{H(t-1)}(\mathbf{r} - \mathbf{F}^H\mathbf{H}\hat{\mathbf{d}}^{(t-1)})\right]$
Update for $Q(\mathbf{d})$	$\hat{\mathbf{d}}^{(t)} = \left[\sigma^2\rho^{-2}\mathbf{I} + \mathbf{H}^H\mathbf{H}\right]^{-1}\mathbf{H}^H\mathbf{F}\hat{\mathbf{P}}^{H(t)}\mathbf{r}$
	End

As derived in Appendix IV, solving $\partial\mathcal{L}/\partial\hat{\mathbf{d}}^* = \mathbf{0}$ and $\partial\mathcal{L}/\partial\hat{\boldsymbol{\theta}} = \mathbf{0}$ leads to

$$\hat{\mathbf{d}} = (\sigma^2\rho^{-2}\mathbf{I} + \mathbf{H}^H\mathbf{H})^{-1}\mathbf{H}^H\mathbf{F}\hat{\mathbf{P}}^H\mathbf{r} \quad (20)$$

$$\hat{\boldsymbol{\theta}} = (\sigma^2\boldsymbol{\Phi}^{-1} + \hat{\mathbf{X}}^H\hat{\mathbf{X}})^{-1} \cdot \mathrm{Im}\left[\hat{\mathbf{X}}^H(\mathbf{r} - \mathbf{F}^H\mathbf{H}\hat{\mathbf{d}})\right] \quad (21)$$

where $\hat{\mathbf{X}} = \mathrm{diag}(\mathbf{F}^H\mathbf{H}\hat{\mathbf{d}})$ and $\hat{\mathbf{P}} = \mathrm{diag}(e^{j\hat{\boldsymbol{\theta}}})$.

The ICM algorithm is shown in Table II. The saving in ICM is that we no longer require the covariance matrices $\mathbf{S}_d$ and $\mathbf{S}_\theta$ of the posterior distribution. The drawback is that point estimates do not model the uncertainties at each iteration, thus ICM, in general, produces inferior results compared to variational inference. It should be noted, however, though ICM resembles the heuristic decision-directed approach where the data symbols and PHN are detected/estimated iteratively until decisions are made for both unknowns, it is, in nature, rather different, as no symbol hard decisions (constrained to the symbol constellation) are made during the iterations.

C. Expectation-Maximization (EM) Algorithm

In the algorithms of Tables I and II, the Q-functions were chosen to be Gaussian and delta functions, respectively. We now show that these choices can be seen as the two extremes in a spectrum of choices that can be related to the EM algorithm [22]. The EM algorithm is used to estimate a vector of parameters, say $\boldsymbol{\phi}$, from observations $\mathbf{y}$ that are termed incomplete data, with the help of some auxiliary or hidden variables, say $\mathbf{x}$. The algorithm iteratively carries out two operations: the E-step and the M-step. The tth iteration effectively computes a probability density $p(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}^{(t-1)})$, where $\boldsymbol{\phi}^{(t-1)}$ is the estimate of $\boldsymbol{\phi}$ in the previous iteration, and then maximizes

$$U\left(\boldsymbol{\phi}, \boldsymbol{\phi}^{(t-1)}\right) = \int p\left(\mathbf{x}|\mathbf{y}, \boldsymbol{\phi}^{(t-1)}\right)\log p(\mathbf{x}, \mathbf{y}|\boldsymbol{\phi})d\mathbf{x} \quad (22)$$

over $\boldsymbol{\phi}$, yielding $\boldsymbol{\phi}^{(t)}$.

In any given problem, the EM algorithm requires us to separate the unknown parameters from the unknown hidden variables. In other words, all unknowns are grouped into two classes. In each iteration, hard estimates are obtained for the parameters, while soft estimates in the form of probability densities are obtained for the hidden variables. The link between EM and variational inference is made in [21], where it was shown that the EM algorithm is equivalent to jointly estimating the hidden variables and parameters by minimizing a single free-energy expression over a postulated distribution for the hidden variables, and over the parameters.

This can be generalized even further: Suppose we are simply faced with a problem with multiple unknowns. There are some unknowns which we want hard estimates of, and the rest we want soft estimates of. Using delta functions for the postulated distributions of the "hard" unknowns, and exact (or postulated) distributions for those of the "soft" unknowns in the variational inference algorithm will lead to the EM algorithm with the hard unknowns as parameters, and the soft unknowns as hidden variables.

With this perspective, the algorithm in Table I can be seen as an EM algorithm with only hidden variables of Gaussian postulated distributions, and the ICM algorithm in Table II is an EM algorithm with only parameters. In between these extremes, we can set one unknown to be a variable and the other a parameter by simply adjusting the corresponding Q-functions. Specifically, we have two options.

- Option I: $Q(\mathbf{d}) = \mathcal{CN}(\mathbf{m}_d, \mathbf{S}_d)$ and $Q(\boldsymbol{\theta}) = \delta(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$. The free energy expression becomes

$$\begin{aligned}\mathcal{F}(\mathbf{m}_d, \mathbf{S}_d, \hat{\boldsymbol{\theta}}) &= \frac{1}{2\rho^2}\left[\mathrm{tr}(\mathbf{S}_d) + \mathbf{m}_d^H\mathbf{m}_d\right] + \frac{1}{2}\hat{\boldsymbol{\theta}}^T\boldsymbol{\Phi}^{-1}\hat{\boldsymbol{\theta}} - \log|\mathbf{S}_d| \\ &\quad + \frac{1}{2\sigma^2}\Big\{\mathrm{tr}[\hat{\mathbf{P}}\mathbf{F}^H\mathbf{H}\mathbf{S}_d\mathbf{H}^H\mathbf{F}\hat{\mathbf{P}}^H] \\ &\qquad + (\hat{\mathbf{P}}\mathbf{F}^H\mathbf{H}\mathbf{m}_d - \boldsymbol{r})^H(\hat{\mathbf{P}}\mathbf{F}^H\mathbf{H}\mathbf{m}_d - \boldsymbol{r})\Big\} \quad (23)\end{aligned}$$

  where $\hat{\mathbf{P}} = \mathrm{diag}(e^{j\hat{\boldsymbol{\theta}}})$. Minimizing this free energy yields the EM I algorithm in Table III.
- Option II: $Q(\mathbf{d}) = \delta(\mathbf{d}, \hat{\mathbf{d}})$ and $Q(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{m}_\theta, \mathbf{S}_\theta)$. The free energy expression becomes

$$\begin{aligned}\mathcal{F}(\hat{\mathbf{d}}, \mathbf{m}_\theta, \mathbf{S}_\theta) &= \frac{1}{2\rho^2}\hat{\mathbf{d}}^H\hat{\mathbf{d}} + \frac{1}{2}\mathrm{tr}(\boldsymbol{\Phi}^{-1}\mathbf{S}_\theta) + \frac{1}{2}\mathbf{m}_\theta^T\boldsymbol{\Phi}^{-1}\mathbf{m}_\theta - \frac{1}{2}\log|\mathbf{S}_\theta| \\ &\quad + \frac{1}{2\sigma^2}\Big\{\hat{\mathbf{d}}^H\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\hat{\mathbf{d}} \\ &\qquad + [\mathbf{P}_m\mathbf{F}^H\mathbf{H}\hat{\mathbf{d}} - \mathbf{r}]^H[\mathbf{P}_m\mathbf{F}^H\mathbf{H}\hat{\mathbf{d}} - \mathbf{r}]\Big\} \quad (24)\end{aligned}$$

  where $\mathbf{P}_m = \mathrm{diag}(e^{j\mathbf{m}_\theta})$. Minimizing this free energy yields the EM II algorithm in Table IV.

We omit the formal derivations for EM I and EM II since they are straightforward simplifications to the original variational inference derivation. The fact that the above algorithms are equivalent to EM may be verified through the conventional EM formulation by setting the hidden variables and parameters appropriately. It should be noted that in the EM algorithm

TABLE III
UPDATING EQUATIONS FOR EM I ALGORITHM

Initialization:	Choose initial values for $\mathbf{S}_d^{(0)}$ and $\mathbf{m}_d^{(0)}$ (e.g. $\mathbf{S}_d^{(0)} = \mathbf{0}$ and $\mathbf{m}_d^{(0)} = \hat{\mathbf{d}}_{init}$).
Iterations:	*For* $t = 1$ *to* n
Update for $Q(\boldsymbol{\theta})$ *M Step*	$\hat{\boldsymbol{\theta}}^{(t)} = \left[\sigma^2\boldsymbol{\Phi}^{-1} + \mathbf{X}_m^{H(t-1)}\mathbf{X}_m^{(t-1)}\right]^{-1} \cdot \mathrm{Im}\left[\mathbf{X}_m^{H(t-1)}(\mathbf{r} - \mathbf{F}^H\mathbf{H}\mathbf{X}_m^{(t-1)})\right]$
Update for $Q(\mathbf{d})$ *E Step*	$\mathbf{S}_d^{(t)} = 2\sigma^2\left[\sigma^2\rho^{-2}\mathbf{I} + \mathbf{H}^H\mathbf{H}\right]^{-1}$ $\mathbf{m}_d^{(t)} = (2\sigma^2)^{-1}\mathbf{S}_d^{(t)}\mathbf{H}^H\mathbf{F}\hat{\mathbf{P}}^{H(t)}\mathbf{r}$
	End

TABLE IV
UPDATING EQUATIONS FOR EM II ALGORITHM

Initialization:	Choose initial values for $\hat{\mathbf{d}}^{(0)}$ (e.g. $\hat{\mathbf{d}}^{(0)} = \hat{\mathbf{d}}_{init}$).
Iterations:	*For* $t = 1$ *to* n
Update for $Q(\boldsymbol{\theta})$ *E Step*	$\mathbf{S}_\theta^{(t)} = \sigma^2\left[\sigma^2\boldsymbol{\Phi}^{-1} + \hat{\mathbf{X}}^{H(t-1)}\hat{\mathbf{X}}^{(t-1)}\right]^{-1}$ $\mathbf{m}_\theta^{(t)} = \sigma^{-2}\mathbf{S}_\theta^{(t)} \cdot \mathrm{Im}\left[\hat{\mathbf{X}}^{H(t-1)}(\mathbf{r} - \mathbf{F}^H\mathbf{H}\hat{\mathbf{d}}^{(t-1)})\right]$
Update for $Q(\mathbf{d})$ *M Step*	$\hat{\mathbf{d}}^{(t)} = \left[\sigma^2\rho^{-2}\mathbf{I} + \mathbf{H}^H\mathbf{H}\right]^{-1}\mathbf{H}^H\mathbf{F}\mathbf{P}_m^{H(t)}\mathbf{r}$
	End

TABLE V
COMPARISON OF VARIANTS OF VARIATIONAL INFERENCE ALGORITHM

Inference Schemes	*Variational*	*EM I*	*EM II*	*ICM*
Hidden Variable	$\mathbf{d}$, $\boldsymbol{\theta}$	$\mathbf{d}$	$\boldsymbol{\theta}$	–
Parameter	–	$\boldsymbol{\theta}$	$\mathbf{d}$	$\mathbf{d}$, $\boldsymbol{\theta}$
$Q(\mathbf{d})$	$\mathcal{N}(\mathbf{m}_d, \mathbf{S}_d)$	$\mathcal{N}(\mathbf{m}_d, \mathbf{S}_d)$	$\delta(\mathbf{d}, \hat{\mathbf{d}})$	$\delta(\mathbf{d}, \hat{\mathbf{d}})$
$Q(\boldsymbol{\theta})$	$\mathcal{N}(\mathbf{m}_\theta, \mathbf{S}_\theta)$	$\delta(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$	$\mathcal{N}(\mathbf{m}_\theta, \mathbf{S}_\theta)$	$\delta(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}})$
Objective Function	$\mathcal{F}(\mathbf{m}_d, \mathbf{S}_d, \mathbf{m}_\theta, \mathbf{S}_\theta)$	$p(\boldsymbol{\theta}\|\mathbf{r})$	$p(\mathbf{d}\|\mathbf{r})$	$p(\boldsymbol{\theta}, \mathbf{d}\|\mathbf{r})$

the posterior distribution of the hidden variables is not necessarily Gaussian.[1] However, our Gaussian parameterization for the Q-functions is correct here because $p(\mathbf{d}|\mathbf{r}, \hat{\boldsymbol{\theta}})$ and $p(\boldsymbol{\theta}|\mathbf{r}, \hat{\mathbf{d}})$ are Gaussian distributions.

Following the conventional convergence analysis of the EM algorithm, it can be shown that EM I monotonically increases $p(\boldsymbol{\theta}|\mathbf{r})$ with each iteration, while EM II monotonically increases $p(\mathbf{d}|\mathbf{r})$. However, the convergence point may not be the global optima of $p(\boldsymbol{\theta}|\mathbf{r})$ or $p(\mathbf{d}|\mathbf{r})$, due to their local maxima.

D. Summary of Variants of Variational Inference

Through the preceding investigation, we have proposed not one, but a spectrum of algorithms, for the joint detection/estimation problem at hand, all under the unified framework of variational inference and all rigorously derived via free energy minimization. Among these, we have covered the EM algorithm as a special case. This framework admitted new insights into the problem and enabled us to look beyond conventional signal processing methods. Table V provides a summary of the above-mentioned variants of the variational inference algorithm, in which the term "objective function" stands for the function that the corresponding algorithm guarantees to improve after each iteration.

[1]If the true distribution is non-Gaussian, but we assume it to be Gaussian, or some other more convenient distribution, we have a *variational EM* algorithm [9].

V. COMPLEXITY ANALYSIS AND REDUCTION

We consider ICM to be the most promising candidate for practical application, since it has the simplest form while retaining almost exactly the same performance as others. However, applying it directly may still substantially increase the complexity of a practical OFDM receiver. Further complexity reduction has to be devised to avoid a full $N \times N$ matrix inversion, which requires a complexity order of $\mathcal{O}(N^3)$.

Observing the expression for $\hat{\mathbf{d}}$ in (20), we find that the evaluation of $\hat{\mathbf{d}}$ only involves the inversion of a diagonal matrix $\sigma^2\rho^{-2}\mathbf{I} + \mathbf{H}^H\mathbf{H}$ and multiplications by diagonal or FFT matrices. Hence, the complexity associated with computing $\hat{\mathbf{d}}$ is $\mathcal{O}(N \log N)$.

The computation for $\hat{\boldsymbol{\theta}}$ in (21) is more involved and its complexity depends on our assumptions about $\boldsymbol{\Phi}$. We now present two simplifying designs for both the Wiener and Gaussian PHN models. The simplifying techniques used here are similar to those for estimating PHN within the channel estimation stage using the JCPCE algorithm [11], suggesting that the same processing unit may be used for both tasks in actual implementation.

TABLE VI
CG ALGORITHM FOR EVALUATING $\hat{\boldsymbol{\theta}}$

Initialization:	
	$\hat{\boldsymbol{\theta}}_0 = \mathbf{0}$
	$\boldsymbol{\gamma}_0 = [\boldsymbol{\Psi}^{-1} + \hat{\mathbf{X}}^H\hat{\mathbf{X}}]\hat{\boldsymbol{\theta}}_0 - \mathbf{q} = -\mathbf{q}$
	$\boldsymbol{\nu}_0 = -\boldsymbol{\gamma}_0 = \mathbf{q}$
For	$k = 0 : i-1$
	$\alpha_k = \boldsymbol{\gamma}_k^H\boldsymbol{\gamma}_k/(\boldsymbol{\nu}_k^H[\boldsymbol{\Psi}^{-1} + \hat{\mathbf{X}}^H\hat{\mathbf{X}}]\boldsymbol{\nu}_k)$
	$\hat{\boldsymbol{\theta}}_{k+1} = \hat{\boldsymbol{\theta}}_k + \alpha_k\boldsymbol{\nu}_k$
	$\boldsymbol{\gamma}_{k+1} = \boldsymbol{\gamma}_k + \alpha_k[\boldsymbol{\Psi}^{-1} + \hat{\mathbf{X}}^H\hat{\mathbf{X}}]\boldsymbol{\nu}_k$
	$\beta_{k+1} = \frac{\boldsymbol{\gamma}_{k+1}^H\boldsymbol{\gamma}_{k+1}}{\boldsymbol{\gamma}_k^H\boldsymbol{\gamma}_k}$
	$\boldsymbol{\nu}_{k+1} = -\boldsymbol{\gamma}_{k+1} + \beta_{k+1}\boldsymbol{\nu}_k$
End	

A. Wiener PHN

The inverse of Wiener PHN covariance matrix $\boldsymbol{\Phi}$ has a convenient tridiagonal structure [23]. If we let $\boldsymbol{\Psi} = \sigma^{-2}\boldsymbol{\Phi}$, $\boldsymbol{\Psi}^{-1} = \sigma^2\boldsymbol{\Phi}^{-1}$ can be written as

$$\boldsymbol{\Psi}^{-1} = \frac{\sigma^2}{\alpha_\phi^2}\begin{bmatrix} 2 & -1 & & & \mathbf{0} \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ \mathbf{0} & & & -1 & 1 \end{bmatrix}. \tag{25}$$

Let $\mathbf{q} = \mathrm{Im}[\hat{\mathbf{X}}^H(\mathbf{r} - \mathbf{F}^H\mathbf{H}\hat{\mathbf{d}})]$, where $\mathbf{q}$ can be computed efficiently using FFT since all matrices involved in calculating $\mathbf{q}$ are either diagonal or FFT matrices. The evaluation of $\hat{\boldsymbol{\theta}}$ is now equivalent to solving a linear equation $[\boldsymbol{\Psi}^{-1} + \hat{\mathbf{X}}^H\hat{\mathbf{X}}]\hat{\boldsymbol{\theta}} = \mathbf{q}$. This problem can be easily tackled by the CG method. The complete algorithm is presented in Table VI.

The tridiagonal form of $\boldsymbol{\Psi}^{-1}$ helps to reduce the dominant complexity in Table VI, $[\boldsymbol{\Psi}^{-1} + \hat{\mathbf{X}}^H\hat{\mathbf{X}}]\boldsymbol{\nu}_k$, to merely $6N$ operations. Thus, the overall complexity of every iteration of the CG algorithm is $\mathcal{O}(N)$. The CG algorithm requires a maximum of N iterations to converge to the exact solution. However, simulations in Section VI show that little performance degradation is introduced by setting $i = 8$. In conclusion, for Wiener PHN, the complexity of evaluating $\hat{\boldsymbol{\theta}}$ is $\mathcal{O}(iN)$, where i is the number of iterations in the CG algorithm.

B. Gaussian PHN

In the case of Gaussian PHN, we notice that $\boldsymbol{\Psi}$, as a Toeplitz matrix, can be approximated by a circulant matrix $\tilde{\boldsymbol{\Psi}}$ [24], [25]. Letting $\boldsymbol{\psi}^T = [\psi_0, \cdots, \psi_{N-1}]$ be the first row of $\boldsymbol{\Psi}$, then the first row of $\tilde{\boldsymbol{\Psi}}$, $\tilde{\boldsymbol{\psi}}^T = [\tilde{\psi}_0, \cdots, \tilde{\psi}_{N-1}]$, may be written as

$$\tilde{\psi}_i = \frac{(N-i)\psi_i + i\psi_{N-i}}{N}. \tag{26}$$

It can be shown that this approximation is asymptotically exact as $N \to \infty$ for an autocorrelation matrix $\boldsymbol{\Psi}$ of a first-order autoregressive process, which is a good fit to the Gaussian PHN process assumed in [26]. Replacing $\boldsymbol{\Psi}$ by $\tilde{\boldsymbol{\Psi}}$, the modified estimator for $\boldsymbol{\theta}$ becomes

$$\hat{\boldsymbol{\theta}} = [\tilde{\boldsymbol{\Psi}}^{-1} + \hat{\mathbf{X}}^H\hat{\mathbf{X}}]^{-1} \cdot \mathrm{Im}\left[\hat{\mathbf{X}}^H(\mathbf{r} - \mathbf{F}^H\mathbf{H}\mathbf{d})\right]. \tag{27}$$

This problem can be treated similar to the Wiener PHN case using the CG method in Table VI by replacing $\boldsymbol{\Psi}$ with $\tilde{\boldsymbol{\Psi}}$. Specifically, the evaluation of $[\tilde{\boldsymbol{\Psi}}^{-1} + \hat{\mathbf{X}}^H\hat{\mathbf{X}}]\boldsymbol{\nu}_k$ requires $4N + 2N\log N$ operations. Therefore, for Gaussian PHN, the overall computational complexity of $\hat{\boldsymbol{\theta}}$ is $\mathcal{O}(iN\log N)$.

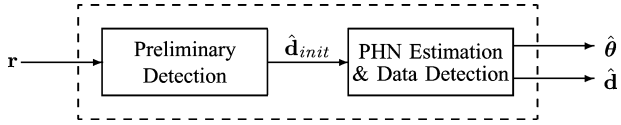

Fig. 2. Structure of the data detection module incorporating PHN mitigation.

VI. SIMULATIONS

To verify the effectiveness of the proposed PHN cancellation schemes, we present a set of simulations as follows. The data detection module is depicted in Fig. 2, where the received signal first goes through a preliminary detection stage that detects the transmitted data $\hat{\mathbf{d}}_{\text{init}}$ ignoring PHN distortion. Such a decision is inaccurate, but is necessary to initialize the next stage, which is the focus of this paper. In the proposed schemes, $\hat{\mathbf{d}}_{\text{init}}$ is used as $\mathbf{m}_d^{(0)}$ or $\hat{\mathbf{d}}^{(0)}$. We also simulate the conventional algorithm (i.e., GPNS in Section III-A) for comparison. It is implemented differently from the original paper [6] since no pilot symbols are allocated in our simulation setting for estimating the common PHN U_0 and ICI-plus-noise power $2\sigma_{\text{tot}}^2$. Instead, we assume perfect knowledge for U_0 and calculate $2\sigma_{\text{tot}}^2$ from the known channel response and phase noise statistics. In effect, this results in the best-achievable performance of GPNS.

The following system parameters are assumed in our simulations: 1) A Rayleigh multipath fading channel with a delay of $L = 10$ taps and an exponentially decreasing power delay profile that has a decay constant of four taps. 2) An OFDM symbol size of $N = 64$ subcarriers with each subcarrier modulated in 64-QAM format. 3) Baseband sampling rate $f_s = 20$ MHz (subcarrier spacing of 312.5 KHz). 4) The Wiener PHN is generated as a random-walk process with incremental PHN of $\alpha_\phi = 0.5°$. The covariance matrix $\boldsymbol{\Phi}$ is as depicted in (2). 5) The Gaussian PHN has a standard deviation of $\theta_{\text{rms}} = 3°$ (i.e.,$R_\theta(0) = (\pi\theta_{\text{rms}}/180)^2$). It is generated, according to the Matlab code recommended for the IEEE 802.11g standard [26], as i.i.d. Gaussian samples passed through a single pole Butterworth filter of 3-dB bandwidth $\Omega_o = 100$ KHz. Hence, the PHN covariance matrix $\boldsymbol{\Phi}$ is $\boldsymbol{\Phi}_{i,j} = (\pi\theta_{\text{rms}}/180)^2 e^{-(2\pi\Omega_o|i-j|/f_s)}$.

Fig. 3 compares the actual PHN profile with the PHN profile estimated using the variational inference algorithm (Table I with $n = 3$ iterations) at 30-dB SNR (E_b/N_o). The average PHN is also plotted. Through variational inference, we have very accurately estimated the phase noise profile, resulting in a much improved BER performance, as will be shown in Figs. 4 and 5. Note that the PHN estimated via variational inference is a distribution rather than a fixed value. Thus, in addition to plotting the mean $\mathbf{m}_\theta$, we also indicate one standard deviation around the mean, extracted from the diagonal elements of $\mathbf{S}_\theta$. The standard deviation quantitatively predicts the reliability of our estimates. Such an accuracy measure is also available in $\mathbf{m}_d$, but is not shown here graphically.

In Fig. 4, we demonstrate the performance of the proposed joint detector/estimator compared to the conventional method. The dotted line indicates the BER of a OFDM receiver free of PHN (the ideal scenario), and the solid line indicates the BER

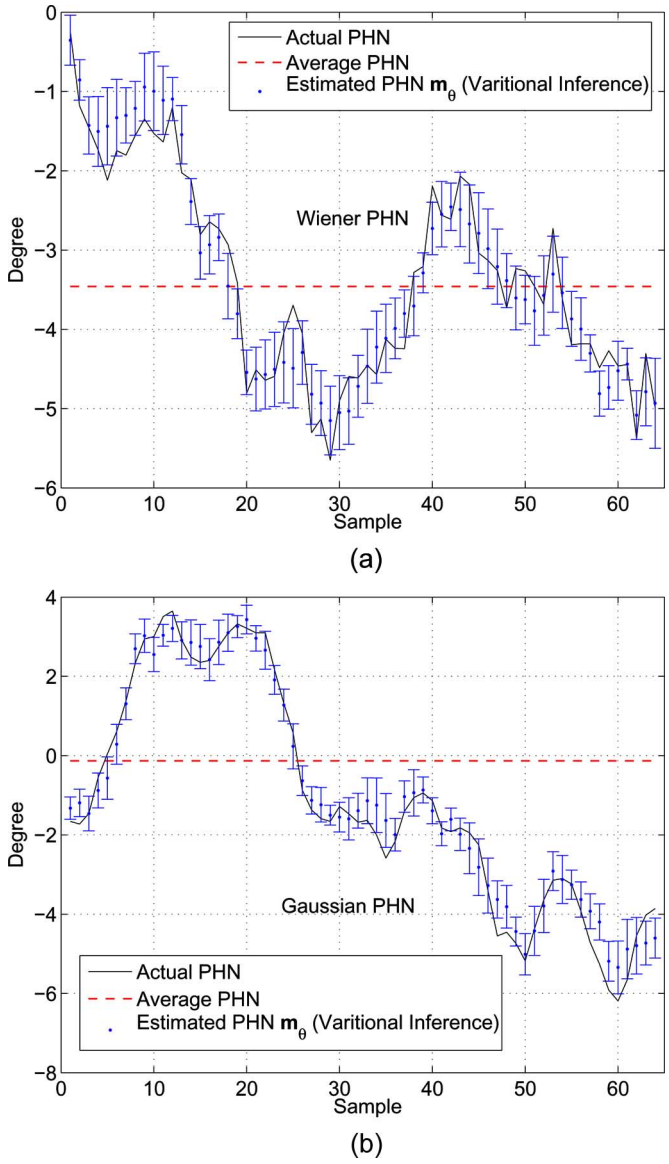

Fig. 3. Instance of PHN sequence estimated using the conventional method and variational inference. (a) Wiener PHN; (b) Gaussian PHN.

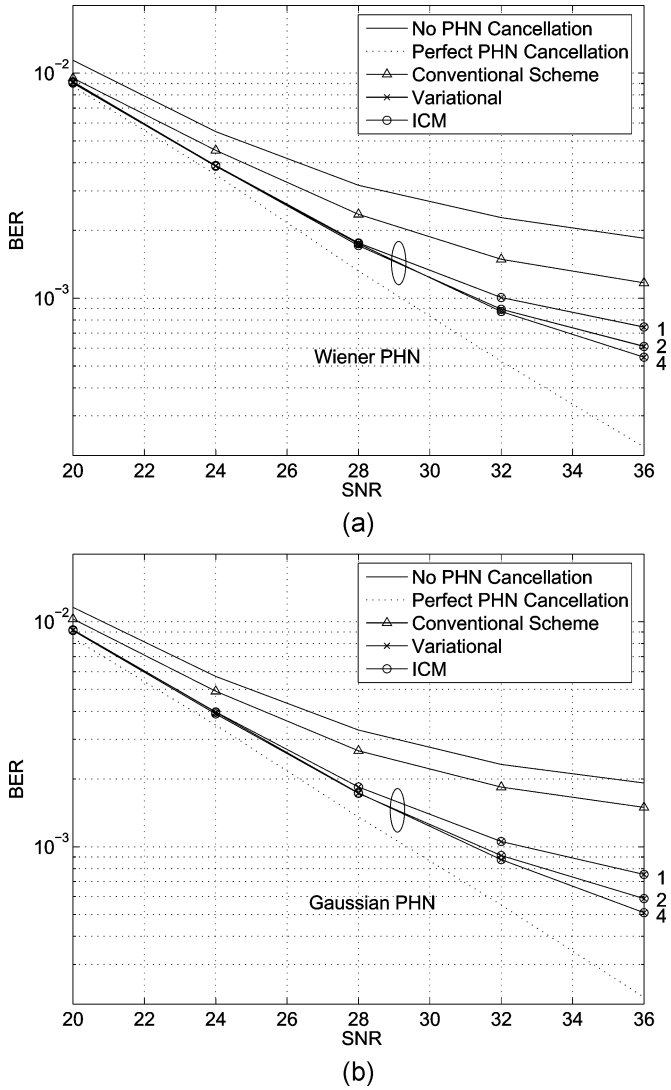

Fig. 4. BER performance comparison between the conventional method and proposed detection schemes. (a) Wiener PHN; (b) Gaussian PHN.

of an OFDM receiver with PHN but without PHN mitigation (the worst case scenario). It should be noted that the system without PHN mitigation does not fail only because of the simulation assumption of good phase synchronization at the beginning of each OFDM symbol. In between these two curves are the BER performance of receivers implementing the conventional PHN cancellation method (triangles) and the proposed schemes (crosses and circles). We plot the performance of the proposed algorithms after one, two, and four iterations (the BER does not improve significantly with more iterations). It is obvious that both the variational and ICM algorithms significantly outperform the conventional one, even though we assume the conventional scheme has perfect knowledge of U_0. The curves for EM I and EM II are not shown here since they overlap with the curves in the plot, implying identical performance. The superiority of variational inference over ICM is not evident here since we only consider an uncoded system, where the extra reliability information on $\mathbf{d}$ is not fully utilized. A coded OFDM system via a turbo detection scheme is outside the scope of this paper but is currently under investigation.

In Fig. 5, we study the performance of the low complexity simplified ICM technique by evaluating $\hat{\boldsymbol{\theta}}$ through $i = 8$ CG iterations as prescribed in Table VI. Again, we plot the performance after one, two, and four ICM iterations as in Table II. Compared to Fig. 4, it is seen that the OFDM receiver performs almost equally well, demonstrating that the simplified ICM scheme can be implemented efficiently in a practical receiver.

VII. Conclusion

This paper, together with [14], presents a complete physical layer design strategy for OFDM receivers in the presence of both CFO and PHN. Assuming the channel response and CFO have been accurately estimated adopting the methodology in [14], here we put forward a novel and low-complexity OFDM detection scheme based on variational inference techniques that combats PHN impairment. While the theoretical framework we have

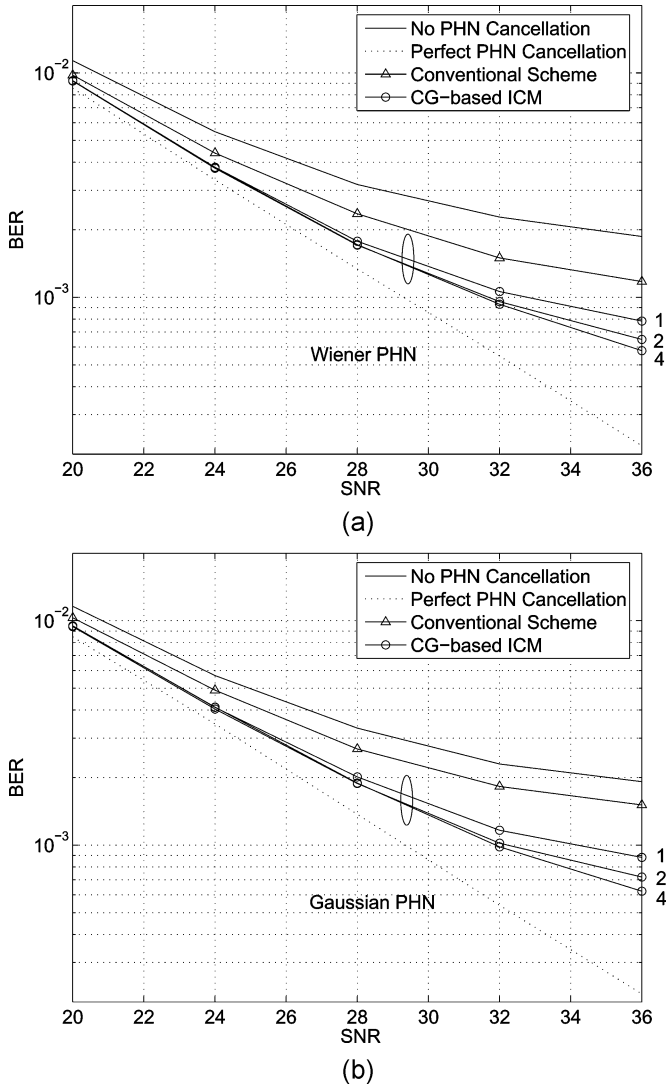

Fig. 5. BER performance of the low complexity detection scheme. (a) Wiener PHN; (b) Gaussian PHN.

introduced is new, the resulting design solution is simple, elegant, and with intuitive appeal. It is in essence an algorithm that updates the estimates for data and PHN iteratively. We have also demonstrated that each step of the iteration can be performed efficiently using FFT-based computations.

The possible extensions to this work are numerous, and we summarize a few of them as follows.

- *Variational EM Channel Estimation*: Additional parameter estimation can be naturally incorporated in the variational PHN cancellation scheme. For example, a variational EM algorithm that iteratively updates the inaccurate channel information can be derived by including $\mathbf{H}$ as an uncertain parameter in the free energy expression (14).
- *Coded OFDM*: In uncoded systems, the advantage of variational inference over ICM is not fully seen. One of variational inference's special characteristics is that the decision for data vector $\mathbf{d}$ is a distribution rather than a value. In this paper, we have not exploited this "soft decision" feature, but it is easy to see that in a coded OFDM system, this feature enables "turbo OFDM detection" where the soft extrinsic information of the data symbols are passed between the detector and a soft-in-soft-out (SISO) decoder [27].
- *OFDMA PHN Cancellation*: In a multiuser OFDM system, each user suffers from a different PHN pattern. The challenge is then to first separate the signals from different users and then cancel the PHN from each user's signal. The signal separation stage is suggested by the E-step in [28].

Appendix I
Closed-Form Expression of $p(\mathbf{r}|\mathbf{d})$

Since $p(\mathbf{r}|\mathbf{d},\boldsymbol{\theta})$ and $p(\boldsymbol{\theta})$ are Gaussian distributed, it can be shown that the distribution of $p(\mathbf{r}|\mathbf{d})$ is also Gaussian. Denoting the mean of $\mathbf{r}$ given $\mathbf{d}$ to be $\mathrm{E}(\mathbf{r}|\mathbf{d})$ and the variance as $\mathrm{V}(\mathbf{r}|\mathbf{d})$, then applying the Iterated Expectation Theorem [29, ch. 14] and its analog in covariance, yields

$$\begin{aligned}\mathrm{E}(\mathbf{r}|\mathbf{d}) &= \mathrm{E}_\theta\left[\mathrm{E}_r(\mathbf{r}|\mathbf{d},\boldsymbol{\theta})\right]\\ \mathrm{V}(\mathbf{r}|\mathbf{d}) &= \mathrm{V}_\theta\left[\mathrm{E}_r(\mathbf{r}|\mathbf{d},\boldsymbol{\theta})\right] + \mathrm{E}_\theta\left[\mathrm{V}_r(\mathbf{r}|\mathbf{d},\boldsymbol{\theta})\right].\end{aligned} \tag{28}$$

Because $p(\mathbf{r}|\mathbf{d},\boldsymbol{\theta}) = \mathcal{CN}(\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{d})(\mathbf{1}+j\boldsymbol{\theta}), 2\sigma^2\mathbf{I})$, it is straightforward to infer that

$$\begin{aligned}\mathrm{E}_r(\mathbf{r}|\mathbf{d},\boldsymbol{\theta}) &= \mathbf{F}^H\mathbf{H}\mathbf{d} + j\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{d})\boldsymbol{\theta}\\ \mathrm{V}_r(\mathbf{r}|\mathbf{d},\boldsymbol{\theta}) &= 2\sigma^2\mathbf{I}.\end{aligned} \tag{29}$$

Given that $p(\boldsymbol{\theta}) = \mathcal{N}(\mathbf{0},\boldsymbol{\Phi})$, with some further manipulation, we obtain

$$\begin{aligned}\mathrm{E}_\theta\left[\mathrm{E}_r(\mathbf{r}|\mathbf{d},\boldsymbol{\theta})\right] &= \mathbf{F}^H\mathbf{H}\mathbf{d}\\ \mathrm{V}_\theta\left[\mathrm{E}_r(\mathbf{r}|\mathbf{d},\boldsymbol{\theta})\right] &= \mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{d})\boldsymbol{\Phi}\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{d})^H\\ \mathrm{E}_\theta\left[\mathrm{V}_r(\mathbf{r}|\mathbf{d},\boldsymbol{\theta})\right] &= 2\sigma^2\mathbf{I}\end{aligned} \tag{30}$$

which implies that

$$\begin{aligned}\mathrm{E}(\mathbf{r}|\mathbf{d}) &= \mathbf{F}^H\mathbf{H}\mathbf{d}\\ \mathrm{V}(\mathbf{r}|\mathbf{d}) &= \mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{d})\boldsymbol{\Phi}\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{d})^H + 2\sigma^2\mathbf{I}.\end{aligned} \tag{31}$$

Therefore

$$p(\mathbf{r}|\mathbf{d}) = \mathcal{CN}\left(\mathbf{F}^H\mathbf{H}\mathbf{d}, \mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{d})\boldsymbol{\Phi}\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{d})^H + 2\sigma^2\mathbf{I}\right). \tag{32}$$

Appendix II
Closed-Form Expression of $\mathcal{F}(Q,p)$

A simple expansion of the variational free energy in (12) shows that

$$\begin{aligned}\mathcal{F}(Q,p) &= \int_{\mathbf{d},\boldsymbol{\theta}} Q(\mathbf{d})Q(\boldsymbol{\theta})\log\frac{Q(\mathbf{d})Q(\boldsymbol{\theta})}{p(\mathbf{d},\boldsymbol{\theta},\boldsymbol{r})}d\mathbf{d}d\boldsymbol{\theta}\\ &= \int_{\mathbf{d},\boldsymbol{\theta}} Q(\mathbf{d})Q(\boldsymbol{\theta})\log\frac{Q(\mathbf{d})Q(\boldsymbol{\theta})}{p(\mathbf{r}|\mathbf{d},\boldsymbol{\theta})p(\mathbf{d})p(\boldsymbol{\theta})}d\mathbf{d}d\boldsymbol{\theta}\\ &= -\int_{\mathbf{d}} Q(\mathbf{d})\log p(\mathbf{d})d\mathbf{d} - \int_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})\log p(\boldsymbol{\theta})d\boldsymbol{\theta}\\ &\quad + \int_{\mathbf{d}} Q(\mathbf{d})\log Q(\mathbf{d})d\mathbf{d} + \int_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})\log Q(\boldsymbol{\theta})d\boldsymbol{\theta}\\ &\quad - \int_{\mathbf{d},\boldsymbol{\theta}} Q(\mathbf{d})Q(\boldsymbol{\theta})\log p(\mathbf{r}|\mathbf{d},\boldsymbol{\theta})d\mathbf{d}d\boldsymbol{\theta}.\end{aligned} \tag{33}$$

Substituting (10) and (13) into the expression above and applying Gaussian expectation properties [30], we may obtain each term of the variational free energy as a function of the Q-distribution parameters (with the constant terms omitted)

$$-\int_{\mathbf{d}} Q(\mathbf{d})\log p(\mathbf{d})d\mathbf{d} = \frac{1}{2\rho^2}\int_{\mathbf{d}} Q(\mathbf{d})(\mathbf{d}^H\mathbf{d})d\mathbf{d} = \frac{1}{2\rho^2}\left(\mathrm{tr}(\mathbf{S}_d)+\mathbf{m}_d^H\mathbf{m}_d\right) \tag{34}$$

$$-\int_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})\log p(\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{1}{2}\int_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})(\boldsymbol{\theta}^T\boldsymbol{\Phi}^{-1}\boldsymbol{\theta})d\boldsymbol{\theta} = \frac{1}{2}\mathrm{tr}(\boldsymbol{\Phi}^{-1}\mathbf{S}_\theta)+\frac{1}{2}\mathbf{m}_\theta^T\boldsymbol{\Phi}^{-1}\mathbf{m}_\theta \tag{35}$$

$$\int_{\mathbf{d}} Q(\mathbf{d})\log Q(\mathbf{d}) = -\log|\mathbf{S}_d| \tag{36}$$

$$\int_{\boldsymbol{\theta}} Q(\boldsymbol{\theta})\log Q(\boldsymbol{\theta}) = -\frac{1}{2}\log|\mathbf{S}_\theta| \tag{37}$$

$$\begin{aligned}
&-2\sigma^2\int_{\mathbf{d},\boldsymbol{\theta}} Q(\mathbf{d})Q(\boldsymbol{\theta})\log p(\mathbf{r}|\mathbf{d},\boldsymbol{\theta})d\mathbf{d}d\boldsymbol{\theta}\\
&=\int_{\mathbf{d},\boldsymbol{\theta}} Q(\mathbf{d})Q(\boldsymbol{\theta})(\mathbf{r}-\mathbf{P}\mathbf{F}^H\mathbf{H}\mathbf{d})^H(\mathbf{r}-\mathbf{P}\mathbf{F}^H\mathbf{H}\mathbf{d})d\mathbf{d}d\boldsymbol{\theta}\\
&=\int_{\mathbf{d},\boldsymbol{\theta}} Q(\mathbf{d})Q(\boldsymbol{\theta})\left[\mathbf{r}-\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{d})(\mathbf{1}+j\boldsymbol{\theta})\right]^H\\
&\quad\times\left[\mathbf{r}-\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{d})(\mathbf{1}+j\boldsymbol{\theta})\right]d\mathbf{d}d\boldsymbol{\theta}\\
&=\int_{\mathbf{d},\boldsymbol{\theta}} Q(\mathbf{d})Q(\boldsymbol{\theta})\left[\mathbf{r}-\mathbf{F}^H\mathbf{H}\mathbf{d}-j\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{d})\boldsymbol{\theta}\right]^H\\
&\quad\times\left[\mathbf{r}-\mathbf{F}^H\mathbf{H}\mathbf{d}-j\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{d})\boldsymbol{\theta}\right]d\mathbf{d}d\boldsymbol{\theta}\\
&=\int_{\mathbf{d}} Q(\mathbf{d})\Big(\mathbf{d}^H\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\mathbf{d}\\
&\quad+\left[(j\mathbf{M}_\theta+\mathbf{I})\mathbf{F}^H\mathbf{H}\mathbf{d}-\mathbf{r}\right]^H\left[(j\mathbf{M}_\theta+\mathbf{I})\mathbf{F}^H\mathbf{H}\mathbf{d}-\mathbf{r}\right]\Big)d\mathbf{d}\\
&=\mathrm{tr}\left[(j\mathbf{M}_\theta+\mathbf{I})\mathbf{F}^H\mathbf{H}\mathbf{S}_d\mathbf{H}^H\mathbf{F}(j\mathbf{M}_\theta+\mathbf{I})^H\right]\\
&\quad+\mathrm{tr}\left[\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\mathbf{S}_d\right]\\
&\quad+\mathbf{m}_d^H\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\mathbf{m}_d\\
&\quad+\left[(j\mathbf{M}_\theta+\mathbf{I})\mathbf{F}^H\mathbf{H}\mathbf{m}_d-\mathbf{r}\right]^H\\
&\quad\times\left[(j\mathbf{M}_\theta+\mathbf{I})\mathbf{F}^H\mathbf{H}\mathbf{m}_d-\mathbf{r}\right]
\end{aligned} \tag{38}$$

where $\mathbf{M}_\theta=\mathrm{diag}(\mathbf{m}_\theta)$. Finally, the complete free energy expression is assembled

$$\begin{aligned}
&\mathcal{F}(\mathbf{m}_d,\mathbf{S}_d,\mathbf{m}_\theta,\mathbf{S}_\theta)\\
&=\frac{1}{2\rho^2}\left[\mathrm{tr}(\mathbf{S}_d)+\mathbf{m}_d^H\mathbf{m}_d\right]+\frac{1}{2}\mathrm{tr}(\boldsymbol{\Phi}^{-1}\mathbf{S}_\theta)\\
&\quad+\frac{1}{2}\mathbf{m}_\theta^T\boldsymbol{\Phi}^{-1}\mathbf{m}_\theta-\frac{1}{2}\log|\mathbf{S}_\theta|-\log|\mathbf{S}_d|\\
&\quad+\frac{1}{2\sigma^2}\left\{\mathrm{tr}\left[(j\mathbf{M}_\theta+\mathbf{I})\mathbf{F}^H\mathbf{H}\mathbf{S}_d\mathbf{H}^H\mathbf{F}(j\mathbf{M}_\theta+\mathbf{I})^H\right]\right.\\
&\quad+\mathrm{tr}\left[\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\mathbf{S}_d\right]\\
&\quad+\mathbf{m}_d^H\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\mathbf{m}_d\\
&\quad+\left[(j\mathbf{M}_\theta+\mathbf{I})\mathbf{F}^H\mathbf{H}\mathbf{m}_d-\mathbf{r}\right]^H\\
&\quad\left.\times\left[(j\mathbf{M}_\theta+\mathbf{I})\mathbf{F}^H\mathbf{H}\mathbf{m}_d-\mathbf{r}\right]\right\}.
\end{aligned} \tag{39}$$

APPENDIX III
DERIVATION OF VARIATIONAL INFERENCE ALGORITHM

In each iteration, we minimize $\mathcal{F}(\mathbf{m}_d,\mathbf{S}_d,\mathbf{m}_\theta,\mathbf{S}_\theta)$ in (39) w.r.t. the variables $\mathbf{S}_\theta$, $\mathbf{m}_\theta$, $\mathbf{S}_d$ and $\mathbf{m}_d$ in turn.

1) For $\mathbf{S}_\theta$

$$\begin{aligned}
&\partial\mathcal{F}/\partial\mathbf{S}_\theta^{-1}\\
&=\frac{\partial}{\partial\mathbf{S}_\theta^{-1}}\left\{\frac{1}{2}\mathrm{tr}(\boldsymbol{\Phi}^{-1}\mathbf{S}_\theta)+\frac{1}{2\sigma^2}\mathrm{tr}\left[\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\mathbf{S}_d\right]\right.\\
&\qquad\left.+\frac{1}{2}\log\left|\mathbf{S}_\theta^{-1}\right|+\frac{1}{2\sigma^2}\mathbf{m}_d^H\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\mathbf{m}_d\right\}\\
&=\frac{\partial}{\partial\mathbf{S}_\theta^{-1}}\left\{\frac{1}{2}\mathrm{tr}(\boldsymbol{\Phi}^{-1}\mathbf{S}_\theta)+\frac{1}{2\sigma^2}\mathrm{tr}\left[\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{S}_d\mathbf{H}^H\mathbf{F})\mathbf{S}_\theta\right]\right.\\
&\qquad\left.+\frac{1}{2}\log\left|\mathbf{S}_\theta^{-1}\right|+\frac{1}{2\sigma^2}\mathrm{tr}\left[\mathbf{X}_m^H\mathbf{S}_\theta\mathbf{X}_m\right]\right\}\\
&=-\frac{1}{2}\mathbf{S}_\theta^T\boldsymbol{\Phi}^{-1}\mathbf{S}_\theta^T+\frac{1}{2}\mathbf{S}_\theta^T\\
&\quad-\frac{1}{2\sigma^2}\left[\mathbf{S}_\theta^T\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{S}_d\mathbf{H}^H\mathbf{F})\mathbf{S}_\theta^T+\mathbf{S}_\theta^T\left(\mathbf{X}_m\mathbf{X}_m^H\right)^T\mathbf{S}_\theta^T\right]\\
&=-\frac{1}{2}\mathbf{S}_\theta^T\left[\boldsymbol{\Phi}^{-1}\mathbf{S}_\theta^T-\mathbf{I}+\frac{1}{\sigma^2}\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{S}_d\mathbf{H}^H\mathbf{F})\mathbf{S}_\theta^T\right.\\
&\qquad\left.+\frac{1}{\sigma^2}\left(\mathbf{X}_m\mathbf{X}_m^H\right)^T\mathbf{S}_\theta^T\right]
\end{aligned} \tag{40}$$

where $\mathbf{X}_m=\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{m}_d)$. Thus, $\partial\mathcal{F}/\partial\mathbf{S}_\theta^{-1}=\mathbf{0}$ leads to (15).

2) For $\mathbf{m}_\theta$

$$\begin{aligned}
&\partial\mathcal{F}/\partial\mathbf{m}_\theta\\
&=\frac{\partial}{\partial\mathbf{m}_\theta}\left\{\frac{1}{2}\mathbf{m}_\theta^T\boldsymbol{\Phi}^{-1}\mathbf{m}_\theta+\frac{1}{2\sigma^2}\right.\\
&\qquad\times\mathrm{tr}\left[(j\mathbf{M}_\theta+\mathbf{I})\mathbf{F}^H\mathbf{H}\mathbf{S}_d\mathbf{H}^H\mathbf{F}(j\mathbf{M}_\theta+\mathbf{I})^H\right]\\
&\qquad+\frac{1}{2\sigma^2}\left[(j\mathbf{M}_\theta+\mathbf{I})\mathbf{F}^H\mathbf{H}\mathbf{m}_d-\mathbf{r}\right]^H\\
&\qquad\left.\times\left[(j\mathbf{M}_\theta+\mathbf{I})\mathbf{F}^H\mathbf{H}\mathbf{m}_d-\mathbf{r}\right]\right\}\\
&=\frac{\partial}{\partial\mathbf{m}_\theta}\left\{\frac{1}{2}\mathbf{m}_\theta^T\boldsymbol{\Phi}^{-1}\mathbf{m}_\theta+\frac{1}{2\sigma^2}\mathbf{m}_\theta^T\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{S}_d\mathbf{H}^H\mathbf{F})\mathbf{m}_\theta\right.\\
&\qquad+\frac{1}{2\sigma^2}(j\mathbf{X}_m\mathbf{m}_\theta+\mathbf{F}^H\mathbf{H}\mathbf{m}_d-\boldsymbol{r})^H\\
&\qquad\left.\times(j\mathbf{X}_m\mathbf{m}_\theta+\mathbf{F}^H\mathbf{H}\mathbf{m}_d-\boldsymbol{r})\right\}\\
&=\boldsymbol{\Phi}^{-1}\mathbf{m}_\theta+\frac{1}{\sigma^2}\left[\mathrm{diag}(\mathbf{F}^H\mathbf{H}\mathbf{S}_d\mathbf{H}^H\mathbf{F})\mathbf{m}_\theta+\mathbf{X}_m^H\mathbf{X}_m\mathbf{m}_\theta\right]\\
&\quad-\frac{1}{\sigma^2}\mathrm{Re}\left[j\mathbf{X}_m^H(\mathbf{F}^H\mathbf{H}\mathbf{m}_d-\boldsymbol{r})\right].
\end{aligned} \tag{41}$$

Thus, $\partial\mathcal{F}/\partial\mathbf{m}_\theta=\mathbf{0}$ leads to (16).

3) For $\mathbf{S}_d$

$$\begin{aligned}
&\partial\mathcal{F}/\partial\mathbf{S}_d^{-1}\\
&=\frac{\partial}{\partial\mathbf{S}_d^{-1}}\left\{\frac{1}{2\rho^2}\mathrm{tr}(\mathbf{S}_d)+\log\left|\mathbf{S}_d^{-1}\right|\right.\\
&\qquad+\frac{1}{2\sigma^2}\mathrm{tr}\left[\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\mathbf{S}_d\right]\\
&\qquad\left.+\frac{1}{2\sigma^2}\mathrm{tr}\left[\mathbf{P}_m\mathbf{F}^H\mathbf{H}\mathbf{S}_d\mathbf{H}^H\mathbf{F}\mathbf{P}_m^H\right]\right\}
\end{aligned}$$

$$= -\frac{1}{2\rho^2}\mathbf{S}_d^T\mathbf{S}_d^T + \mathbf{S}_d^T - \frac{1}{2\sigma^2}\left(\mathbf{S}_d\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\mathbf{S}_d\right)^T - \frac{1}{2\sigma^2}\left(\mathbf{S}_d\mathbf{H}^H\mathbf{F}\mathbf{P}_m^H\mathbf{P}_m\mathbf{F}^H\mathbf{H}\mathbf{S}_d\right)^T$$
$$= -\mathbf{S}_d\left[\frac{1}{2\rho^2}\mathbf{S}_d - \mathbf{I} + \frac{1}{2\sigma^2}\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\mathbf{S}_d + \frac{1}{2\sigma^2}\mathbf{H}^H\mathbf{H}\mathbf{S}_d\right] \quad (42)$$

where $\mathbf{P}_m = \mathrm{diag}(e^{j\mathbf{m}_\theta}) \approx (j\mathbf{M}_\theta + \mathbf{I})$. Thus, $\partial\mathcal{F}/\partial\mathbf{S}_d^{-1} = \mathbf{0}$ leads to (17).

4) For $\mathbf{m}_d$

$$\partial\mathcal{F}/\partial\mathbf{m}_d^* = \frac{\partial}{\partial\mathbf{m}_d^*}\left\{\frac{1}{2\rho^2}\mathbf{m}_d^H\mathbf{m}_d + \frac{1}{2\sigma^2}\mathbf{m}_d^H\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\mathbf{m}_d + \frac{1}{2\sigma^2}[\mathbf{P}_m\mathbf{F}^H\mathbf{H}\mathbf{m}_d - \mathbf{r}]^H[\mathbf{P}_m\mathbf{F}^H\mathbf{H}\mathbf{m}_d - \mathbf{r}]\right\}$$
$$= \frac{1}{2\rho^2}\mathbf{m}_d + \frac{1}{2\sigma^2}\mathbf{H}^H\mathbf{F}\mathrm{diag}(\mathbf{S}_\theta)\mathbf{F}^H\mathbf{H}\mathbf{m}_d + \frac{1}{2\sigma^2}\mathbf{H}^H\mathbf{F}\mathbf{P}_m^H[\mathbf{P}_m\mathbf{F}^H\mathbf{H}\mathbf{m}_d - \mathbf{r}]. \quad (43)$$

Thus, $\partial\mathcal{F}/\partial\mathbf{m}_d^* = \mathbf{0}$ leads to (18).

Appendix IV
Derivation of ICM Algorithm

Given the complete negative log-likelihood function in (19), we again take the gradient descent approach and in each step minimize $\mathcal{L}(\hat{\mathbf{d}}, \hat{\boldsymbol{\theta}})$ w.r.t. $\hat{\mathbf{d}}$ and $\hat{\boldsymbol{\theta}}$ in turn.

1) For $\hat{\mathbf{d}}$

$$\partial\mathcal{L}/\partial\hat{\mathbf{d}}^* = \frac{1}{2\rho^2}\hat{\mathbf{d}} + \frac{1}{2\sigma^2}\mathbf{H}^H\mathbf{F}\hat{\mathbf{P}}^H(\hat{\mathbf{P}}\mathbf{F}^H\mathbf{H}\hat{\mathbf{d}} - \mathbf{r}) \quad (44)$$

where $\hat{\mathbf{P}} = \mathrm{diag}(e^{j\hat{\boldsymbol{\theta}}})$. Equating $\partial\mathcal{L}/\partial\hat{\mathbf{d}}^* = \mathbf{0}$ and we obtain (20).

2) For $\hat{\boldsymbol{\theta}}$

$$\partial\mathcal{L}/\partial\hat{\boldsymbol{\theta}} = \boldsymbol{\Phi}^{-1}\hat{\boldsymbol{\theta}} + \frac{1}{\sigma^2}\left[\hat{\mathbf{X}}^H\hat{\mathbf{X}}\hat{\boldsymbol{\theta}} - \mathrm{Re}\left(j\hat{\mathbf{X}}^H(\mathbf{F}^H\mathbf{H}\hat{\mathbf{d}} - \boldsymbol{r})\right)\right] \quad (45)$$

where $\hat{\mathbf{X}} = \mathrm{diag}(\mathbf{F}^H\mathbf{H}\hat{\mathbf{d}})$. Equating $\partial\mathcal{L}/\partial\hat{\boldsymbol{\theta}} = \mathbf{0}$ and we obtain (21).

Acknowledgment

The authors would like to thank M. Moldoveanu, Redline Communications, for initiating this research, and Y. Zhao, A. Kannan, and R. Pacheco for insightful discussions.

References

[1] R. van Nee, *OFDM Wireless Multimedia Communications*. Boston, MA: Artech House, 2000.
[2] T. Pollet, M. V. Bladel, and M. Moeneclaey, "BER sensitivity of OFDM systems to carrier frequency offset and Wiener phase noise," *IEEE Trans. Commun.*, vol. 43, no. 2, pp. 191–193, Feb. 1995.
[3] L. Tomba, "On the effect of Wiener phase noise in OFDM systems," *IEEE Trans. Commun.*, vol. 46, no. 5, pp. 580–583, May 1998.
[4] P. Robertson and S. Kaiser, "Analysis of the effects of phase noise in orthogonal frequency division multiplex (OFDM) systems," in *Proc. IEEE Int. Conf. Communications*, Jun. 1995, pp. 1652–1657.
[5] K. Nikitopoulos and A. Polydoros, "Phase-impairment effects and compensation algorithms for OFDM systems," *IEEE Trans. Commun.*, vol. 53, no. 4, pp. 698–707, Apr. 2005.
[6] S. Wu and Y. Bar-Ness, "OFDM systems in the presence of phase noise: Consequences and solutions," *IEEE Trans. Commun.*, vol. 52, no. 11, pp. 1988–1996, Nov. 2004.
[7] F. R. Kschischang, B. J. Frey, and H. A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, no. 2, pp. 498–519, Feb. 2001.
[8] G. Colavolpe, A. Barbieri, and G. Caire, "Algorithms for iterative decoding in the presence of strong phase noise," *IEEE J. Sel. Areas Commun.*, vol. 23, no. 9, pp. 1748–1757, Sep. 2005.
[9] B. J. Frey and N. Jojic, "A comparison of algorithms for inference and learning in probabilistic graphical models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 9, pp. 1392–1416, Sep. 2005.
[10] L. Piazzo and P. Mandarini, "Analysis of phase noise effects in OFDM modems," *IEEE Trans. Commun.*, vol. 50, no. 10, pp. 1696–1705, Oct. 2002.
[11] D. D. Lin, R. A. Pacheco, T. J. Lim, and D. Hatzinakos, "Optimal OFDM channel estimation with carrier frequency offset and phase noise," in *Proc. IEEE Wireless Communications Networking Conf.*, Apr. 2006, vol. 2, pp. 1050–1055.
[12] P. Chevillat, D. Maiwald, and G. Ungerboeck, "Rapid training of a voiceband data-modem receiver employing an equalizer with fractional-T spaced coefficients," *IEEE Trans. Commun.*, vol. 35, no. 9, pp. 869–876, Sep. 1987.
[13] T. Schmidl and D. Cox, "Robust frequency and timing synchronization for OFDM," *IEEE Trans. Commun.*, vol. 45, no. 12, pp. 1613–1621, Dec. 1997.
[14] D. D. Lin, R. A. Pacheco, T. J. Lim, and D. Hatzinakos, "Joint estimation of channel response, frequency offset and phase noise in OFDM," *IEEE Trans. Signal Process.*, vol. 54, no. 9, pp. 3542–3554, Sep. 2006.
[15] S. M. Kay, *Fundamentals of Statistical Signal Processing—Estimation Theory*. Upper Saddle River, NJ: Prentice-Hall PTR, 1993, ch. 10.7, pp. 328–330.
[16] D. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[17] V. Dotsenko, *The Theory of Spin Glasses and Neural Networks*. Singapore: World Scientific, 1994.
[18] B. J. Frey, *Graphical Models for Machine Learning and Digital Communication*. Cambridge, MA: MIT Press, 1998.
[19] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
[20] J. Nocedal and S. J. Wright, *Numerical Optimization*. New York: Springer-Verlag, 1999.
[21] R. M. Neal and G. E. Hinton, "A view of the EM algorithm that justifies incremental, sparse, and other variants," in *Learning in Graphical Models*, M. I. Jordan, Ed. Norwell, MA: Kluwer, 1998.
[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from imcomplete data via the EM algorithm," *J. Roy. Statist. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.
[23] A. Kavčić and J. M. F. Moura, "Matrices with banded inverses: Inversion algorithms and factorization of Gauss–Markov processes," *IEEE Trans. Inf. Theory*, vol. 46, no. 4, pp. 1495–1509, Jul. 2000.
[24] C. R. Vogel, *Computational Methods for Inverse Problems*. Philadelphia, PA: SIAM, 2002.
[25] T. F. Chan, "An optimal circulant preconditioner for Toeplitz systems," *SIAM J. Sci. Statist. Comput.*, vol. 9, no. 4, pp. 766–771, Jul. 1988.
[26] IEEE P802.11-Task Group G., 2000, Phase Noise Matlab Model [Online]. Available: http://www.grouper.ieee.org/groups/802/11/Reports/tgg_update.htm
[27] X. Wang and H. V. Poor, "Iterative (turbo) soft interference cancellation and decoding for coded CDMA," *IEEE Trans. Commun.*, vol. 47, no. 7, pp. 1046–1061, Jul. 1999.
[28] M. Feder and E. Weinstein, "Parameter estimation of superimposed signals using the EM algorithm," *IEEE Trans. Acoust. Speech, Signal, Process.*, vol. 36, no. 4, pp. 477–489, Apr. 1988.
[29] M. I. Jordan, *An Introduction to Probabilistic Graphical Models,* to be published.
[30] *"Matrix Reference Manual"* 2005 [Online]. Available: http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/expect.html, M. Brookes.

Darryl Dexu Lin (S'01) received the B.A.Sc. degree (with Honors) in engineering science and the M.A.Sc. degree in electrical engineering from the University of Toronto, Toronto, ON, Canada, in 2001 and 2003, respectively, where he is currently pursuing the Ph.D. degree.

His research interests include probabilistic inference and its applications in wireless communications. His work currently focuses on the mitigation time-varying distortions in multicarrier modulation and efficient joint detection and decoding in multiple access channels.

Teng Joon Lim (S'92–M'95–SM'02) received the B.Eng. degree from the National University of Singapore in 1992 and the Ph.D. degree from the University of Cambridge, Cambridge, U.K., in 1996.

He joined the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada, in December 2000, where he is currently an Associate Professor. In the five years prior to that, he was a Member of Technical Staff at the Centre for Wireless Communications, Singapore, serving as the Leader of the Digital Communications and Signal Processing Group.

His research interests are in wireless transceiver design, in particular multiuser detection, OFDM and OFDMA receiver structures, MIMO techniques, precoding in downlink channels, and cross-layer aspects of cooperative network design, and he has published widely in these areas. He contributes regularly in organizing conferences, serving on technical program committees, and organizing seminars for the IEEE Toronto Communications Chapter.