Abstract—In this paper, we propose a Bayesian channel estimator for intelligent reflecting surface-aided (IRS-aided) millimeter wave (mmWave) massive multiple-input multiple-output (MIMO) systems with semi-passive elements that can receive the signal in the active sensing mode. Ultimately, our goal is to minimize the channel estimation error using the received signal at the base station and additional information acquired from a small number of active sensors at the IRS. Unlike recent works on channel estimation with semi-passive elements that require both uplink and downlink training signals to estimate the UE-IRS and IRS-BS links, we only use uplink training signals to estimate all the links. To compute the minimum mean squared error (MMSE) estimates of all the links, we propose a novel variational inference-sparse Bayesian learning (VI-SBL) channel estimator that performs approximate posterior inference on the channel using VI with the mean-field approximation under the SBL framework. The simulation results show that VI-SBL outperforms the state-of-the-art baselines for IRS with passive reflecting elements in terms of the channel estimation accuracy and training overhead. Furthermore, VI-SBL with semi-passive elements is shown to be more spectral- and energy-efficient than the baselines with passive reflecting elements.

Index Terms—Channel estimation, intelligent reflecting surface (IRS), semi-passive element, variational inference (VI), sparse Bayesian learning (SBL).

I. INTRODUCTION

5G WIRELESS communications support high data rates by communicating in the millimeter wave (mmWave) band [1]. The high carrier frequency in the range of 30-300 GHz offers a large bandwidth, which results in a significant throughput gain. The problem, however, is that the severe path loss renders mmWave communications vulnerable to blockages. To overcome such an issue, intelligent reflecting surface (IRS) was recently proposed [2], which is an array of metamaterial-based passive reflecting elements capable of adjusting the amplitude and phase of the impinging signal. For the IRS to generate a favorable detour around a blockage, the reflection amplitude and phase shift must align with the channel, which necessitates accurate channel state information (CSI).

The distinct feature of channel estimation for IRS with passive reflecting elements is that the UE-IRS link of size \( NK \) and IRS-BS link of size \( MN \) where \( M, N, \) and \( K \) are the numbers of base station antennas, IRS elements, and single-antenna users. Since passive reflecting elements cannot observe the UE-IRS and IRS-BS links, the UE-IRS-BS link must be estimated from the reflected signal, which results in a significant training overhead.

Next, we review prior works [3], [4], [5], [6], [7], [8], [9], [10], [11] on channel estimation for IRS with passive reflecting elements that focus on training overhead reduction. In [3], a three-phase channel estimator is proposed inspired by the correlation between the channels of different users. In particular, [3] exploits the fact that the UE-IRS-BS links of different users share the same IRS-BS link. As a result, the three-phase channel estimator proceeds by estimating the UE-BS links of all the users in the first phase, UE-IRS-BS link of a particular user in the second phase, and UE-IRS-BS links of the remaining users in the third phase where most of the training overhead reduction occurs. Moving on to [4], channel estimation for IRS with discrete phase shifts is considered, for which a reflection design is proposed. In particular, the reflection design is optimized based on grouping IRS elements to reduce the channel estimation error. In [5], a channel estimator is proposed for the orthogonal frequency division multiple access (OFDMA) scenario based on the line-of-sight (LoS) assumption on the UE-BS link. Again, the

Digital Object Identifier 10.1109/TWC.2023.3273284

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
channel estimator exploits the fact that the UE-IRS-BS links of different users share the same IRS-BS link to reduce the training overhead as in [3]. In [6], the channel estimation problem is reformulated as a matrix factorization problem, which is solved using the message passing (MP) algorithm. In [7], a dual link training signal-based channel estimator is proposed. In particular, the dual link training signal transmits downlink training signals to the IRS, whose reflected version is used as uplink training signals to estimate the IRS-BS link. In [8], an atomic norm minimization-based channel estimator is proposed that extracts the angle parameters of the channel. In [9], a channel estimator is proposed based on the single-path approximation of the channel. In addition, [9] develops another channel estimator that exploits IRS phase shifts and training signals. In [10], the mmWave channel estimation is reformulated as a compressed sensing problem to capture the channel sparsity. In [11], the double-structured orthogonal matching pursuit (DS-OMP) channel estimator is proposed that exploits the common IRS-BS link shared by all the users. In particular, DS-OMP reduces the channel estimation error and training overhead by identifying the common rows and columns that the UE-IRS-BS links of all the users share by taking into account the double sparsity structure inherent in the IRS links.

The training overhead in [3], [4], [5], [6], [7], [8], [9], [10], and [11] is still high because estimating the UE-IRS-BS link with passive reflecting elements that cannot receive the signal is a challenging task. To overcome such an issue, IRS with semi-passive elements that can be switched to the active sensing mode was recently proposed in [12] and [13]. In particular, the received signal at the active sensors is leveraged to estimate the UE-IRS and IRS-BS links in the first coherence block. Then, [12] and [13] can replace UE-IRS-BS link estimation of size $MNK$ with UE-IRS link estimation of size $NK$ in the subsequent coherence blocks because the IRS-BS link remains constant over multiple coherence blocks, which results in significant channel estimation overhead reduction. The problem of [12] and [13], however, is that both uplink and downlink training signals are necessary to estimate all the links. Furthermore, [12] and [13] are semi-passive element-driven rather than aided because only the received signal at the active sensors is leveraged, while the received signal at the base station is discarded.

In this paper, a Bayesian channel estimator is proposed for IRS-aided mmWave massive multiple-input multiple-output (MIMO) systems with semi-passive elements. In particular, the semi-passive elements are activated to the active sensing mode in the channel estimation phase and deactivated to the passive reflecting mode in the data transmission phase. For the first time in the literature, we estimate the UE-IRS and IRS-BS links using only uplink training signals, which is in contrast to recent works [12], [13] on channel estimation with semi-passive elements that rely on both uplink and downlink training signals to estimate all the links. To compute the minimum mean squared error (MMSE) estimate of the channel from the received signal at the base station and additional information acquired from the active sensors at the IRS, we perform posterior inference on the channel under the sparse Bayesian learning (SBL) framework [14], [15]. Since exact posterior inference is intractable, we use the variational inference (VI) approach with the mean-field approximation [16], [17]. The proposed VI-SBL-based channel estimator enables us to compute the approximate MMSE estimates of all the links iteratively. In addition, we reduce the complexity of VI-SBL by converting a large matrix inversion to many small matrix inversions using the mean-field approximation. The simulation results show that VI-SBL outperforms the state-of-the-art channel estimators for IRS with passive reflecting elements in terms of the channel estimation error, training overhead, and energy efficiency, which is defined as the spectral efficiency normalized by the total power consumption to make a fair comparison between IRS with semi-passive elements and passive reflecting elements.

The rest of the paper is organized as follows. The channel model and signal model are introduced in Section II. In Section III, the proposed VI-SBL-based channel estimator is developed, which is followed by the mean-field approximation-based complexity reduction scheme. The performance of VI-SBL is assessed in Section IV based on various performance metrics, and a concluding remark follows in Section V.

**Notation:** $a$, $a_i$, and $A$ denote a scalar, vector, and matrix. The complex conjugate, transpose, and conjugate transpose of $A$ are written as $A^*$, $A^T$, and $A_H$. $[a]$ and $\angle a$ are the magnitude and phase of $a$. The $i$-th element of $a$ is $a_i$, while the $(i, j)$-th element and $i$-th column of $A$ are $A_{i,j}$ and $[A]_{i,:}$. The elementwise product, Kronecker product, and Khatri-Rao product of $A$ and $B$ are written as $A \odot B$, $A \otimes B$, and $kr(A, B)$. The $n \times 1$ all-zero vector, $n \times 1$ all-one vector, and $n \times n$ identity matrix are written as $0_n$, $1_n$, and $I_n$. $\text{vec}(A)$ is the vectorization of $A$, while $\text{diag}(a)$ is the diagonal matrix with $a$ on the main diagonal. The probability density function (PDF) of a complex Gaussian random vector $x \sim \mathcal{CN}(m, C)$ with mean $m$ and covariance $C$ is written as $\mathcal{CN}(x|m, C)$. The probability measure of a random vector $x$ is $p(x)$. The set difference between sets $A$ and $B$ is $A \setminus B$. $[N]$ denotes $\{1, \ldots, N\}$.

**II. System Model and Problem Formulation**

Consider the uplink of an IRS-aided mmWave massive MIMO system with an $M$-antenna base station and $K$ single-antenna users as illustrated in Fig. 1. The IRS is equipped with $N$ elements to compensate for the path loss in the mmWave band. In particular, $N_p$ passive reflecting elements reflect the impinging signal as usual [18], [19], [20]. Meanwhile, $N_A = N - N_p \ll N$ semi-passive elements act as active sensors in the channel estimation phase and passive reflecting elements in the data transmission phase. The active sensors are capable of receiving the signal, whose quantized version is forwarded to the base station. In practice, the active sensors are implemented by connecting the IRS elements to $N_r$ radio frequency (RF) chains with $B$-bit analog-to-digital converters (ADCs) [12], whose connection can be switched using base station-controlled switching networks [21], [22]. In this paper, a small number of active sensors and low-resolution ADCs...
A. Channel Model

The channel is composed of the UE-IRS link $\bar{f}_k \in \mathbb{C}^N$ for $k \in [K]$, IRS-BS link $\bar{G} \in \mathbb{C}^{M \times N}$, and UE-BS link $\bar{h}_k \in \mathbb{C}^M$ for $k \in [K]$. Since the scatterers are limited in the mmWave band as the path loss is severe, the links are typically modeled as [23]

$$
\bar{f}_k = \sqrt{\frac{N}{1 + \kappa_{UL,k}}} \alpha_{UL,k,0} a_{UL,k,0} (-1)^{\alpha_{UL,k,0}} (-1)^{\alpha_{UL,k,0}},
\bar{G} = N \sqrt{\frac{M}{1 + \kappa_{IB}}} \alpha_{IB,0} a_{IB,0} (-1)^{\alpha_{IB,0}} (-1)^{\alpha_{IB,0}},
\bar{h}_k = \sqrt{\frac{M}{1 + \kappa_{UB,k}}} \alpha_{UB,k,0} a_{UB,k,0} (-1)^{\alpha_{UB,k,0}} (-1)^{\alpha_{UB,k,0}},
$$

where $\alpha_{UL,k,0}, \alpha_{IB,0}, \alpha_{UB,k,0}$ are $\{-1, 1\}$-valued random variables. Let $\kappa_{UL,k}, \kappa_{IB}, \kappa_{UB,k}$ be the distance- and frequency-dependent path loss, $\theta_{UL,k}, \theta_{IB,0}, \theta_{UB,k}$ be the $\ell$-th azimuth/zenith angles of arrival/departure, $\mathbb{C}$ is the Rician $K$-factor, and $L$ is the number of non-LoS (NLoS) paths. In addition, $a_{UL,k,0} \in \mathbb{C}^M$ and $a_{IB,0} \in \mathbb{C}^N$ are the array response vectors of the base station and IRS.

B. Signal Model

Let $\Omega[t] \in \{0, 1\}^N$ and $\Omega'[t] = I_N - \Omega[t]$ denote the index vectors of the active sensors and passive reflecting elements that constitute the IRS at time $t$. In addition, define $\bar{F} = [\bar{f}_1, \ldots, \bar{f}_K]$ and $\bar{H} = [\bar{h}_1, \ldots, \bar{h}_K]$ for notational simplicity.

Then, the received signal $y[t] \in \mathbb{C}^M$ at the base station is

$$
y[t] = \bar{G} \left( \Omega'[t] \odot v[t] \odot \bar{F} x[t] \right) + \bar{H} x[t] + \bar{n}_B[t]
$$

where $v[t] \in \mathbb{C}^N$ is the passive reflection vector with the reflection amplitude $|v_n[t]| \leq 1$ and phase shift $\angle v_n[t] \in [0, 2\pi]$, $x[t] \in \mathbb{C}^K$ is the transmit signal of the users under the transmit power constraint $\mathbb{E} \{|x_k[t]|^2\} \leq P_k[t]$, and $\bar{n}_B[t] \sim \mathcal{CN}(0, \sigma_{B}^2 \mathbb{I}_N)$ is the additive white Gaussian noise (AWGN) at the base station. Likewise, the quantized received signal $z[t] \in \mathbb{C}^N$ at the active sensors forwarded to the base station is

$$
z[t] = \Omega[t] \odot Q(\bar{F} x[t] + \bar{n}_I[t])
$$

where $\bar{n}_I[t] \sim \mathcal{CN}(0, \sigma^2 \mathbb{I}_N)$ is the AWGN at the active sensors. The $B$-bit quantizer $Q(\cdot)$ is applied to both real and imaginary parts elementwise as

$$
z = \text{Q}(u) \iff \text{Re}(z) = \text{Q}(u) \quad \text{and} \quad \text{Im}(z) = \text{Q}(u)
$$

where $z \in \mathbb{C}$ and $z \in \mathbb{C}$ are the lower and upper thresholds associated with $z \in \mathbb{C}$. In other words, the real and imaginary parts of $z$, $z^\text{lo}$, and $z^\text{up}$ correspond to one of the $2^B$ quantization intervals.

A coherence block of length $T_c$ in the uplink consists of the channel estimation phase of length $T$ and data transmission phase of length $T_c - T$. To proceed, let $T_c = \lceil T \rceil$ and $T = \lfloor T \rfloor$, the channel estimation phase and data transmission phase. Then, the received signals $Y \in \mathbb{C}^{M \times T}$ and $Z \in \mathbb{C}^{N \times T}$ at the base station and active sensors over the channel estimation phase of length $T$ are expressed as

$$
Y = [y[1] \ldots y[T]] = \bar{G}(S \odot FX) + \bar{H}X + \bar{n}_B,
Z = [z[1] \ldots z[T]] = \Omega \odot Q(FX + N_I)
$$

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
using the notations $S = [s[1], \ldots, s[T]]$, $\Omega = [\Omega[1], \ldots, \Omega[T]]$, $X = [x[1], \ldots, x[T]]$, $N_B = [n_{B[1]}, \ldots, n_{B[T]}]$, and $N_I = [n_{I[1]}, \ldots, n_{I[T]}]$. In addition, the semi-passive elements are configured as

$$
||\Omega[t]||_0 = \begin{cases} N_a & \text{for } t \in T_c \\ 0 & \text{for } t \in T_d \end{cases},
$$

which means that $N_a$ semi-passive elements are activated to the active sensing mode in the channel estimation phase and deactivated to the passive reflecting mode in the data transmission phase. The switching performance of switching networks that connect the IRS elements to RF chains is determined by the switching period $T_{sn} \geq 1$ defined as the time required for the 0-1 pattern of $\Omega[t]$ to change such that

$$
\Omega[(i-1)T_{sn} + 1] = \cdots = \Omega[iT_{sn}] \text{ for } i \in \mathbb{N},
$$

and we define the switching frequency as $f_{sn} = 1/T_{sn} \leq 1$. Therefore, the 0-1 pattern of $\Omega[t]$ can change according to (7) and (8).

C. Problem Formulation via Virtual Channel Representation

We formulate the channel estimation problem using the virtual channel representation in conjunction with the channel model and signal model introduced in the previous subsections. To proceed, define the overcomplete dictionaries

$$
A_B = [a_B(\hat{\theta}_1) \cdots a_B(\hat{\theta}_{M_B})] \in \mathbb{C}^{M \times M_B},
$$

$$
A_I = [a_I(\hat{\theta}^A, \hat{\theta}^I_1) \cdots a_I(\hat{\theta}^A, \hat{\theta}^I_{N_B})] \in \mathbb{C}^{N \times N_B},
$$

over the predefined grids $\{\hat{\theta}_m\}_{m \in [M_B]}$ and $\{(\hat{\theta}^A, \hat{\theta}^I_n)\}_{n \in [N_B]}$, where $M_B \geq M$ and $N_B \geq N$ are the grid resolutions. Then, the virtual channel representation admits the transformations [24]

$$
\hat{F} = A_I F, \quad \hat{G} = A_B G A^H, \quad \hat{H} = A_B H
$$

where $F \in \mathbb{C}^{N \times K}$, $G \in \mathbb{C}^{M \times N_B}$, and $H \in \mathbb{C}^{M \times M_B}$ are the equivalent angular-domain channels. In practice, the angular-domain channels are sparse because there are limited scatterers that constitute the channels in the mmWave band [25], [26].

Now, we can reexpress (5) and (6) using the virtual channel representation as

$$
Y = A_B G A^H (S \odot A_I F X) + A_B H X + N_B,
$$

$$
Z = \Omega \odot Q(A_I F X + N_I),
$$

and our goal is to estimate $\{F, G, H\}$ from $\{Y, Z\}$. Since all the links are sparse, (14) and (15) is a combination of low-rank matrix factorization [27] and low-rank matrix completion [28] problems, which are NP-hard in general. To the best of our knowledge, our work is the first attempt to jointly exploit the received signal at the base station and additional information acquired from the active sensors at the IRS to estimate the UE-IRS and IRS-BS links. In contrast, recent works [12], [13] on channel estimation with semi-passive elements cannot be considered as semi-passive element-aided but rather driven because only the received signal at the semi-passive elements is leveraged. As a result, [12], [13] require both uplink and downlink training signals to estimate the UE-IRS and BS-IRS links, whereas we only use uplink training signals to estimate all the links.

III. PROPOSED VI-SBL-BASED CHANNEL ESTIMATOR

In this section, a VI-SBL-based channel estimator is proposed that performs approximate posterior inference on $\{F, G, H\}$ from $\{Y, Z\}$ under the SBL framework [14], [15]. In particular, we solve SBL via the variational free energy principle with the mean-field approximation [16], [17] to derive the posterior distributions of $\{F, G, H\}$. Then, we can compute the posterior means from the approximate posterior distributions, or equivalently the MMSE estimates that minimize the channel estimation error.

A. Pseudo-Measurement Model

To facilitate the analysis, we propose a pseudo-measurement model

$$
\hat{Z} = Q(U) = Q(\Omega \odot A_I F X + N_I)
$$

where the elements of $\hat{Z} \in \mathbb{C}^{N \times T}$ corresponding to the nonzero pattern of $\Omega$ are equal to the nonzero elements of $Z$. In addition, the lower and upper pseudo-thresholds $Z_l^o \in \mathbb{C}^{N \times T}$ and $Z_u^o \in \mathbb{C}^{N \times T}$ are associated with $\hat{Z}$ as defined in (4). The elements of $Z$ and $\hat{Z}$ corresponding to the zeros of $\Omega$, however, are not the same. Nevertheless, (15) and (16) are statistically equivalent because the elements of $Z$ and $\hat{Z}$ corresponding to the nonzero pattern of $\Omega$ bear the same information about $F$, while those corresponding to the zeros of $\Omega$ have no meaningful information about $F$ as evident from (15) and (16). Therefore, the elements of $Z$, $Z_l^o$, and $Z_u^o$ corresponding to the zeros of $\Omega$—the elements that cannot be determined from $Z$—can be assigned arbitrarily from the $2^B$ quantization intervals of the $B$-bit quantizer without loss of generality. Since the pseudo-measurement model is more convenient to deal with, we estimate $\{F, G, H\}$ from $\{Y, Z\}$ in (14) and (16) in the sequel.

B. Hierarchical Bayesian Model of SBL Framework

To account for the interaction among $\{F, G, H, U, Y, \hat{Z}\}$, we treat all the variables as random variables that constitute a hierarchical Bayesian model as shown in Fig. 2. In addition, we introduce the random variables $\Gamma_F \in \mathbb{C}^{N \times K}$, $\Gamma_G \in \mathbb{C}^{M \times N_B}$, and $\Gamma_H \in \mathbb{C}^{M \times M_B}$ to capture the sparse nature of $\{F, G, H\}$, which is the well-known SBL framework [14], [15]. Before moving on, we introduce the equivalent vector forms of the measurement $Y = \{y, z\}$, hidden variable $\mathcal{X} = \{f, g, h, \gamma_f, \gamma_g, \gamma_h, u\}$, and pseudo-threshold $\{\hat{z}_l^o, \hat{z}_u^o\}$ to facilitate the analysis. For example, $f = \text{vec}(F)$, $\hat{z}_l^o = \text{vec}(\hat{Z}_l^o)$, and $\hat{z}_u^o = \text{vec}(\hat{Z}_u^o)$.

In essence, the goal of SBL is to perform posterior inference on $\mathcal{X}$ from $Y$ where the interaction among $\{\mathcal{X}, Y\}$ is
captured by the conditional distributions listed as follows. First, \( \{f, g, h\} \) conditioned on \( \{\gamma_f, \gamma_g, \gamma_h\} \) are assumed to be Gaussian distributed as

\[
p(f|\gamma_f) = \mathcal{N}(f|0, \Gamma_f^{-1}), \tag{17}
\]

\[
p(g|\gamma_g) = \mathcal{N}(g|0, \Gamma_g^{-1}), \tag{18}
\]

\[
p(h|\gamma_h) = \mathcal{N}(h|0, \Gamma_h^{-1}), \tag{19}
\]

where \( \Gamma_f = \text{diag}(\gamma_f) \), \( \Gamma_g = \text{diag}(\gamma_g) \), and \( \Gamma_h = \text{diag}(\gamma_h) \) are the precision matrices of the Gaussian distributions above. Meanwhile, the hyperpriors of the hyperparameters \( \{\gamma_f, \gamma_g, \gamma_h\} \) are modeled as independent and identically distributed (i.i.d.) Gamma distributions

\[
p(\gamma_f) = \prod_i \text{Gamma}(\gamma_{f,i}|a, b)
\]

\[
= \prod_i \frac{b^a}{\Gamma(a)} \gamma_{f,i}^{a-1} e^{-b\gamma_{f,i}}, \tag{20}
\]

\[
p(\gamma_g) = \prod_i \text{Gamma}(\gamma_{g,i}|a, b)
\]

\[
= \prod_i \frac{b^a}{\Gamma(a)} \gamma_{g,i}^{a-1} e^{-b\gamma_{g,i}}, \tag{21}
\]

\[
p(\gamma_h) = \prod_i \text{Gamma}(\gamma_{h,i}|a, b)
\]

\[
= \prod_i \frac{b^a}{\Gamma(a)} \gamma_{h,i}^{a-1} e^{-b\gamma_{h,i}}, \tag{22}
\]

where \( \Gamma(\cdot) \) is the Gamma function, and \( a \) and \( b \) are the shape and rate to be chosen. The reason for choosing the Gaussian-Gamma distribution is because \( \{p(\gamma_f), p(\gamma_g), p(\gamma_h)\} \) are the conjugate priors for the likelihood functions \( \{p(f|\gamma_f), p(g|\gamma_g), p(h|\gamma_h)\} \) that make posterior inference tractable [29]. Furthermore, the Gaussian-Gamma distribution captures the sparse nature of \( \{f, g, h\} \) by making the marginal priors \( \{p(f), p(g), p(h)\} \) Student-\( t \)-distributions, which are sparsity-promoting priors under the appropriate choice of \( a \) and \( b \) [14], [15]. In the simulation, we adhere to the convention that assumes uninformative priors by setting \( a = b = 10^{-6} \).

Moving on to the measurement model, the conditional distribution of \( y \) is

\[
p(y|f, g, h)
\]

\[
= \mathcal{CN}(y|\text{vec}(A_BGA_f^H(S \circ A_fFX) + A_BHX), \sigma_y^2 I), \tag{23}
\]

which follows from (14). Likewise, the conditional distributions of \( u \) and \( \hat{z} \) in the pseudo-measurement model are

\[
p(u|f) = \mathcal{CN}(u|\text{vec}(\Omega \circ A_fFX), \sigma_u^2 I), \tag{24}
\]

\[
p(\hat{z}|u) = \mathbb{I} (\hat{z} = Q(u))
\]

\[
= \mathbb{I} (\text{Re}(\hat{z}^L) \leq \text{Re}(u) \ll \text{Re}(\hat{z}^{\text{up}})) \times \mathbb{I} (\text{Im}(\hat{z}^L) \leq \text{Im}(u) \ll \text{Im}(\hat{z}^{\text{up}})) \tag{25}
\]

where (24) and (25) come from (16) and (4). The indicator function \( \mathbb{I} (\cdot) \) is equal to one if the argument is true and zero otherwise.

Now, recall that the goal of SBL is to perform exact posterior inference on \( x \) from \( y \) using (17)-(25). The problem, however, is that computing \( p(x|y) = \int p(x, y) / p(y) dx \) is intractable in general. Therefore, we focus on finding the approximate posterior distribution using the VI approach.

### C. VI Approach to SBL

First, we explain the idea behind VI with the mean-field approximation [16], [17]. Then, we derive the approximate posterior distributions, which enable us to compute the posterior means. To proceed, consider the decomposition of the log-evidence

\[
\log p(y) = \int q(x) \log \frac{q(x)p(x|y)}{p(y)} dx + \int q(x) \log \frac{p(y|x)}{q(x)} dx \tag{26}
\]

where \( q(x) \) is any probability distribution that approximates \( p(x|y) \), \( D_{\text{KL}}(q||p) \) is the Kullback-Leibler (KL) divergence that measures the distance between \( q(x) \) and \( p(x|y) \), and \( \mathcal{L}(q) \) is known as the negative variational free energy. Then, we can minimize \( D_{\text{KL}}(q||p) \) with respect to \( q(x) \) by maximizing \( \mathcal{L}(q) \) because \( \log p(y) \) is constant.

In general, the variational free energy minimization problem is intractable for a general class of \( q(x) \). Therefore, we adopt the mean-field approximation by considering the class of \( q(x) \) that assumes independence among the partition of \( x \) such that \( q(x) = \prod_i q_i(x_i) \). Then, \( \{q_i(x_i)\}_{vi} \) is the global minimum of the variational free energy minimization problem if and only if [16], [17]

\[
q_i(x_i) = \frac{1}{Z_i} \exp \left\{ \mathbb{E}_{\prod_{j \neq i} q_j(x_j)} \log p(y|x)p(x) | x_i \right\} \tag{27}
\]

where \( Z_i \) is the normalization constant that makes \( q_i(x_i) \) a valid probability distribution. Here, \( \langle \cdot \rangle_{X_i} \) indicates the expectation with respect to \( \prod_{j \neq i} q_j(x_j) \). In addition, we use the notation \( \langle \cdot \rangle \) to indicate the expectation with respect to \( q(x_i) \) in the sequel.

Then, our goal is to derive the functional forms of \( \{q_i(x_i)\}_{vi} \) from (27) that enable us to deduce the distributions of \( \{q_i(x_i)\}_{vi} \) by inspection. Since the distributions of \( \{q_i(x_i)\}_{vi} \) are coupled in an intertwined manner as evident from (27), however, \( q_i(x_i) \) is updated for fixed \( \{q_j(x_j)\}_{j \neq i} \)
by cycling through \( i \), which leads us to a local minimum. Now, we derive the functional forms of \( \{ q(\lambda_i) \}_i \) to determine the approximate posterior distributions for the partition \( \lambda = \{ f, g, h, \gamma_f, \gamma_g, \gamma_h, u \} \).

1) Derivation of \( q(u) \): To derive \( q(u) \), we only need to plug in (24) and (25) to (27). The reason for not plugging in (17)-(23) to (27) is because the expectations of (17)-(23) with respect to \( \{ \lambda_i \} \) are constants that do not affect the functional form of \( q(u) \). In the posterior inference jargon, \( \{ f, \tilde{z} \} \) is said to be the Markov blanket of \( u \) [30]. In the sequel, we identify the Markov blanket by inspection without further explanation.

To proceed, define

\[
\mathbf{b}_{uf} = \text{vec}(\mathbf{\Omega} \otimes \mathbf{A}_1 \mathbf{F}) \\
= \text{diag}(\text{vec}(\mathbf{\Omega}))(\mathbf{X}^T \otimes \mathbf{A}_1)\mathbf{f}
\]  

(28)

from (24). Then, plugging in (24) and (25) to (27) yields

\[
q(u) \propto \exp\{\log p(y|f, g, h) + \log p(\mathbf{u}|f) + \log p(\mathbf{\gamma}_f)|f)\} \\
\times \mathcal{N}(\langle \mathbf{u}_0 \rangle | \mathbf{u}, \sigma_1^2 \mathbf{I})
\]

(29)

or equivalently

\[
q(\text{Re}(u_i)) \propto \mathcal{N}(\text{Re}(\tilde{z}_i^{\text{lo}}) \leq \text{Re}(u_i) \leq \text{Re}(\tilde{z}_i^{\text{up}})) \\
\]  

(30)

\[
q(\text{Im}(u_i)) = \mathcal{N}(\text{Im}(\tilde{b}_{uf,i}) \leq \text{Im}(u_i) \leq \text{Im}(\tilde{b}_{uf,i}))
\]  

(31)

where \( \phi(\cdot) \) and \( \Phi(\cdot) \) are the standard normal PDF and cumulative distribution function (CDF), \( \alpha_i = (\tilde{z}_i^{\text{lo}} - \langle \mathbf{b}_{uf} \rangle) / (\sigma_1 \sqrt{2}) \), and \( \beta_i = (\tilde{z}_i^{\text{up}} - \langle \mathbf{b}_{uf} \rangle) / (\sigma_1 \sqrt{2}) \). To compute the expressions above, we need \( \langle \mathbf{b}_{uf} \rangle \), which we can obtain from (28) by replacing \( \mathbf{f} \) with \( \langle \mathbf{f} \rangle \).

2) Derivation of \( q(f) \): First, define \( \mathbf{A}_{fg}, \mathbf{b}_{hn}, \) and \( \mathbf{A}_f \) from (23) and (24) in (32) and (33), shown at the bottom of the page. Then, plugging in (17), (23), and (24) to (27) with some straightforward but tedious algebra leads to

\[
q(f) \propto \exp\{\log p(y|f, g, h) + \log p(\mathbf{u}|f) + \log p(\mathbf{\gamma}_f)|f)\} \\
\times \mathcal{N}(\langle \mathbf{u} \rangle | \mathbf{u}, \sigma_2^2 \mathbf{I})
\]

(34)

which is a Gaussian distribution with

\[
\mathbf{m}_f = C_f \left( \frac{1}{\sigma_B^2} \langle \mathbf{A}_{fg} \rangle^\dagger (y - \langle \mathbf{b}_{hn} \rangle) + \frac{1}{\sigma_f^2} \mathbf{A}_f^\dagger \mathbf{A}_f + \langle \mathbf{T}_f \rangle \right)^{-1}
\]  

(35)

The posterior mean and covariance above require \( \langle \mathbf{A}_{fg} \rangle \), \( \langle \mathbf{b}_{hn} \rangle \), and \( \langle \mathbf{A}_{fg}^\dagger \mathbf{A}_{fg} \rangle \) to be computed, from which the first two can be obtained from (32) by substituting \( \mathbf{G} \) and \( \mathbf{h} \) with \( (\mathbf{G}) \) and \( (\mathbf{h}) \). The expression for \( \langle \mathbf{A}_{fg}^\dagger \mathbf{A}_{fg} \rangle \) is provided in Appendix A as the derivation is more involved.

3) Derivation of \( q(g) \): First, we define \( \mathbf{A}_{gf} \) and \( \mathbf{b}_{gh} \) from (23) in (37), shown at the bottom of the page. Then, expanding (27) after plugging in (18) and (23) gives

\[
q(g) \propto \exp\{\log p(y|f, g, h) + \log p(\mathbf{g}|\mathbf{\gamma}_g)\} \\
\times \mathcal{N}(\langle \mathbf{u} \rangle | \mathbf{u}, \sigma_2^2 \mathbf{I})
\]

(38)

which is recognized as a Gaussian distribution associated with the posterior mean and covariance

\[
\mathbf{m}_g = C_g \left( \frac{1}{\sigma_B^2} \langle \mathbf{A}_{gf} \rangle^\dagger (y - \langle \mathbf{b}_{gh} \rangle) \right)
\]  

(39)

\[
\mathbf{C}_g = \left( \frac{1}{\sigma_B^2} \langle \mathbf{A}_{gf} \rangle^\dagger \langle \mathbf{A}_{gf} \rangle + \langle \mathbf{T}_g \rangle \right)^{-1}
\]  

(40)
Again, computing the expressions above requires $\langle A_{gf} \rangle$ and $\langle b_{gh} \rangle$ as well as $\langle A_{gf}^H A_{gf} \rangle$, from which the terms corresponding to the first moment are given by (37) after replacing $F$ and $h$ with $\langle F \rangle$ and $\langle h \rangle$. The explicit form of $\langle A_{gf}^H A_{gf} \rangle$ with a detailed derivation is provided in Appendix B.

4) Derivation of $q(h)$: First, define $A_h$ and $b_{hfg}$ from (23) in (41), shown at the bottom of the previous page. By plugging in (19) and (23) to (27), we arrive at

$$q(h) \propto \exp\{\log p(y|f, g, h) + \log p(h|\gamma_h)\}_h$$

$$\propto \exp\{-\|y - b_{hfg} - A_h h\|^2_{\sigma B_h^2}\}_h$$

$$\times \exp\{-h^H \Gamma_h h\}_h$$

$$\propto \exp\{-h - m_h\}^H C_h^{-1}(h - m_h)$$,

which implies that $h$ is Gaussian distributed parameterized by

$$m_h = C_h \left( \frac{1}{\sigma B_h^2} A_h^T (y - \langle b_{hfg} \rangle) \right),$$

$$C_h = \left( \frac{1}{\sigma B_h^2} A_h^T A_h + \langle \Gamma_h \rangle \right)^{-1}.$$  

The posterior mean and covariance above require $\langle b_{hfg} \rangle$ to be computed, which can be obtained from (41) by substituting $F$ and $g$ with $\langle F \rangle$ and $\langle g \rangle$.

5) Derivation of $q(\gamma_f)$: Let us plug in (17) and (20) to (27) to obtain

$$q(\gamma_f) \propto \exp\{\log p(f|\gamma_f)\}_f + \log p(\gamma_f)\}_f$$

$$\propto \exp\{\log \det(\Gamma_f) - (\Gamma_f)^{-1} f_f + \log p(\gamma_f)\}$$

$$\propto \prod_i \gamma_{f,i} \times \exp\{-|f_i|^2 \gamma_{f,i} \} \times \gamma_{f,i}^{-1} \exp\{-b_{f,i} \gamma_{f,i} \}$$

$$\propto \prod_i \gamma_{f,i}^{-1} \exp\{-b + (|f_i|^2) \gamma_{f,i} \},$$

which is recognized as a product of Gamma distributions associated with the same posterior shape $\tilde{a} = a + 1$ and different rates $\tilde{b}_{f,i} = b + \langle |f_i|^2 \rangle = b + |C_f + m_f m_f^H|_{1,i}$. Therefore, the posterior mean of $\gamma_f$ is

$$\langle \gamma_{f,i} \rangle = \frac{\tilde{a}}{\tilde{b}_{f,i}}.$$  

6) Derivation of $q(\gamma_g)$: Since the priors $\{p(f|\gamma_f), p(g|\gamma_g)\}$ as well as the hyperpriors $\{p(\gamma_f), p(\gamma_g)\}$ are i.i.d., the derivation of $q(\gamma_g)$ leads to the same functional form, which is a product of Gamma distributions with the same posterior shape $\tilde{a} = a + 1$ and different rates $\tilde{b}_{g,i} = b + \langle |g_i|^2 \rangle = b + |C_g + m_g m_g^H|_{1,i}$. Therefore, the posterior mean of $\gamma_g$ is

$$\langle \gamma_{g,i} \rangle = \frac{\tilde{a}}{\tilde{b}_{g,i}}.$$  

7) Derivation of $q(\gamma_h)$: By the same argument as in the derivation of $q(\gamma_g)$, we conclude that $q(\gamma_h)$ is a product of Gamma distributions parameterized by the same posterior shape $\tilde{a} = a + 1$ and different rates $\tilde{b}_{h,i} = b + \langle |h_i|^2 \rangle = b + |C_h + m_h m_h^H|_{1,i}$, from which the posterior mean of $\gamma_h$ is deduced as

$$\langle \gamma_{h,i} \rangle = \frac{\tilde{a}}{\tilde{b}_{h,i}}.$$  

The VI-SBL-based channel estimator developed until now is summarized in Algorithm 1. In essence, Algorithm 1 performs approximate posterior inference by updating $q(f), q(g), q(h), q(\gamma_f), q(\gamma_g), q(\gamma_h)$, and $q(u)$ iteratively. The MMSE estimates of all the links are given by

$$\hat{F} = A_f \text{reshape}(m_f, N_g, K_f),$$

$$\hat{G} = A_g \text{reshape}(m_g, N_g, N_g, A_f^H),$$

$$\hat{H} = A_B \text{reshape}(m_h, M_g, K_f)$$

where $\text{reshape}(a, m, n)$ reshapes $a$ to a matrix of size $m \times n$ that preserves the columnwise ordering. To investigate the convergence of Algorithm 1, recall that Algorithm 1 tackles (27) by solving $q(X_i)$ for fixed $\{q(X_j)\}_{j \neq i}$ in a cycling manner instead of jointly solving for $\{q(X_i)\}_{i \neq i}$. Since (27) defines the global minimum of the variational free energy minimization problem [16], [17], each line in Algorithm 1 at least reduces the variational free energy. By noting that the variational free energy is lower-bounded by the negative log-evidence as evident from (26), we conclude that Algorithm 1 converges to a local minimum of the variational free energy problem.

The problem, however, is that poor initialization can result in a bad local minimum. The simulation results showed that a good initialization strategy is to perform “initial” posterior inference on a subset of the parameters in advance, and then use the coarse estimates obtained in this stage to initialize the parameters. The initial posterior inference we consider is the UE-RIS link estimation and UE-BS link estimation. In particular, the UE-RIS link initialization proceeds by performing initial posterior inference on $\{f, \gamma_f, u\}$ based on only the measurement $z$ to initialize $\{m_f, C_f, \langle \gamma_f \rangle, \langle u \rangle\}$ via Lines 8-10. For the UE-BS link initialization, we first turn off the passive reflecting elements for a short duration during the channel estimation phase, as shown in (57). Then, the UE-BS link initialization proceeds by performing initial posterior inference on $\{h, \gamma_h\}$ based on a subset of $y$ corresponding to the time slots the passive reflecting elements were turned off as in (57) to initialize $\{m_h, C_h, \langle \gamma_h \rangle\}$ via Lines 6-7. Since each initial posterior inference is equivalent to the conventional direct link channel estimation, i.e., UE-RIS link and UE-BS link, the initial posterior inference can be performed without much difficulty (even though the estimates obtained in this

---

**Algorithm 1 VI-SBL-Based Channel Estimator**

**Input:** $y, z$

**Output:** $m_f, m_g, m_h$, or equivalently $(f), (g), (h)$

1. Set the parameters $a$ and $b$ for the hyperpriors
2. Initialize $m_f, C_f, m_h, \langle \gamma_f \rangle, \langle \gamma_g \rangle, \langle \gamma_h \rangle$, and $q(u)$
3. while termination condition do
   4. Update $m_g$ and $C_g$ according to (39) and (40)
   5. Update $\langle \gamma_f \rangle$ according to (47)
   6. Update $m_h$ and $C_h$ according to (43) and (44)
   7. Update $\langle \gamma_h \rangle$ according to (48)
   8. Update $m_f$ and $C_f$ according to (35) and (36)
   9. Update $\langle \gamma_f \rangle$ according to (46)
   10. Update $(u)$ according to (30) and (31)
4. end while
stage are very coarse due to the extremely small size of the measurements used, thus using them only for initialization and preventing us from using them as final estimates). The remaining parameter \(\gamma_g\), which is not associated with neither the UE-IRS nor UE-BS link initialization, can be initialized as its prior mean \(\mathbb{E} [\gamma_g] = a/b \times 1\). By adopting the proposed initialization strategy, Algorithm 1 was demonstrated to produce good results.

D. Complexity Reduction via Mean-Field Approximation

The complexity of VI-SBL is mainly attributed to the matrix inversions in (36), (40), and (44) where the matrices to be inverted are of sizes \(N_x K_x N_y K_y\), \(M_x N_x\), and \(M_y K_y\). Therefore, the complexity of VI-SBL is \(O(N^3_x K^3 + M^3_x N^3_x + M^3_y K^3)\). To reduce the complexity of VI-SBL, a mean-field approximation-based solution is proposed that converts a large matrix inversion to many small matrix inversions.

In particular, we consider the mean-field approximation that splits \(\{f, \gamma_f\}, \{g, \gamma_g\}, \{h, \gamma_h\}\) to \(S_r, S_g\), and \(S_h\) subvectors. For example, the partition of \(\{f, \gamma_f\}\) can be written as \(f^T = [f^T_1, \ldots, f^T_{S_r}]\) and \(\gamma_f^T = [\gamma^T_f_1, \ldots, \gamma^T_f_{S_r}]\), where \(f^T_{[i]} \in \mathbb{C}^{N_x K} / S_r\) and \(\gamma^T_f_{[i]} \in \mathbb{C}^{N_x K} / С_r\), and the same logic holds for \(\{g, \gamma_g\}\) and \(\{h, \gamma_h\}\). Then, VI-SBL performs posterior inference on the partition

\[
q(X) = q(u) \times \prod_{i=1}^{S_r} q(f_{[i]}) q(\gamma_f_{[i]}) \prod_{j=1}^{S_g} q(g_{[j]}) q(\gamma_g_{[j]}) \prod_{k=1}^{S_h} q(h_{[k]}) q(\gamma_h_{[k]})
\]

from \(\{y, \hat{z}\}\).

Now, we present how the update rules in (35), (36), (39), (40), (43), and (44) are modified after the proposed mean-field approximation is applied. As a preliminary, let us introduce some shorthand notations. First, \(f^T_{[i]}\) is defined as the vector obtained by removing \(f_{[i]}\) from \(f\). Likewise, \(A^T_{[i]}\) and \(A^c_{[i]}\) are defined as the matrices obtained by retaining the columns of \(A_g\) in (32) corresponding to \(f_{[i]}\) and \(f^c_{[i]}\). The same logic can be extended to denote any vectors and matrices obtained by pruning. As an exception, the pruned versions of the precision matrices are denoted by \(G_f_{[i]} = \text{diag}(\gamma_f_{[i]})\), \(G_g_{[i]} = \text{diag}(\gamma_g_{[i]})\), and \(G_h_{[i]} = \text{diag}(\gamma_h_{[i]})\).

Then, the new update rules for (35), (36), (39), (40), (43), and (44) are given by in (53)–(55), shown at the bottom of the page. Since (53) repeats an \(N^2_x K^2 / S_r \times N^2_x K^2 / S_r\) matrix inversion \(S_r\) times, the overall complexity is \(O(N^3_x K^3 / S_r^2)\). The same logic applies to (54) and (55). As a result, the overall complexity of VI-SBL reduces to \(O(N^3_x K^3 / S_r^2 + M^3_x N^3_x / S^2_r + M^3_y K^3 / S^2_h)\).

IV. Simulation Results

In this section, we evaluate the performance of the proposed VI-SBL-based channel estimator based on the channel estimation error and energy efficiency. The baselines are state-of-the-art compressed sensing-based channel estimators for IRS-aided mmWave massive MIMO systems with passive reflecting elements. To implement the baselines, the UE-IRS-BS link estimation problem is reformulated as a sparse recovery problem as proposed in [10]. Then, the problem is solved using the generalized approximate MP (GAMP) [34], vector approximate MP (VAMP) [35], SBL [14], and generalized expectation consistent-signal recovery (GEC-SR) [36] algorithms, which are compressed sensing algorithms widely adopted in the mmWave channel estimation domain [32], [37], [38]. In addition, we also adopt the DS-OOMP channel estimator as our baseline, which is an algorithm that exploits the double sparsity structure of the UE-IRS-BS link by taking into account the fact that all the users share a common IRS-BS link. DS-OOMP attempts to improve the channel estimation accuracy by identifying the common rows and columns that the UE-IRS-BS links of all the users share.

The channel parameters in (1) are as follows. The bandwidth and path loss are \(W = 80\) MHz and

\[
PL = \begin{cases} 
31.4 + 20 \log_{10} d & \text{for LoS} \\
42 + 29.2 \log_{10} d & \text{for NLoS} 
\end{cases}
\]

in \(\text{dB}\) where \(d\) is the distance of the link in meters. The Rician K-factors are \(K_{UL} = K_{IB} = 13.2\) dB for the UE-IRS and IRS-BS links in LoS and \(K_{UB} = -\infty\) dB for the UE-BS link in NLoS. The number of NLoS paths is \(L_{UL} = L_{IB} = L_{UB} = 4\) for all the links.

In addition, the system parameters are configured as specified in the Dense Urban-eMBB scenario in ITU-R M.2412-0 [39]. In particular, the transmit power is \(P_k[t] = 23\) dBm. Meanwhile, the noise figure of the base station and active
sensors is NF = 7 dB. Therefore, the noise power is \( \sigma_n^2 = W \times N_0 \times NF \) where \( N_0 = -174 \) dBm/Hz is the noise spectral density. There are \( M = 16 \) antennas at the base station, \( N = 8 \times 8 \) elements at the IRS with \( N_a = 4 \) semi-passive elements, and \( K = 4 \) users. The parameters for the active sensors are \( B = 4 \) bit for the ADC resolution and \( f_s = 1 \) for the switching frequency. In addition, the uniform \( B \)-bit quantizer proposed in [37] and [40] is adopted. The passive reflecting elements are initially turned off for UE-BS link estimation and turned on as

\[
|v_n[t]| = \begin{cases} 0 & \text{for } t \in [50] \\ 1 & \text{for } t \in [T] \setminus [50] \end{cases}
\]

with random phase shifts in the channel estimation phase. The base station and IRS are at (0, 0) m and (20, 10) m, while the users are around the circle with center \((40, 0)\) m and radius 5 m. The lengths of the coherence block and channel estimation phase are \( T_c = 1800 \) and \( T = 400 \). Throughout the simulation, the system parameters are fixed unless stated otherwise. In particular, the parameters for VI-SBL are \( M_g = M, N_g = N, S_f = 1, S_g = 8, S_h = 1 \), and \( a = b = 10^{-6} \). To understand the reason behind \( S_g = 8 \), recall that the complexity of VI-SBL is \( O(N_a^2 K^3/S_g^2 + M_a^3 N_a^3/S_g^2 + M_a^2 K^3/S_h^2) \). Since massive MIMO and IRS imply large \( M_a \) and \( N_a \), most of the complexity comes from \( S_g \) that repeats an \( M_g N_g / S_g \times M_g N_g / S_g \) matrix inversion \( S_g \) times. Therefore, we convert a large matrix inversion to many small matrix inversions by setting \( S_g = 8 \).

Now, we explain the power consumption model that accounts for the power dissipated at the base station, IRS, and fronthaul link through which the received signal at the active sensors is forwarded. Since the base station is equipped with a local oscillator, \( M \) RF chains, and \( M \) pairs of \( B_{\infty} \)-bit ADCs where \( B_{\infty} \gg 1 \), the power consumed by the base station is modeled as \( P_{BS} = P_{LO} + M(P_{RF} + 2P_{ADC}(B_{\infty})) \). Here, \( P_{ADC}(B) = FOM \times f_s \times 2^B \) [44] is the power consumption of a \( B \)-bit ADC where FOM and \( f_s \) are the figure of merit and sampling frequency. Likewise, the power consumed by the active sensors at the IRS is modeled as \( P_{IRS} = P_{LO} + N_a(P_{RF} + 2P_{ADC}(B)) \). Moving on to the fronthaul link that forwards \( 2BN_a W \) bits per second, the power consumption model is \( P_{FH} = 2BN_a W P_T \) [41] where \( P_T \) is the traffic-dependent power consumption. Then, the total power consumption is \( P_{BS} + P_{FH} = 2BN_a W P_T \) [41] where the first case corresponds to VI-SBL with semi-passive elements, while the second case corresponds to the baselines with passive reflecting elements. Throughout the simulation, we mainly compare the NMSE of the UE-IRS-BS link because the baselines can only estimate the UE-IRS-BS and UE-BS links unlike VI-SBL that estimates all the links. Moreover, the reason for omitting the UE-BS link is because the UE-BS link estimation problem can be converted to the conventional channel estimation problem as pointed out in [3].

In the first simulation, we compare the performance of VI-SBL and the baselines in terms of the NMSE of the UE-IRS-BS link and energy efficiency for various transmit powers in the channel estimation phase. According to Fig. 3, VI-SBL outperforms the baselines in terms of the channel estimation accuracy. The superior performance of VI-SBL is expected
because VI-SBL exploits the additional information acquired from a small number of active sensors at the IRS, whereas only the received signal at the base station is available to the baselines. As a result, we see a significant energy efficiency gap from Fig. 4 due to the high channel estimation accuracy of VI-SBL that outweighs the additional power consumption.

In the second simulation, we evaluate the performance of various channel estimators for different $T$. According to Fig. 5, $T \geq 200$ is sufficient for VI-SBL to yield an accurate channel estimate, whereas the baselines require at least $T \geq 400$. As a result, the energy efficiency gap is large in the low training overhead regime as evident from Fig. 6. The reason for the dramatic training overhead gap comes from the fact that the baselines cannot observe the UE-IRS link, whereas VI-SBL has direct access to the UE-IRS link through a small number of active sensors. Therefore, the training overhead issue problematic in IRS-aided mmWave massive MIMO systems incurred by the large size of the UE-IRS-BS link can be resolved by adopting a small number of active sensors at the cost of additional but marginal power consumption.

In the third simulation, we evaluate the performance of VI-SBL in terms of the NMSEs of all the links and energy efficiency for various $N_a$. From Fig. 7, observe that the NMSE of the UE-IRS link decreases as $N_a$ increases, which is not as surprising. In contrast, the NMSE of the IRS-BS link increases unlike the UE-IRS link. The reason for such a phenomenon is because the IRS-BS link can only be observed through the reflected signal. Therefore, more active sensors means less passive reflecting elements, or equivalently less IRS-BS link measurements. In addition, another interesting point is that the NMSE of the UE-IRS-BS link composed of the UE-IRS and IRS-BS links is lower-bounded by the worst NMSE of the UE-IRS and IRS-BS links as evident from Fig. 7. From the discussion until now, we recommend to keep $N_a$ small because increasing $N_a$ results in additional power consumption, while degrading the quality of the UE-IRS-BS link estimate. By recalling that the impact of the UE-IRS-BS link on the spectral efficiency is significant, we conclude that increasing $N_a$ offers no benefit other than increasing the accuracy of the UE-IRS link estimate. In fact, the energy efficiency falls below the baselines as $N_a$ increases as evident from Fig. 8.

In the fourth simulation, we assess the performance of VI-SBL in terms of the NMSEs of all the links and energy efficiency.
Fig. 8. Energy efficiency vs. $N_a$. The baselines with passive reflecting elements are shown as references.

Fig. 9. NMSEs of all the links vs. $B$. The baselines are not shown.

efficiency for various $B$. According to Fig. 9, the NMSEs of all the links except the UE-IRS link saturate at $B \geq 4$. In contrast, the NMSE of the UE-IRS link hits the floor at $B \geq 12$. To understand such a phenomenon, recall that the UE-IRS link is observed through the quantized received signal at the active sensors. Then, we conclude that the impact of $B$ on the quality of the UE-IRS link estimate must be significant. Also, another interesting point is that the NMSE of the UE-IRS-BS link is lower-bounded by the worst NMSE of the UE-IRS and IRS-BS links as previously observed in Fig. 7. Moving on to Fig. 10, we see that there is a diminishing return beyond $B \geq 6$ at the cost of additional power consumption. Therefore, we can find the most energy-efficient operating point as a function of $B$, which is $B = 6$ in our simulation setup.

In the fifth simulation, we investigate the energy-spectral efficiency of VI-SBL as a function of $B$ for various $N_a$. According to Fig. 11, it is evident that increasing the ADC resolution beyond a certain limit, i.e., approximately $B \geq 4$ for any $N_a$, yields only a small spectral efficiency gain at the cost of large additional power consumption. Therefore, it is not worth it to deploy high-resolution ADCs at the active sensors of the IRS since the gain is marginal. In addition, an interesting observation is that increasing the number of active sensors at the IRS actually decreases the spectral efficiency. Therefore, the combined effect of reduced spectral efficiency and increased power consumption results in significantly degraded system performance. This is in line with Figs. 7 and 8, and this phenomenon can be attributed to the fact that increasing the number of active sensors decreases the amount of signals reflected at the IRS towards the base station, thus decreasing the IRS-BS link measurements. Therefore, we conclude that an energy-efficient architecture that attains high energy efficiency requires both the number of the active sensors and ADC resolution to be low.

Before moving on, we emphasize that the superior performance of VI-SBL with active sensors over the baselines with passive reflecting elements is due to the availability of the additional information acquired at the active sensors. That is, the simulation results until now do not imply that VI-SBL proposed in this paper is algorithmically superior over the existing compressed sensing algorithms. The main contribution of the proposed VI-SBL is that it suggests a systematic way of how to jointly process the pilots received at the base station and active sensors effectively to yield good channel estimation accuracy/energy efficiency. This is in
TABLE II
PER-ITERATION COMPLEXITIES OF VARIOUS CHANNEL ESTIMATORS

| Algorithm | Complexity |
|-----------|------------|
| Proposed  | \(O(N_e^2K^3/S_f^2 + M_g^2N_e^2S_g^2 + M_g^2K^3/S_h^2)\) |
| GAMP      | \(O(MT N_e N_g K)\) |
| VAMP      | \(O(MT N_e N_g K)\) |
| SBL       | \(O(M_g^2N_e^3K^3)\) |
| GEC-SR    | \(O(M_g^2N_e^3K^3)\) |
| DS-OMP    | \(O(MTK + NTKL_{1b}L_{1,k}^3)\) |

contrast to the recent works on IRS with semi-passive elements [12], [13] that only use the pilots received at the active sensors due to the complicated nature of jointly exploiting the pilots received at the base station and active sensors. The channel estimators in [12] and [13] indeed require a dedicated channel estimator protocol involving both uplink and downlink pilots, which make them to be only applicable to limited scenarios.

Next, we investigate the complexities of various channel estimators. According to Table II, SBL and GEC-SR that perform an \(M_gN_eK \times M_gN_eK\) matrix inversion have the highest complexities. In contrast, GAMP and VAMP have relatively low complexities that involve a matrix-vector multiplication of size \(MT \times M_gN_eK\), while DS-OMP has the lowest complexity slightly larger than that of the conventional OMP algorithm [47]. Meanwhile, VI-SBL performs many small matrix inversions, whose complexity is controlled by \(S_f\), \(S_g\), and \(S_h\) as discussed earlier. The complexity of VI-SBL is demanding but not as SBL and GEC-SR. Therefore, we conclude that VI-SBL offers a significant performance gain at the expense of not-so-low complexity.

V. CONCLUSION

A Bayesian channel estimator was proposed for IRS-aided mmWave massive MIMO systems with semi-passive elements. Unlike recent works on channel estimation with semi-passive elements that require both uplink and downlink signals to estimate the UE-IRS and IRS-BS links, the proposed channel estimator aimed to estimate all the links using only uplink training signals. To perform approximate posterior inference on the channel, the channel estimation problem was recast as an SBL framework. Then, SBL was solved using the variational free energy principle and mean-field approximation, from which the VI-SBL-based channel estimator was obtained. The simulation results showed that the proposed channel estimator requires low training overhead by taking advantage of the active sensors. Also, VI-SBL was capable of estimating all the links, which reduces the channel estimation overhead in the long run by replacing UE-IRS-BS link estimation as the IRS-BS link is quasi-static in practice.

APPENDIX A

DERIVATION OF \(\langle A_{gf}^H A_{gf} \rangle\) FOR UPDATING \(q(f)\)

As a preliminary, define \(B_f\) from

\[
A_{gf} = (I_T \otimes A_B G A_f^H) \text{diag} (\text{vec}(S)) (X^T \otimes A_1) = B_f
\]

in (32). Then, expanding \(\langle A_{gf}^H A_{gf} \rangle\) leads to

\[
\langle A_{gf}^H A_{gf} \rangle = \langle B_f^H (I_T \otimes A_1 G^H A_f^H) (I_T \otimes A_B G A_f^H) B_f \rangle
\]

\[
= \langle B_f^H (I_T \otimes A_1 G^H A_f^H) (I_T \otimes A_B G A_f^H) B_f \rangle
\]

where the mixed-product property of the Kronecker product was used in (a). Meanwhile, the vectorized version of \(G^H A_f^H A_B G\) in the middle of the last equality can be expressed as

\[
\text{vec} (\langle G^H A_f^H A_B G \rangle) = (G^T \otimes G^H) \text{vec} (A_f^H A_B).
\]

Now, let us focus on the conjugate transpose of \(G^T \otimes G^H\), or equivalently \(G^* \otimes G\), which contains \(M_g \times N_g\) submatrices of sizes \(M_g \times N_g\). The \((i,j)\)-th submatrix of \(G^* \otimes G\) can be vectorized as

\[
\text{vec} (g_{i+j-1} \otimes (j-1) N_g) = C_{fg} + m_{fg} m_{fg}^T
\]

which can be reshaped and assembled to compute \(A_{gf}^H A_{gf}\).

APPENDIX B

DERIVATION OF \(\langle A_{gf}^H A_{gf} \rangle\) FOR UPDATING \(q(g)\)

First, we expand \(\langle A_{gf}^H A_{gf} \rangle\) using (37) to obtain in (62), shown at the bottom of the page, where the mixed-product property of the Kronecker product was applied in (a). Now, let us focus on the transpose of \(\langle (S^* \otimes A_f^T F X^* ) (S^T \otimes X^T F A_f^T) \rangle\) in the middle of the last line of (62). To proceed, we define \(B_g\) from

\[
\text{vec}(S \otimes A_f^T F X^*) = \text{diag} (\text{vec}(S)) (X^T \otimes A_1) f = B_g
\]

Then, the \((i,j)\)-th element of \(\langle (S \otimes A_f^T F X^* ) (S^H \otimes X^H F A_f^H) \rangle\) is

\[
\langle (S \otimes A_f^T F X^* ) (S^H \otimes X^H F A_f^H) \rangle_{i,j}
\]

\[
= \sum_{k=1}^{M_g} [B_g \langle f f^H \rangle B_g^H]_{i+j-1} + (k-1) N_g, j \rangle + (1-1) N_g)
\]

\[
= \sum_{k=1}^{M_g} [B_g \langle C_f + m_f m_f^T \rangle B_g^H]_{i+j-1} + (k-1) N_g, j \rangle + (1-1) N_g)
\]

which can be assembled and rearranged to compute \(A_{gf}^H A_{gf}\).
In-Soo Kim (Member, IEEE) received the B.S. degree in electrical engineering from Hanyang University, Seoul, South Korea, in 2017, the M.S. degree in electrical engineering from the Pohang University of Science and Technology (POSTECH), Pohang, South Korea, in 2019, and the Ph.D. degree in electrical engineering from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2022. He is currently a Senior Systems Engineer. His research interests include transceiver architectures for low-power wireless communication systems with low-resolution ADCs. In particular, he develops channel estimators, data detectors, precoders, and resource allocation schemes for (but not limited to) hardware-constrained mmWave/cell-free/IRS-aided massive MIMO systems.

Mehdi Bennis (Fellow, IEEE) is currently a Full (tenured) Professor with the Centre for Wireless Communications, University of Oulu, Finland, and the Head of the Intelligent Connectivity and Networks/Systems Group (ICON). His main research interests include radio resource management, game theory, and distributed AI in 5G/6G networks. He has published more than 200 research papers in international conferences, journals, and book chapters. He has received several prestigious awards, including the 2015 Fred W. Ellersick Prize from the IEEE Communications Society, the 2017 EURASIP Best paper Award for the Journal of Wireless Communications and Networks, the all-University of Oulu Award for research, the 2019 IEEE ComSoc Radio Communications Committee Early Achievement Award, and the 2020 Clarivate Highly Cited Researcher with the Web of Science. He is an Editor of IEEE TRANSACTIONS ON COMMUNICATIONS and the Specialty Chief Editor for Data Science for Communications in the Frontiers in Communications and Networks journal.

Jaehoon Chung received the B.S. degree from Yonsei University in 1997 and the M.S.E.E. and Ph.D. degrees from the Korea Advanced Institute of Science and Technology (KAIST) in 1999 and 2005, respectively.

Junil Choi (Senior Member, IEEE) received the B.S. (Hons.) and M.S. degrees in electrical engineering from Seoul National University in 2005 and 2007, respectively, and the Ph.D. degree in electrical and computer engineering from Purdue University in 2015.

Junil Choi is currently an Ewon Associate Professor with the School of Electrical Engineering, KAIST. From 2007 to 2011, he was a member of Technical Staff with the Samsung Advanced Institute of Technology (SAIT) and Samsung Electronics Company Ltd., South Korea, where he contributed to advanced codebook and feedback framework designs for the 3GPP LTE/LTE-Advanced and IEEE 802.16m standards. Before joining KAIST, he was a Post-Doctoral Fellow with The University of Texas at Austin from 2015 to 2016 and an Assistant Professor with POSTECH from 2016 to 2019. His research interests include the design and analysis of massive MIMO, mmWave communication systems, distributed reception, and vehicular communication systems.

Dr. Choi was a co-recipient of the 2022 IEEE Vehicular Technology Society Best Vehicular Electronics Paper Award, the 2021 IEEE Vehicular Technology Society Neal Shepherd Memorial Best Propagation Award, the 2019 IEEE Communications Society Stephen O. Rice Prize, the 2015 IEEE Signal Processing Society Best Paper Award, and the 2013 Global Communications Conference (GLOBECOM) Signal Processing for Communications Symposium Best Paper Award. He was awarded the Michael and Katherine Birck Fellowship from Purdue University in 2011, the Korean Government Scholarship Program for Study Overseas (2011–2013), the Purdue University ECE Graduate Student Association (GSA) Outstanding Graduate Student Award in 2013, the Purdue College of Engineering Outstanding Student Research Award in 2014, the IEEE ComSoc AP Region Outstanding Young Researcher Award in 2017, the NSF Korea and Elsevier Young Researcher Award in 2018, the KICS Haedong Young Researcher Award in 2019, and the IEEE Communications Society Communication Theory Technical Committee Early Achievement Award in 2021. He is an Associate Editor of IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, and IEEE OPEN JOURNAL OF THE COMMUNICATIONS SOCIETY.

Jaeky Oh received the B.S. and M.S. degrees from Hanyang University, South Korea, in 2002 and 2005, respectively. He was a Wireless Communication Researcher with Hanyang University. He is currently working on next generation communication and advanced AI/ML research with the Information Communication Technology Center, CTO Division, LG Electronics, as a Principal Research Engineer. He has focused on applying advanced AI/ML technology to wireless environment and 6G.