Training for Channel Estimation in Nonlinear Multi-Antenna Transceivers

Kang Gao, J. Nicholas Laneman, N. J. Estes, Jonathan Chisum, Bertrand Hochwald

Department of Electrical Engineering, University of Notre Dame, Notre Dame, IN, 46556
Email: {kgao,jnl,nestes,jchisum,bhochwald}@nd.edu

***Abstract*—Recent efforts to obtain high data rates in wireless systems have focused on what can be achieved in systems that have nonlinear or coarsely quantized transceiver architectures. Estimating the channel in such a system is challenging because the nonlinearities distort the channel estimation process. It is therefore of interest to determine how much training is needed to estimate the channel sufficiently well so that the channel estimate can be used during data communication. We provide a way to determine how much training is needed by deriving a lower bound on the achievable rate in a training-based scheme that can be computed and analyzed even when the number of antennas is very large. This lower bound can be tight, especially at high SNR. One conclusion is that the optimal number of training symbols may paradoxically be smaller than the number of transmitters for systems with coarsely-quantized transceivers. We show how the training time can be strongly dependent on the number of receivers, and give an example where doubling the number of receivers reduces the training time by about 37 percent.**

I. Introduction

Nonlinear transceivers are being considered for high-frequency wireless communication because of their low cost and power advantages. Examples, include systems with low-resolution (especially one-bit) analog-to-digital converters (ADCs) at the receiver [1]–[7] or digital-to-analog converters (DACs) at the transmitter [8]–[11] or both [12]–[16]. Performance analysis has shown that high data rates with low error probability can be achieved with multiple antennas and channel state information (CSI) at the receiver [6].

Training-based schemes are often used in practice to obtain channel information, where part of coherence interval is used for training and the rest for data. In [17], a lower bound of the capacity in a training-based scheme for a linear system is provided, and the optimal training length is analyzed by maximize the lower bound. Simple analysis using a worst-case noise analysis allows optimum training rules to be derived.

In [3], [4], lower bounds on the channel capacity with training-based schemes are provided for systems with linear transmitters and one-bit quantized receivers. The authors formulate the quantized output as the combination of signal, Gaussian noise, and quantization noise uncorrelated with the signal, and provide a lower bound by following [17] and considering the worst-case additive noise that minimizes the input-output mutual information in a linear system at low SNR. We consider a more general model where either the transmitter or the receiver can have an arbitrary nonlinearity.

We consider a general system with nonlinear transceivers and provide a lower bound on the achievable rate (channel capacity). We do not attempt to approximate or linearize the transceiver architecture, but instead directly deal with the nonlinearity. A step-by-step method is provided to compute the lower bound. When we apply our bound to a system with linear transmitters and one-bit receivers, we can improve upon existing known training-based results. When we apply our bound to a system with one-bit transceivers, we find that the optimal number of training symbols can be smaller than the number of transmitters when we have more receivers than transmitters. We find that the optimal training length can decrease strongly as the number of receivers increases.

II. Training-based scheme and capacity lower bound

We assume a classical discrete-time block-fading channel [17], where the channel is constant for some discrete time interval T_b, after which it changes to an independent value that holds for another interval T_b, and so on. We divide the interval into two phases: $T_{\rm t}$ for training and $T_{\rm d}$ for data transmission, where $T_{\rm t} + T_{\rm d} = T_b$. For a general system with nonlinear transceivers, within one block of T_b symbols, the channel can be modeled by

$$Y_{\rm t} = f\left(\sqrt{\frac{\rho}{M}} H X_{\rm t} + V_{\rm t}\right), {\rm tr} X_{\rm t}^{\rm H} X_{\rm t} = M T_{\rm t}, \quad (1)$$

$$Y_{\rm d} = f\left(\sqrt{\frac{\rho}{M}} H X_{\rm d} + V_{\rm d}\right), \mathbb{E}{\rm tr} X_{\rm d}^{\rm H} X_{\rm d} = M T_{\rm d}, \quad (2)$$

where $X_{\rm t} \in \mathcal{X}^{M\times T_{\rm t}}$ and $X_{\rm d} \in \mathcal{X}^{M\times T_{\rm d}}$ are matrices of transmitted training signal and data signal, $Y_{\rm t} \in \mathcal{Y}^{N\times T_{\rm t}}$ and $Y_{\rm d} \in \mathcal{Y}^{N\times T_{\rm d}}$ are the corresponding matrices of

received signal, M and N are the number of transmitters and receivers, $\mathcal{X}$ and $\mathcal{Y}$ are the alphabet of the transmitted signal and received signal, H is a random channel matrix, which is fixed in one coherent time interval, $V_{\rm t}$ and $V_{\rm d}$ are additive noise. Elements of $V_{\rm t}, V_{\rm d}$ are independent and identically distributed (*iid*) complex Gaussian $\mathcal{CN}(0,\sigma^2)$. Elements of H are *iid* complex Gaussian $\mathcal{CN}(0,1)$. ρ is the expected signal power at the receiver. $f(\cdot)$ is an elementwise nonlinear function which models the nonlinearity of each receiver.

The capacity per transmitter per channel-use is

$$C_{\rm t} = \sup_{p_{X_{\rm d}}(\cdot), X_{\rm d}\in\mathcal{X}^{M\times T_{\rm d}}} \frac{1}{T_b M}\mathbf{I}(X_{\rm t}, Y_{\rm t}, Y_{\rm d}; X_{\rm d}),$$

where $\mathbf{I}(\cdot;\cdot)$ is the mutual information notation. In general, this optimization is difficult to compute, especially for large T, M, and N. We use a series of now-standard inequalities to obtain a tractable lower bound on this capacity. Let $\mathbf{x}_{\rm d}(k)$ and $\mathbf{y}_{\rm d}(k)$ be the kth column of $X_{\rm d}$ and $Y_{\rm d}$. A lower bound on $C_{\rm t}$ can be obtained by considering $\mathbf{x}_{\rm d}(k)$ to be *iid* with some distribution $p_{\mathbf{x}_{\rm d}}(\mathbf{x})$. Then,

$$\begin{aligned}
&T_b M C_{\rm t} \geq \mathbf{I}(X_{\rm d}; Y_{\rm d}|X_{\rm t}, Y_{\rm t})\\
&= \mathbf{H}(X_{\rm d}|X_{\rm t}, Y_{\rm t}) - \mathbf{H}(X_{\rm d}|X_{\rm t}, Y_{\rm t}, Y_{\rm d})\\
&= \sum_{k=1}^{T_{\rm d}} \mathbf{H}(\mathbf{x}_{\rm d}(k)|X_{\rm t}, Y_{\rm t}) - \sum_{k=1}^{T_{\rm d}} \mathbf{H}(\mathbf{x}_{\rm d}(k)|X_{\rm t}, Y_{\rm t}, Y_{\rm d},\\
&\qquad \mathbf{x}_{\rm d}(1), \cdots, \mathbf{x}_{\rm d}(k-1))\\
&\geq \sum_{k=1}^{T_{\rm d}} (\mathbf{H}(\mathbf{x}_{\rm d}(k)|X_{\rm t}, Y_{\rm t}) - \mathbf{H}(\mathbf{x}_{\rm d}(k)|X_{\rm t}, Y_{\rm t}, \mathbf{y}_{\rm d}(k)))\\
&= (T_b - T_{\rm t})\mathbf{I}(\mathbf{x}_{\rm d}; \mathbf{y}_{\rm d}|X_{\rm t}, Y_{\rm t}). \qquad (3)
\end{aligned}$$

Then, a lower bound on $C_{\rm t}$ is

$$C_{\rm t} \geq C_{\rm bound} = \max_{T_{\rm t}} \frac{T_b - T_{\rm t}}{T_b} R_{\rm eff}(T_{\rm t}), \qquad (4)$$

where

$$R_{\rm eff}(T_{\rm t}) = \frac{1}{M}\mathbf{I}(\mathbf{x}_{\rm d}; \mathbf{y}_{\rm d}|X_{\rm t}, Y_{\rm t}) \qquad (5)$$

is the effective achievable rate per transmitter in each channel-use. The corresponding training time that optimizes the lower bound is

$$T_{\rm t,opt} = \operatorname*{argmax}_{T_{\rm t}} \frac{T_b - T_{\rm t}}{T_b} R_{\rm eff}(T_{\rm t}). \qquad (6)$$

The complexity of finding $R_{\rm eff}(T_{\rm t})$ is lower than $C_{\rm t}$ since we do not need to find the optimizing $p_{X_{\rm d}}(\cdot)$. Nevertheless, $R_{\rm eff}(T_{\rm t})$ is still non-trivial to find since the amount of averaging needed to compute $\mathbf{I}(\mathbf{x}_{\rm d}; \mathbf{y}_{\rm d}|X_{\rm t}, Y_{\rm t})$ is generally exponential in M and N.

As a result, a tight approximation of $R_{\rm eff}(T_{\rm t})$ is needed that works for any M and N. Some efforts to make such an approximation include modeling nonlinear receivers as linear receivers with an equivalent extra noise, and then using a worst-case noise analysis to obtain a lower bound. For example, linear transmitters and coarsely quantized (one-bit) receivers are considered in [3] and [4], and a "Bussgang decomposition" is applied to find an equivalent uncorrelated quantization noise in a worst-case noise analysis. We do not employ such methods, but compare the lower bound we obtain with those obtained with such methods.

We approximate the bound $C_{\rm bound}$ with its limit when $M, N, T_b \to \infty$. We then use a novel equivalent channel, that can be analyzed using the so-called "replica method". The replica method, a tool used in statistical mechanics [18], has been applied in many communication system contexts [7], [19]–[23], neural networks [24], [25], error-correcting codes [26], and image restoration [27]. Although the replica method is used for a large scale limit, [16] and [28] also show that results obtained from the replica method provide a good approximation for small M and N (≈ 8). A mathematically rigorous justification of the replica method is elusive, but the success of the method maintains its popularity.

Because we look at the large M and N limit, we define

$$\alpha = \frac{N}{M}, \qquad \beta = \frac{T_b}{M}, \qquad (7)$$

and (4) and (6) become

$$\mathcal{C}_{\rm t} \geq \mathcal{C}_{\rm bound} = \max_{\beta_{\rm t}} \frac{\beta - \beta_{\rm t}}{\beta}\mathcal{R}_{\rm eff}(\beta_{\rm t}), \quad \beta_{\rm t} = \frac{T_{\rm t}}{T_b}, \quad (8)$$

$$\beta_{\rm t,opt} = \operatorname*{argmax}_{\beta_{\rm t}} \frac{\beta - \beta_{\rm t}}{\beta}\mathcal{R}_{\rm eff}(\beta_{\rm t}), \qquad (9)$$

where

$$\mathcal{R}_{\rm eff}(\beta_{\rm t}) = \lim_{M\to\infty} \frac{1}{M}\mathbf{I}(\mathbf{x}_{\rm d}; \mathbf{y}_{\rm d}|X_{\rm t}, Y_{\rm t}), \qquad (10)$$

$$\mathcal{C}_{\rm bound} = \lim_{M\to\infty} C_{\rm bound}, \quad \mathcal{C}_{\rm t} = \lim_{M\to\infty} C_{\rm t}.$$

We assume that the elements of the training matrix $X_{\rm t}$ are *iid* with 0 mean and unit variance. We find an equivalent channel that has the same achievable rate per transmitter as $\mathcal{R}_{\rm eff}$. The following claim summarizes the method we use.

Claim: Consider a nonlinear system with an unknown channel, with training and data phase as shown in (1) and (2). Let the elements of $X_{\rm t}$ be *iid* with zero mean and unit variance. Let the elements of $X_{\rm d}$ also be *iid* zero-mean unit-variance but not necessarily the same distribution as $X_{\rm t}$. As $M, N, T_{\rm t} \to \infty$, we have

$$\mathcal{R}_{\rm eff}(\beta_{\rm t}) = \lim_{M\to\infty} \frac{1}{M}\mathbf{I}(\mathbf{x}_{\rm d}; \hat{\mathbf{y}}_{\rm d}|\hat{H}), \qquad (11)$$

where $\mathcal{R}_{\rm eff}$ is defined in (10), $\mathbf{x}_{\rm d}, \hat{\mathbf{y}}_{\rm d}$ are the input and

output of a new channel

$$\hat{\mathbf{y}}_{\mathrm{d}} = f\left(\sqrt{\frac{\rho}{M}}\hat{H}\mathbf{x}_{\mathrm{d}} + \hat{\mathbf{v}}_{\mathrm{d}}\right), \tag{12}$$

where the channel $\hat{H}$ is known by the receiver, the elements of $\hat{H}$ are *iid* $\mathcal{CN}(0, q_{\mathrm{h}})$, and the elements of $\hat{\mathbf{v}}_{\mathrm{d}}$ are *iid* $\mathcal{CN}(0, \sigma^2 + \rho(1 - q_{\mathrm{h}}))$, and $1 - q_{\mathrm{h}}$ is the mean square error of the minimum mean square error (MMSE) estimator of H based on the training phase:

$$1 - q_{\mathrm{h}} = \lim_{M\to\infty} \frac{1}{MN}\mathbb{E}\left(\|H - \mathbb{E}(H|X_{\mathrm{t}}, Y_{\mathrm{t}})\|_{\mathrm{F}}^2\right), \tag{13}$$

where $\|\cdot\|_{\mathrm{F}}$ is the notation of Frobenius norm. The $\mathcal{R}_{\mathrm{eff}}$ depends on β_{t} through q_{h}.

The proof is shown in Appendix A and uses some common assumptions used for replica methods [7], [19], [20], [22]. We omit these here.

III. Step-by-step computation of lower bound

We partition the computational methods into small and large M, N, T_b. These "recipes" are then applied to a particular one-bit transceiver architecture in Section IV.

1) Derive $g(z, y, \sigma_0^2)$ as a function of z, y, σ_0^2 from

$$g(z, y, \sigma_0^2) = P(f(z + v) = y), \quad v \sim \mathcal{CN}(0, \sigma_0^2), \tag{14}$$

when $\mathcal{Y}$ is discrete, and from

$$g(z, y, \sigma_0^2) = \prod_{o\in\{\mathrm{R,I}\}} \frac{d}{dy}P([f(z + v)]_o \le [y]_o), \tag{15}$$

where $v \sim \mathcal{CN}(0, \sigma_0^2)$, $[\cdot]_R$ and $[\cdot]_I$ are the real and imaginary part of the enclosed value, when $\mathcal{Y}$ is continuous.

Steps 2A) to 7A) below compute C_{bound} directly, while steps 2B) to 10B) compute $\mathcal{C}_{\mathrm{bound}}$ defined in (8). The former works for small M and N when the computation complexity is affordable. Otherwise, the latter is appropriate. We now describe the former.

A. Small M, N, T_b

In the following steps, we need to run steps from 2A) to 6A) for $T_{\mathrm{t}} = 1, 2, \cdots, T_b$. Then, run step 7A) to compute C_{bound}.

2A) Derive

$$d_1(\mathbf{x}_{\mathrm{d}}, \mathbf{y}_{\mathrm{d}}, X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}}) = \prod_{k=1}^{N} \mathbb{E}_{\mathbf{h}}\Bigg[g\Big(\sqrt{\frac{\rho}{M}}\mathbf{h}^T\mathbf{x}_{\mathrm{d}}, y_{\mathrm{d},k}, \sigma^2\Big) \prod_{p=1}^{T_{\mathrm{t}}} g\left(\sqrt{\frac{\rho}{M}}\mathbf{h}^T\mathbf{x}_{\mathrm{t},p}, y_{\mathrm{t},kp}, \sigma^2\right)\Bigg],$$

as a function of $\mathbf{x}_{\mathrm{d}}, \mathbf{y}_{\mathrm{d}}, X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}}$, where $\mathbf{h} \sim \mathcal{CN}(0, I)$. If both $\mathcal{X}$ and $\mathcal{Y}$ are discrete, the numerical values can be computed and stored. Otherwise, simplify the expression if possible. This process of storing or simplification is the same for the following steps.

3A) Derive

$$d_2(X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}}) = \prod_{k=1}^{N} \mathbb{E}_{\mathbf{h}} \prod_{p=1}^{T_{\mathrm{t}}} g\left(\sqrt{\frac{\rho}{M}}\mathbf{h}^T\mathbf{x}_{\mathrm{t},p}, y_{\mathrm{t},kp}, \sigma^2\right)$$

as a function of $X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}}$, where $\mathbf{h} \sim \mathcal{CN}(0, I)$.

4A) Derive

$$d_3(\mathbf{x}_{\mathrm{d}}, \mathbf{y}_{\mathrm{d}}, X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}}) = \frac{d_1(\mathbf{x}_{\mathrm{d}}, \mathbf{y}_{\mathrm{d}}, X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}})}{d_2(X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}})} \tag{16}$$

as a function of $\mathbf{x}_{\mathrm{d}}, \mathbf{y}_{\mathrm{d}}, X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}}$.

5A) Derive

$$d_4(X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}}) = -\sum_{\mathbf{x}_{\mathrm{d}},\mathbf{y}_{\mathrm{d}}} \Bigg(p_{\mathbf{x}_{\mathrm{d}}}(\mathbf{x}_{\mathrm{d}}) d_3(\mathbf{x}_{\mathrm{d}}, \mathbf{y}_{\mathrm{d}}, X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}}) \ln \frac{\sum_{\mathbf{x}_{\mathrm{d}}} p_{\mathbf{x}_{\mathrm{d}}}(\mathbf{x}_{\mathrm{d}}) d_3(\mathbf{x}_{\mathrm{d}}, \mathbf{y}_{\mathrm{d}}, X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}})}{d_3(\mathbf{x}_{\mathrm{d}}, \mathbf{y}_{\mathrm{d}}, X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}})}\Bigg)$$

as a function of $X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}}$, where

$$p_{\mathbf{x}_{\mathrm{d}}}(\mathbf{x}_{\mathrm{d}}) = \prod_{k=1}^{M} p_{x_{\mathrm{d}}}(x_{\mathrm{d},k}),$$

$x_{\mathrm{d},k}$ is the kth element of $\mathbf{x}_{\mathrm{d}}$. Here we used summation, but if $\mathcal{X}, \mathcal{Y}$ is continuous, the summation should be replaced with integral. This also applies to the following steps.

6A) Compute $R_{\mathrm{eff}}(T_{\mathrm{t}})$ from

$$R_{\mathrm{eff}}(T_{\mathrm{t}}) = \frac{1}{M}\sum_{X_{\mathrm{t}},Y_{\mathrm{t}}} p_{X_{\mathrm{t}}}(X_{\mathrm{t}}) d_2(X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}}) d_4(X_{\mathrm{t}}, Y_{\mathrm{t}}, T_{\mathrm{t}}),$$

where

$$p_{X_{\mathrm{t}}}(X_{\mathrm{t}}) = \prod_{m=1}^{M}\prod_{p=0}^{T_{\mathrm{t}}} p_{x_{\mathrm{t}}}(x_{\mathrm{t},mp}),$$

$x_{\mathrm{t},mp}$ is the mth row and the pth column of X_{t}.

7A) Compute $T_{\mathrm{t,opt}}$ and C_{bound} from

$$T_{\mathrm{t,opt}} = \operatorname*{argmax}_{T_{\mathrm{t}}} \frac{T_b - T_{\mathrm{t}}}{T_b} R_{\mathrm{eff}}(T_{\mathrm{t}}).$$

$$C_{\mathrm{bound}} = \frac{T_b - T_{\mathrm{t,opt}}}{T_b} R_{\mathrm{eff}}(T_{\mathrm{t,opt}}),$$

Steps 2A) to 7A) come directly from standard information-theoretic arguments. Unfortunately the computational complexity increases very rapidly with M, N, T_b because all of the dimensions of the quantities involved grow. This technique therefore is limited to small values.

B. Large M, N, T_b

We use the Claim above to show how a bound C_{bound} defined in (4) may be computed. Discretize β_{t}

in the range $(0, \beta)$ to a target accuracy for optimization, denoted as $\beta_{\mathrm{t},k}, k = 1, 2, \cdots$. Compute steps from 2B) to 9B) for every k. Step 10B) then computes $\mathcal{C}_{\mathrm{bound}}$.

2B) Derive

$$\mathcal{G}_1(n, q) = -\beta_{\mathrm{t},k} \ln \left(\sum_{y \in \mathcal{Y}} \mathbb{E}_{\mathbf{u}} \prod_{a=0}^{n} g(\sqrt{\rho} u_a, y, \sigma^2) \right), \tag{17}$$

as a function of n, q in a form so that $\mathcal{G}_1(n, q)$ is differentiable at $n = 0$, where $\mathbf{u} = [u_0, u_1, \cdots, u_n]^T$, the real and imaginary part of $\mathbf{u}$ are $[\mathbf{u}]_{\mathrm{R}}$ and $[\mathbf{u}]_{\mathrm{I}}$, which are independent $\mathcal{N}(0, \frac{1}{2} Q_{\mathrm{h}})$. Q_{h} is a $(n+1) \times (n+1)$ matrix with 1 as diagonal elements and q as off-diagonal elements. This special structure of Q enables us to simplify the expression to be differentiable at $n = 0$ for any $g(\cdot, \cdot, \cdot)$ through a simple trick. The details of the trick can be found in our examples shown in next section. We consider $\mathcal{Y}$ as discrete here. If $\mathcal{Y}$ is continuous, we just need to change summation into integral. This also applies to the following steps.

3B) Derive

$$\mathcal{F}_1(q, \hat{q}) = \lim_{n \to 0} \frac{\partial}{\partial n} \mathcal{G}_1(n, q) + q\hat{q} + \ln(1 + \hat{q}) - \hat{q} \tag{18}$$

as a function of q and $\hat{q}$.

4B) Solve

$$\frac{\partial \mathcal{F}_1}{\partial q} = 0, \frac{\partial \mathcal{F}_1}{\partial \hat{q}} = 0, \tag{19}$$

and get solution $q = q_{\mathrm{h}}, \hat{q} = \hat{q}_{\mathrm{h}}$. If there are multiple solutions, select the solution $(q_{\mathrm{h}}, \hat{q}_{\mathrm{h}})$ that minimize $\mathcal{F}_1(q, \hat{q})$. We only require numerical solution and numerical methods such as Newton's method and secant method can be applied to solve such equation efficiently [29]. When we have expectation over random variables (often Gaussian) in the equation, we may need to Monte Carlo method. In many cases, only one or two Gaussian random variables is shown in the equation, which does not require high computation power to get an accurate solution. Now, with q_{h}, we are ready to derive $\mathcal{R}_{\mathrm{eff}}$ according to (11). q_{h} depends on $\beta_{\mathrm{t},k}$, and the in the rest steps, only q_{h} is needed.

5B) Let $\hat{\rho} = q_{\mathrm{h}} \rho, \hat{\sigma}^2 = \sigma^2 + (1 - q_{\mathrm{h}})\rho$. Derive

$$\mathcal{G}_2(n, r) = -\alpha \ln \left(\sum_{y \in \mathcal{Y}} \mathbb{E}_{\mathbf{w}} \prod_{a=0}^{n} g(\sqrt{\hat{\rho}} w_a, y, \hat{\sigma}^2) \right) \tag{20}$$

as a function of n and r in a form so that $\mathcal{G}_2(n, r)$ is differentiable at $n = 0$, where $\mathbf{w} = [w_0, w_1, \cdots, w_n]^T$, the real and imaginary part of $\mathbf{w}$ are $[\mathbf{w}]_{\mathrm{R}}$ and $[\mathbf{w}]_{\mathrm{I}}$, which are independent $\mathcal{N}(0, \frac{1}{2} Q_{\mathrm{x}})$, where Q_{x} is a $(n+1) \times (n+1)$ matrix with 1 as diagonal elements and r as off-diagonal elements. This step is very similar to 2B).

6B) Derive

$$\mathcal{G}_3(n, \hat{r}) = -\ln \mathbb{E}_{x_0, \cdots, x_n} \exp \Big(\sum_{0 \le a < b \le n} 2\hat{r} \cdot ([x_a]_{\mathrm{R}} [x_b]_{\mathrm{R}} + [x_a]_{\mathrm{I}} [x_b]_{\mathrm{I}}) \Big) \tag{21}$$

as a function of n and $\hat{r}$ in a form so that $\mathcal{G}_3(n, \hat{r})$ is differentiable at $n = 0$, where $[x_a]_{\mathrm{R}}$ and $[x_a]_{\mathrm{I}}$ are the real and imaginary part of x_a, and x_a are *iid* with distribution $p_{\mathrm{x}}(x_a)$. In many cases, the real and imaginary part of x_a are *iid* and it simplifies the expression by only considering the expectation on the real part. For continuous input (often Gaussian), the tricks we used on examples shown in the next section can be applied to make the expression differentiable at $n = 0$. For discrete input, the tricks used in [20] can be applied. We also used the trick in our example, but the details are omitted.

7B) Derive

$$\mathcal{F}_2(r, \hat{r}) = \lim_{n \to 0} \frac{\partial}{\partial n} \left(\mathcal{G}_2(n, r) + \mathcal{G}_3(n, \hat{r}) \right) + r\hat{r} \tag{22}$$

as a function of r and $\hat{r}$.

8B) Solve

$$\frac{\partial \mathcal{F}_2}{\partial r} = 0, \frac{\partial \mathcal{F}_2}{\partial \hat{r}} = 0, \tag{23}$$

and get solution $r = q_{\mathrm{x}}, \hat{r} = \hat{q}_{\mathrm{x}}$. If there are multiple solutions, select the solution $(q_{\mathrm{x}}, \hat{q}_{\mathrm{x}})$ that minimize $\mathcal{F}_2(r, \hat{r})$.

9B) Compute $\mathcal{R}_{\mathrm{eff}}(\beta_{\mathrm{t,k}})$ through

$$\mathcal{R}_{\mathrm{eff}}(\beta_{\mathrm{t},k}) = \frac{1}{\ln 2} \Big[\mathcal{F}_2(r, \hat{r}) + \alpha \sum_{y \in \mathcal{Y}} \mathbb{E}_u \Big(g(\sqrt{\hat{\rho}} u, y, \hat{\sigma}^2) \cdot \ln g(\sqrt{\hat{\rho}} u, y, \hat{\sigma}^2) \Big) \Big], u \sim \mathcal{CN}(0, 1) \tag{24}$$

Note that $\mathcal{R}_{\mathrm{eff}}(\beta_{\mathrm{t},k})$ depends on $\beta_{\mathrm{t},k}$ through q_{h}.

10B) Compute $\beta_{\mathrm{t,opt}}$ and $\mathcal{C}_{\mathrm{bound}}$ through

$$\beta_{\mathrm{t,opt}} = \underset{\beta_{\mathrm{t},k}}{\operatorname{argmax}} \frac{\beta - \beta_{\mathrm{t},k}}{\beta} \mathcal{R}_{\mathrm{eff}}(\beta_{\mathrm{t},k})$$

$$\mathcal{C}_{\mathrm{bound}} = \frac{\beta - \beta_{\mathrm{t,opt}}}{\beta} \mathcal{R}_{\mathrm{eff}}(\beta_{\mathrm{t,opt}}) \tag{25}$$

The proof is shown in Appendix B.

IV. Application of the Lower Bound

We consider two nonlinear systems as examples of the step-by-step methods presented in the previous section. We focus on large M and N, and compute $\mathcal{C}_{\mathrm{bound}}$. One system uses linear transmitters with $\mathcal{X}$ arbitrary complex, and one-bit receivers with $\mathcal{Y} = \{\pm 1 \pm j\}$. The other

system uses one-bit transmitters with $\mathcal{X} = \{\frac{\pm 1 \pm j}{\sqrt{2}}\}$ and one-bit receivers with $\mathcal{Y} = \{\pm 1 \pm j\}$. The nonlinear function is $f(z) = \text{sign}(z)$ for both cases, where the output of $\text{sign}(z)$ is a complex number with the sign of the real and imaginary parts of z as its real and imaginary parts. This model mimicks having a highly-nonlinear single-bit quantizer in the transceiver chain.

We let $\sigma^2 = 1$, and therefore ρ is the SNR at each receiver. For linear transmitters, we assume each element in $\mathbf{x}_\text{d}$ are *iid* $\mathcal{CN}(0,1)$, and for one-bit transmitters we assume each element in $\mathbf{x}_\text{d}$ are *iid* uniform distributed in $\{\frac{\pm 1 \pm j}{\sqrt{2}}\}$. With given M, N, T_t, ρ, we treat β_t as a variable and discretize it in increments of 0.1 for numerical accuracy.

For many of the steps, the distinction between linear transmitters and one-bit transmitters is not needed. In the steps where the distinction is important, we use the subscripts "L" and "O" to indicate "linear" or "one-bit" at the transmitter.

1) Derive $g(z, y, \sigma_0^2)$ according to (14), we have

$$g(z, y, \sigma_0^2) = \prod_{o \in \{\text{R},\text{I}\}} Q(-\frac{\sqrt{2}}{\sigma_0}[z]_o \cdot [y]_o), \quad (26)$$

where $[\cdot]_\text{R}$ and $[\cdot]_\text{I}$ are the real and imaginary part of the enclosed value.

2B) Let $\alpha = \frac{N}{M}$. Derive $\mathcal{G}_1(n, q)$ according to (17):

$$\begin{aligned}
&\mathcal{G}_1(n, q) \\
&= -\beta_\text{t} \ln \prod_{o \in \{\text{R},\text{I}\}} \sum_{y \in \{\pm 1\}} \mathbb{E}_{[\mathbf{u}]_o} \prod_{a=0}^{n} Q(\sqrt{2\rho}[u_a]_o y) \\
&= -2\beta_\text{t} \ln \left(2\mathbb{E}_\mathbf{z} \prod_{a=0}^{n} Q(\sqrt{\rho} z_a) \right),
\end{aligned}$$

where $\mathbf{z} = [z_0, \cdots, z_n]^T \sim \mathcal{N}(0, Q_\text{h})$, the diagonal elements of Q_h are 1 and the off-diagonal elements of Q_h are q.
We can consider $z_a = \sqrt{q}u + \sqrt{1-q}t_a$ with $u, t_0, t_1, \cdots, t_n$ to be *iid* $\mathcal{N}(0,1)$. Then, we have

$$\begin{aligned}
&\mathcal{G}_1(n, q) \\
&= -2\beta_\text{t} \ln 2\mathbb{E}_u[\mathbb{E}_{t_0} Q(\sqrt{\rho q}u + \sqrt{\rho(1-q)}t_0)]^{n+1} \\
&= -2\beta_\text{t} \ln \left(2\mathbb{E}_u \left[Q(\sqrt{\frac{\rho q}{\rho(1-q)+1}}u) \right]^{n+1} \right).
\end{aligned} \quad (27)$$

3B) Derive $\mathcal{F}_1(q, \hat{q})$ according to (18):

$$\begin{aligned}
\mathcal{F}_1(q, \hat{q}) = &-4\beta_\text{t}\mathbb{E}_u \Big[Q\left(\sqrt{\frac{\rho q}{\rho(1-q)+1}}u \right) \\
&\cdot \ln Q\left(\sqrt{\frac{\rho q}{\rho(1-q)+1}}u \right) \Big] + q\hat{q} + \ln(1+\hat{q}) - \hat{q}.
\end{aligned}$$

4B) Solve q_h according to (19):

By solving (19), we eventually get that q_h is the solution of

$$\frac{q}{1-q} = \frac{\beta_\text{t} B^2}{\pi} \mathbb{E}_u \left(\frac{\exp(-B^2 q u^2)}{Q(B\sqrt{q}u)} \right), \quad (28)$$

where $B = \sqrt{\frac{\rho}{1+\rho(1-q)}}, u \sim \mathcal{N}(0,1)$. The solution q_h depends on β_t, and in the rest steps, only q_h is needed.

5B) Let $\hat{\rho} = q_\text{h}\rho, \hat{\sigma}^2 = 1 + (1 - q_\text{h})\rho$. Derive $\mathcal{G}_2(n, r)$ according to (20):
$\mathcal{G}_2(n, r)$ is very similar to $\mathcal{G}_1(n, q)$. Similar to (27), we have

$$\begin{aligned}
&\mathcal{G}_2(n, r) \\
&= -2\alpha \ln \left(2\mathbb{E}_u \left[Q(\sqrt{\frac{\rho_\text{eff} r}{\rho_\text{eff}(1-r)+1}}u) \right]^{n+1} \right),
\end{aligned} \quad (29)$$

where

$$\rho_\text{eff} = \frac{\rho q_\text{h}}{1 + \rho(1 - q_\text{h})}. \quad (30)$$

6B) Derive $\mathcal{G}_3(n, \hat{r})$ according to (21):
For linear transmitter, we consider each element of $\mathbf{x}_\text{d}$ are i.i.d $\mathcal{CN}(0,1)$.

$$\mathcal{G}_{3,\text{L}}(n, \hat{r}) = -2 \ln \mathbb{E}_\mathbf{w} \exp(\sum_{a<b} \hat{r} w_a w_b),$$

where $\mathbf{w} = [w_0, w_1, \cdots, w_n]^T \sim \mathcal{N}(0, I)$.
Since

$$\begin{aligned}
&\mathbb{E}_\mathbf{w} \exp(\sum_{a<b} \hat{r} w_a w_b) = \int_{\mathcal{R}^{n+1}} \frac{d\mathbf{w}}{(2\pi)^{\frac{n+1}{2}}} \exp(-\frac{1}{2}\mathbf{w}^T D \mathbf{w}) \\
&= |D|^{-\frac{1}{2}},
\end{aligned}$$

where the diagonal elements of D are 1 and off-diagonal elements are $-\hat{r}$. Therefore, we have

$$\mathcal{G}_{3,\text{L}}(n, \hat{r}) = (n+1)\ln(1+\hat{r}) + \ln\left(1 - \frac{(n+1)\hat{r}}{1+\hat{r}}\right).$$

For one-bit transmitter, we consider each element of x_d are *iid* uniform among $\{\frac{\pm 1 \pm j}{\sqrt{2}}\}$.

$$\mathcal{G}_{3,\text{O}}(n, \hat{r}) = -2 \ln \mathbb{E}_\mathbf{m} \exp(\sum_{a<b} \hat{r} m_a m_b),$$

where $\mathbf{m} = [m_0, \cdots, m_n]^T$ with each element *iid* uniform in $\{\pm 1\}$. According to [20], we have

$$\mathcal{G}_{3,\text{O}}(n, \hat{r}) = n\hat{r} - 2 \ln \mathbb{E}_u \cosh^n(\hat{r} + \sqrt{\hat{r}}u), \quad (31)$$

where $u \sim \mathcal{N}(0,1)$.

7B) Derive $\mathcal{F}_2(r, \hat{r})$ according to (22):
For linear transmitters, we have

$$\begin{aligned}
\mathcal{F}_{2,\text{L}}(r, \hat{r}) = &-4\alpha\mathbb{E}_u Q(A\sqrt{r}u) \ln Q(A\sqrt{r}u) \\
&+ \ln(1+\hat{r}) - \hat{r} + r\hat{r},
\end{aligned}$$

where

$$A = \sqrt{\frac{\rho_{\text{eff}}}{\rho_{\text{eff}}(1-r)+1}}. \tag{32}$$

For one-bit transmitters, we have

$$\mathcal{F}_{2,\text{O}}(r,\hat{r}) = -4\alpha\mathbb{E}_u Q(A\sqrt{r}u)\ln Q(A\sqrt{r}u) + \hat{r} - 2\mathbb{E}_u \ln\cosh(\hat{r}+\sqrt{\hat{r}}u) + r\hat{r}.$$

8B) Solve for $(q_{\text{x}}, \hat{q}_{\text{x}})$ according to (23):
For linear transmitters, we get that $(q_{\text{x,L}}, \hat{q}_{\text{x,L}})$ are the solution of

$$\hat{q}_x = \frac{\alpha A^2}{\pi}\mathbb{E}_u \frac{\exp\left(-A^2 q_x u^2\right)}{Q(A\sqrt{q_x}u)}, \tag{33}$$

$$q_x = \frac{\hat{q}_x}{1+\hat{q}_x}, A = \sqrt{\frac{\rho_{\text{eff}}}{1+\rho_{\text{eff}}(1-q_x)}}. \tag{34}$$

For one-bit transmitters, we get that $(q_{\text{x,O}}, \hat{q}_{\text{x,O}})$ are the solution of

$$\hat{q}_{\text{x}} = \frac{\alpha A^2}{\pi}\mathbb{E}_u \frac{\exp\left(-A^2 q_{\text{x}} u^2\right)}{Q(A\sqrt{q_{\text{x}}}u)}, \tag{35}$$

$$q_{\text{x}} + 1 = \mathbb{E}_u \tanh(\sqrt{\hat{q}_{\text{x}}}u + \hat{q}_{\text{x}})(2 + \frac{u}{\sqrt{\hat{q}_{\text{x}}}}), \tag{36}$$

$$A = \sqrt{\frac{\rho_{\text{eff}}}{1+\rho_{\text{eff}}(1-q_{\text{x}})}}.$$

9B) Compute $\mathcal{R}_{\text{eff}}(\beta_{\text{t,k}})$ according to (24):
For linear transmitters, we have

$$\mathcal{R}_{\text{eff,L}}(\beta_{\text{t,k}}) = \frac{1}{\ln 2}\Big[\mathcal{F}_{2,\text{L}}(q_{\text{x,L}}, \hat{q}_{\text{x,L}}) + 4\alpha\mathbb{E}_u Q(\sqrt{\rho_{\text{eff}}}u)\ln Q(\sqrt{\rho_{\text{eff}}}u)\Big], u \sim \mathcal{N}(0,1).$$

For one-bit transmitters, we have

$$\mathcal{R}_{\text{eff,O}}(\beta_{\text{t,k}}) = \frac{1}{\ln 2}\Big[\mathcal{F}_{2,\text{O}}(q_{\text{x,O}}, \hat{q}_{\text{x,O}}) + 4\alpha\mathbb{E}_u Q(\sqrt{\rho_{\text{eff}}}u)\cdot\ln Q(\sqrt{\rho_{\text{eff}}}u)\Big], u \sim \mathcal{N}(0,1). \tag{37}$$

This result matches that in [16].

10B) Compute $\mathcal{C}_{\text{bound}}$ and $\beta_{\text{t,opt}}$
For linear transmitters, we have

$$\beta_{\text{t,opt,L}} = \underset{\beta_{\text{t},k}}{\text{argmax}} \frac{\beta-\beta_{\text{t},k}}{\beta}\mathcal{R}_{\text{eff,L}}(\beta_{\text{t},k}). \tag{38}$$

$$\mathcal{C}_{\text{bound,L}} = \frac{\beta-\beta_{\text{t,opt,L}}}{\beta}\mathcal{R}_{\text{eff,L}}(\beta_{\text{t,opt,L}}). \tag{39}$$

For one-bit transmitters, we have

$$\beta_{\text{t,opt,O}} = \underset{\beta_{\text{t},k}}{\text{argmax}} \frac{\beta-\beta_{\text{t},k}}{\beta}\mathcal{R}_{\text{eff,O}}(\beta_{\text{t},k}). \tag{40}$$

$$\mathcal{C}_{\text{bound,O}} = \frac{\beta-\beta_{\text{t,opt,O}}}{\beta}\mathcal{R}_{\text{eff,O}}(\beta_{\text{t,opt,O}}), \tag{41}$$

V. Numerical Results

We provide some numerical results based on the expressions we obtained for two types of nonlinear systems in previous section. We select an accuracy of 0.1 in our optimization over β_{t}.

A. Linear transmitters and one-bit receivers

For a system with linear transmitters and one-bit receivers, a training-based capacity lower bound using the Bussgang theorem is derived in [3], [4]. LMMSE channel estimation is used and the channel estimation error part is approximated at low SNR to obtain a closed-form solution. However, with our equivalent channel (12), the Bussgang decomposition can be applied to get a simplified solution for any SNR. In our equivalent channel model, MMSE channel estimation is used, which provides better performance than LMMSE channel estimation in general. The lower bound thereby obtained is

$$\mathcal{C}_{\text{t}} \geq \mathcal{C}_{\text{Bussgang}} = \max_{\beta_{\text{t}}} \frac{\beta-\beta_{\text{t}}}{\beta}\log_2\left(1+\frac{2\alpha\rho_{\text{eff}}}{\pi(1+\rho_{\text{eff}})}\right), \tag{42}$$

where ρ_{eff} is defined in (30).

For comparison, we also consider the achievable rate when the receiver knows the channel and each element of $\mathbf{x}_{\text{d}}$ are *iid* $\mathcal{CN}(0,1)$, denoted as $\mathcal{R}_{\text{CSIR}}$. $\mathcal{R}_{\text{CSIR}}$ can be directly derived using steps from 5B) to 9B) with $q_{\text{h}} = 1$. And the solution of $\mathcal{R}_{\text{CSIR}}$ is

$$\mathcal{R}_{\text{CSIR}} = \frac{1}{\ln 2}\big[4\alpha\mathbb{E}_u\big(Q(\sqrt{\rho}u)\ln Q(\sqrt{\rho}u) - Q(\hat{A}\sqrt{q_{\text{x}}}u)\ln Q(\hat{A}\sqrt{q_{\text{x}}}u)\big) + \ln(1+\hat{q}_{\text{x}}) - \hat{q}_{\text{x}} + q_{\text{x}}\hat{q}_{\text{x}}\big], \tag{43}$$

where $\hat{A}, q_{\text{x}}, \hat{q}_{\text{x}}$ are the solution of

$$\hat{q}_x = \frac{\alpha\hat{A}^2}{\pi}\mathbb{E}_u \frac{\exp\left(-\hat{A}^2 q_x u^2\right)}{Q(\hat{A}\sqrt{q_x}u)}, \tag{44}$$

$$q_x = \frac{\hat{q}_x}{1+\hat{q}_x}, \hat{A} = \sqrt{\frac{\rho}{1+\rho(1-q_x)}}. \tag{45}$$

The comparison between $\mathcal{C}_{\text{Bussgang}}, \mathcal{C}_{\text{bound,L}}$ and $\mathcal{R}_{\text{CSIR}}$ is shown in Fig. 1. We observe that $\mathcal{C}_{\text{Bussgang}}$ is generally smaller (less tight as a lower bound) than $\mathcal{C}_{\text{bound,L}}$.

When $\rho_{\text{eff}} \to 0$, which can be caused by small β_{t} or small ρ, $\mathcal{C}_{\text{Bussgang}}$ and $\mathcal{C}_{\text{bound,L}}$ meet. From (33) and (34), we have

$$q_{\text{x,L}} = \frac{2\alpha}{\pi}\rho_{\text{eff}} + o(\rho_{\text{eff}}), \hat{q}_{\text{x,L}} = \frac{2\alpha}{\pi}\rho_{\text{eff}} + o(\rho_{\text{eff}}). \tag{46}$$

Then, (39) becomes

$$\mathcal{C}_{\text{bound,L}} = \max_{\beta_{\text{t}}} \frac{\beta-\beta_{\text{t}}}{\beta}\frac{2\alpha}{\pi\ln 2}\rho_{\text{eff}} + o(\rho_{\text{eff}}), \tag{47}$$

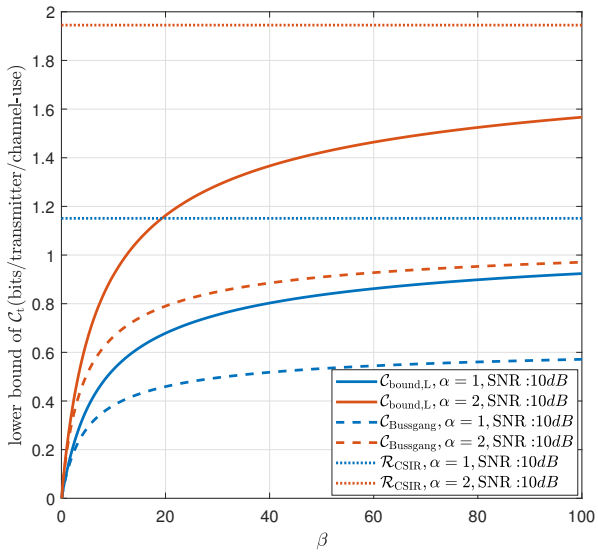

Fig. 1. The comparison between $\mathcal{C}_{\text{bound,L}}$ (39) and $\mathcal{C}_{\text{Bussgang}}$ (42) at $\alpha = 1, 2$ with different β for systems with linear transmitters and one-bit receivers. Both $\mathcal{C}_{\text{Bussgang}}$ and $\mathcal{C}_{\text{bound,L}}$ are maximized over β_{t}. $\mathcal{R}_{\text{CSIR}}$ (43) is the achievable rate when the receiver knows the channel and the elements in $\mathbf{x}_{\text{d}}$ are *iid* $\mathcal{CN}(0, 1)$.

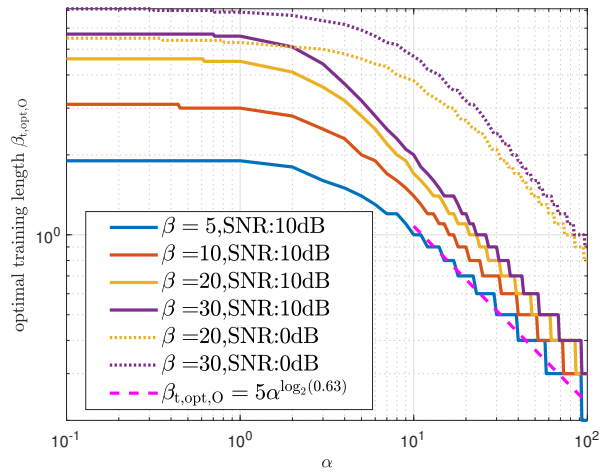

Fig. 2. The optimal training time $\beta_{\text{t,opt,O}}$ (40) versus $\alpha = \frac{N}{M}$ in a wide range. We can see that it's possible to have $\beta_{\text{t,opt,O}} < 1$. Both α and $\beta_{\text{t,opt,O}}$ are plotted in a log domain. When α is large, $\beta_{\text{t,opt,O}}$ decreases by 37 percent when we double α.

which is the same as $\mathcal{C}_{\text{Bussgang}}$. According to (28), for small SNR with $\rho \to 0$, we have

$$q_{\text{h}} = \frac{2\beta_{\text{t}}\rho}{\pi} + o(\rho). \tag{48}$$

Therefore, (30) becomes

$$\rho_{\text{eff}} = \frac{2\beta_{\text{t}}\rho^2}{\pi} + o(\rho^2). \tag{49}$$

And we have

$$\beta_{\text{t,opt,L}} \approx \frac{\beta}{2}, \quad \mathcal{C}_{\text{bound,L}} \approx \frac{\alpha\beta}{\pi^2 \ln 2}\rho^2,$$

when $\rho \to 0$. Hence, half of T_b is used for training at low SNR with linear transmitters and one-bit receivers.

B. One-bit transmitters and one-bit receivers

For systems with one-bit transceivers, we show that it is possible to have optimal training time smaller than the number of transmitter ($\beta_{\text{t,opt}} < 1$) even when the total time is larger than twice of the number of transmitters ($\beta > 2$). This result is shown in Fig. 2, where we show the relationship between $\beta_{\text{t,opt,O}}$ and α with different β. We see that when α is large, $\beta_{\text{t,opt,O}}$ can be smaller than 1 and $\beta_{\text{t,opt,O}}$ decreases approximately 37 percent when we double α as α gets large. When α is small, according to [16],

$$\mathcal{R}_{\text{eff,O}} \approx \alpha c(\rho_{\text{eff}}), \tag{50}$$

where $c(\rho)$ is the capacity of a single pair of one-bit transceivers in a Rayleigh channel with channel information at the receiver at SNR ρ, defined as

$$c(\rho) = 2(1 - \mathbb{E}_z\left(\mathcal{H}_2(Q(\sqrt{\rho}z))\right)), z \sim \mathcal{N}(0, 1), \tag{51}$$

where $\mathcal{H}_2(p) = -(p \log_2 p + (1-p)\log_2(1-p))$ is the binary entropy function.

The $\beta_{\text{t,opt,O}}$ obtained for small α maximize $\frac{\beta-\beta_{\text{t}}}{\beta}c(\rho_{\text{eff}})$, where ρ_{eff} is solved from (30). The corresponding bound $\mathcal{C}_{\text{bound,O}}$ is shown in Fig. 3. Because of the one-bit quantization at the transmitter, $\mathcal{C}_{\text{bound,O}}$ saturates at $\mathcal{C}_{\text{bound,O}} = 2$ as α gets larger. With many more receivers than the transmitters (α very large), saturation is achieved with a small number of training symbols, and hence $\beta_{\text{t,opt,O}}$ can be smaller than 1.

At low SNR, $\rho_{\text{eff}} \to 0$, according to (35) and (36), we have

$$q_{\text{x,O}} = \frac{2\alpha}{\pi}\rho_{\text{eff}} + o(\rho_{\text{eff}}), \hat{q}_{\text{x,O}} = \frac{2\alpha}{\pi}\rho_{\text{eff}} + o(\rho_{\text{eff}}), \tag{52}$$

and (41) becomes

$$\mathcal{C}_{\text{bound,O}} = \max_{\beta_{\text{t}}} \frac{\beta - \beta_{\text{t}}}{\beta} \frac{2\alpha}{\pi \ln 2}\rho_{\text{eff}} + o(\rho_{\text{eff}}). \tag{53}$$

According to (49), we have

$$\beta_{\text{t,opt,O}} \approx \frac{\beta}{2}, \quad \mathcal{C}_{\text{bound,O}} \approx \frac{\alpha\beta}{\pi^2 \ln 2}\rho^2.$$

Thus, as with linear transmitters, half of T_b is used for training at low SNR with one-bit transmitters and receivers.

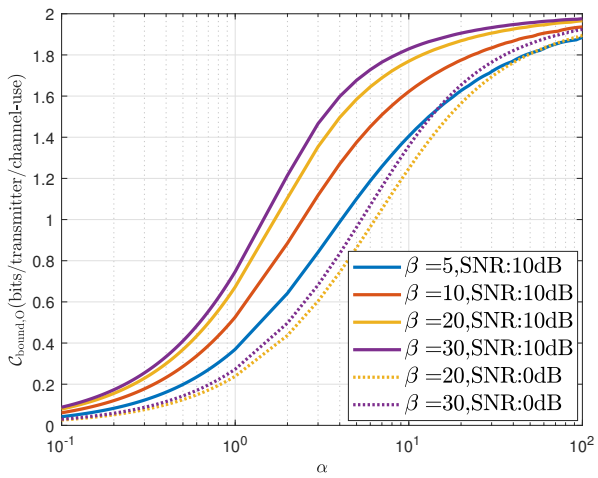

Fig. 3. The lower bound $\mathcal{C}_{\mathrm{bound,O}}$ (41) versus $\alpha = \frac{N}{M}$ in a wide range. Because of the one-bit quantization at the transmitter, the rate per transmitter can not be more than 2. Therefore, we can clearly see the saturation effect when we increase α.

Appendix A
Mutual Information Equivalence

We show the main steps to prove our claim using replica method. Some techniques we use are similar to those used in [7]. According to (10), we have

$$\mathcal{R}_{\mathrm{eff}} = \lim_{M\to\infty} \frac{1}{M}\left(\mathbf{H}(\mathbf{y}_{\mathrm{d}}|X_{\mathrm{t}}, Y_{\mathrm{t}}) - \mathbf{H}(\mathbf{y}_{\mathrm{d}}|X_{\mathrm{t}}, Y_{\mathrm{t}}, \mathbf{x}_{\mathrm{d}})\right) \tag{54}$$

Similar to [7], we apply "replica trick" and have

$$\lim_{M\to\infty} \frac{1}{M}\mathbf{H}(\mathbf{y}_{\mathrm{d}}|X_{\mathrm{t}}, Y_{\mathrm{t}}) = \lim_{M\to\infty} \frac{-1}{M\ln 2}\mathbb{E}\ln\frac{p(\mathbf{y}_{\mathrm{d}}, Y_{\mathrm{t}}|X_{\mathrm{t}})}{p(Y_{\mathrm{t}}|X_{\mathrm{t}})}$$
$$= -\lim_{M\to\infty}\frac{1}{M\ln 2}\lim_{n\to 0}\frac{\partial}{\partial n}\left[\ln\Xi_n - \ln\hat{\Xi}_n\right],$$

where

$$\Xi_n = \mathbb{E}[p^n(\mathbf{y}_{\mathrm{d}}, Y_{\mathrm{t}}|X_{\mathrm{t}})], \hat{\Xi}_n = \mathbb{E}[p^n(Y_{\mathrm{t}}|X_{\mathrm{t}})].$$

We assume the limit of n and M can commute and we have

$$\lim_{M\to\infty}\frac{1}{M\ln 2}\mathbf{H}(\mathbf{y}_{\mathrm{d}}|X_{\mathrm{t}}, Y_{\mathrm{t}}) =$$
$$-\lim_{n\to 0}\frac{\partial}{\partial n}\lim_{M\to\infty}\frac{1}{M\ln 2}\left[\ln\Xi_n - \ln\hat{\Xi}_n\right]. \tag{55}$$

Also, we consider n as integer to derive Ξ_n and $\hat{\Xi}_n$ as a function of n, and we assume the expression still holds for real number n. Then, for integer n, we have

$$\Xi_n = \mathbb{E}[p^n(\mathbf{y}_{\mathrm{d}}, Y_{\mathrm{t}}|X_{\mathrm{t}})]$$
$$= \mathbb{E}_{X_{\mathrm{t}}}\sum_{Y_{\mathrm{t}},\mathbf{y}_{\mathrm{d}}}\left(\mathbb{E}_{\mathbf{x}_{\mathrm{d}},H}p(\mathbf{y}_{\mathrm{d}}, Y_{\mathrm{t}}|\mathbf{x}_{\mathrm{d}}, X_{\mathrm{t}}, H)\right)^{n+1}$$
$$= \mathbb{E}_{X_{\mathrm{t}},[H]_0^n,[\mathbf{x}_{\mathrm{d}}]_0^n}\prod_{k=1}^{N}\sum_{y_{\mathrm{d},k}}\prod_{a=0}^{n} g\left(\sqrt{\frac{\rho}{M}}\mathbf{h}_k^{(a)T}\mathbf{x}_{\mathrm{d}}^{(a)}, y_{\mathrm{d},k}, \sigma^2\right)$$
$$\cdot\prod_{k=1}^{N}\prod_{p=1}^{T_{\mathrm{t}}}\sum_{y_{\mathrm{t},kp}}\prod_{a=0}^{n} g\left(\sqrt{\frac{\rho}{M}}\mathbf{h}_k^{(a)T}\mathbf{x}_{\mathrm{t},p}, y_{\mathrm{t},kp}, \sigma^2\right)$$
$$= \mathbb{E}_{X_{\mathrm{t}},[\mathbf{x}_{\mathrm{d}}]_0^n}\left[\mathbb{E}_{[\mathbf{h}]_0^n}\left[\sum_{y_{\mathrm{d}}}\prod_{a=0}^{n} g\left(\sqrt{\frac{\rho}{M}}\mathbf{h}^{(a)T}\mathbf{x}_{\mathrm{d}}^{(a)}, y_{\mathrm{d}}, \sigma^2\right)\right.\right.$$
$$\left.\left.\cdot\prod_{p=1}^{T_{\mathrm{t}}}\sum_{y_{\mathrm{t},p}}\prod_{a=0}^{n} g\left(\sqrt{\frac{\rho}{M}}\mathbf{h}^{(a)T}\mathbf{x}_{\mathrm{t},p}, y_{\mathrm{t},p}, \sigma^2\right)\right]\right]^N$$

where $y_{\mathrm{t},kp}$ is the kth row and pth column of Y_{t}, $\mathbf{x}_{\mathrm{t},p}$ is the pth column of X_{t}, $\mathbf{h}_k^{(a)T}$ is the kth row of $H^{(a)}$, $[H]_0^n = \{H^{(0)}, \cdots, H^{(n)}\}$,$[\mathbf{h}]_0^n = \{\mathbf{h}^{(0)}, \cdots, \mathbf{h}^{(n)}\}$,$[\mathbf{x}_{\mathrm{d}}]_0^n = \{\mathbf{x}_{\mathrm{d}}^{(0)}, \cdots, \mathbf{x}_{\mathrm{d}}^{(n)}\}$, which are collection of replicas. We drop k at the last step because each element in H are *iid*.

We introduce two $(n+1)\times(n+1)$ matrices $Q_{\mathrm{h}} = [q_{\mathrm{h}}^{ab}]$ and $Q_{\mathrm{x}} = [q_{\mathrm{x}}^{ab}]$ whose elements are defined as $q_{\mathrm{h}}^{ab} = \frac{(\mathbf{h}^{(a)})^H\mathbf{h}^{(b)}}{M}$, $q_{\mathrm{x}}^{ab} = \frac{(\mathbf{x}_{\mathrm{d}}^{(a)})^H\mathbf{x}_{\mathrm{d}}^{(b)}}{M}$. Let $z_p^{(a)} = \sqrt{\frac{1}{M}}\mathbf{h}^{(a)T}\mathbf{x}_{\mathrm{t},p}$, $\mathbf{z}_p = [z_p^{(0)}, \cdots, z_p^{(n)}]^T$, $z_{\mathrm{d}}^{(a)} = \sqrt{\frac{1}{M}}\mathbf{h}^{(a)T}\mathbf{x}_{\mathrm{d}}^{(a)}$, $\mathbf{z}_{\mathrm{d}} = [z_{\mathrm{d}}^{(0)}, \cdots, z_{\mathrm{d}}^{(n)}]^T$, then $\mathbf{z}_p \sim \mathcal{CN}(0, Q_{\mathrm{h}})$, $\mathbf{z}_{\mathrm{d}} \sim \mathcal{CN}(0, Q_z)$ for large M where $Q_z = (Q_{\mathrm{h}} \circ Q_{\mathrm{x}})$ is the Hadamard product between Q_{h} and Q_{x}. Since $\mathbf{x}_{\mathrm{d}}^{(a)}$ is independent of X_{t}, and elements of X_{t} are *iid*, we have

$$\Xi_n = \mathbb{E}_{[\mathbf{x}_{\mathrm{d}}]_0^n}\left[\mathbb{E}_{[\mathbf{h}]_0^n}\left[\mathbb{E}_{\mathbf{z}_{\mathrm{d}}}\sum_{y_{\mathrm{d}}}\prod_{a=0}^{n} g\left(\sqrt{\rho}z_{\mathrm{d}}^{(a)}, y_{\mathrm{d}}, \sigma^2\right)\right.\right.$$
$$\left.\left.\cdot\prod_{p=1}^{T_{\mathrm{t}}}\mathbb{E}_{\mathbf{z}_p}\sum_{y_{\mathrm{t,p}}}\prod_{a=0}^{n} g\left(\sqrt{\rho}z_p^{(a)}, y_{\mathrm{t,p}}, \sigma^2\right)\right]\right]^N$$
$$= \mathbb{E}_{[\mathbf{x}_{\mathrm{d}}]_0^n}\left[\mathbb{E}_{[\mathbf{h}]_0^n}\left[\mathbb{E}_{\mathbf{z}_{\mathrm{d}}}\sum_{y_{\mathrm{d}}}\prod_{a=0}^{n} g\left(\sqrt{\rho}z_{\mathrm{d}}^{(a)}, y_{\mathrm{d}}, \sigma^2\right)\right.\right.$$
$$\left.\left.\cdot\left[\mathbb{E}_{\mathbf{z}}\sum_{y_{\mathrm{t}}}\prod_{a=0}^{n} g\left(\sqrt{\rho}z^{(a)}, y_{\mathrm{t}}, \sigma^2\right)\right]^{T_{\mathrm{t}}}\right]\right]^N,$$

where $\mathbf{z} \sim \mathcal{CN}(0, Q_{\mathrm{h}})$.

Let

$$\mathcal{J}_1(Q_{\mathrm{h}}) = \mathbb{E}_{\mathbf{z}}\sum_{y_{\mathrm{t}}}\prod_{a=0}^{n} g\left(\sqrt{\rho}z^{(a)}, y_{\mathrm{t}}, \sigma^2\right),$$
$$\mathcal{J}_2(Q_{\mathrm{h}}, Q_{\mathrm{x}}) = \mathbb{E}_{\mathbf{z}_{\mathrm{d}}}\sum_{y_{\mathrm{d}}}\prod_{a=0}^{n} g\left(\sqrt{\rho}z_{\mathrm{d}}^{(a)}, y_{\mathrm{d}}, \sigma^2\right). \tag{56}$$

Then

$$\Xi_n = \mathbb{E}_{[\mathbf{x}_{\mathrm{d}}]_0^n}\left[\mathbb{E}_{[\mathbf{h}]_0^n}\left[\left[\mathcal{J}_1(Q_{\mathrm{h}})\right]^{T_{\mathrm{t}}}\mathcal{J}_2(Q_{\mathrm{h}}, Q_{\mathrm{x}})\right]\right]^N. \tag{57}$$

Similarly, we have

$$\begin{aligned}\hat{\Xi}_n &= \mathbb{E}[p^n(Y_{\rm t}|X_{\rm t})]\\ &= \mathbb{E}_{X_{\rm t}}\left[\mathbb{E}_{[\mathbf{h}]_0^n}\prod_{p=1}^{T_{\rm t}}\sum_{y_{{\rm t},p}}\prod_{a=0}^{n} g\left(\sqrt{\frac{\rho}{M}}\mathbf{h}^{(a)T}\mathbf{x}_{{\rm t},p}, y_{{\rm t},p}, \sigma^2\right)\right]^N\\ &= \left[\mathbb{E}_{[\mathbf{h}]_0^n}\left[\mathbb{E}_{\mathbf{z}}\sum_{y_{\rm t}}\prod_{a=0}^{n} g\left(\sqrt{\rho}z^{(a)}, y_{\rm t}, \sigma^2\right)\right]^{T_{\rm t}}\right]^N\\ &= \left[\mathbb{E}_{[\mathbf{h}]_0^n}[\mathcal{J}_1(Q_{\rm h})]^{T_{\rm t}}\right]^N.\end{aligned}$$

Therefore,

$$\ln\Xi_n - \ln\hat{\Xi}_n = \ln\mathbb{E}_{[\mathbf{x}_{\rm d}]_0^n}\left[\frac{\mathbb{E}_{[\mathbf{h}]_0^n}\left[[\mathcal{J}_1(Q_{\rm h})]^{T_{\rm t}}\mathcal{J}_2(Q_{\rm h}, Q_{\rm x})\right]}{\mathbb{E}_{[\mathbf{h}]_0^n}[\mathcal{J}_1(Q_{\rm h})]^{T_{\rm t}}}\right]^N$$

When $T_{\rm t}\to\infty$, based on the saddle point method,

$$\frac{\mathbb{E}_{[\mathbf{h}]_0^n}\left[[\mathcal{J}_1(Q_{\rm h})]^{T_{\rm t}}\mathcal{J}_2(Q_{\rm h}, Q_{\rm x})\right]}{\mathbb{E}_{[\mathbf{h}]_0^n}[\mathcal{J}_1(Q_{\rm h})]^{T_{\rm t}}} \to \mathcal{J}_2(\tilde{Q}_{\rm h}, Q_{\rm x}), \quad (58)$$

where $\tilde{Q}_{\rm h}$ is the saddle point of $\mathbb{E}_{[\mathbf{h}]_0^n}[\mathcal{J}_1(Q_{\rm h})]^{T_{\rm t}}$. We still need to keep it in mind that the saddle point should be considered at the derivative of n with $n\to 0$. Therefore, $\tilde{Q}_{\rm h}$ can be obtained by solving

$$\mathcal{F}_{\rm h} = -\lim_{n\to 0}\frac{\partial}{\partial n}\lim_{M\to\infty}\frac{1}{M}\ln\mathbb{E}_{[\mathbf{h}]_0^n}[\mathcal{J}_1(Q_{\rm h})]^{T_{\rm t}}. \quad (59)$$

It is not hard to show

$$-\lim_{M\to\infty}\frac{1}{MN}\mathbb{E}_{X_{\rm t},Y_{\rm t}}\ln p(Y_{\rm t}|X_{\rm t}) = \mathcal{F}_{\rm h} \quad (60)$$

through regular steps used in replica method with similar assumptions.

Now, we use replica symmetry (RS) assumption by assuming the off-diagonal elements of the saddle point $\tilde{Q}_{\rm h}$ are equal, denoted as $q_{\rm h}$. The diagonal elements of $\tilde{Q}_{\rm h}$ are 1, which is the variance of the elements of channel. According to [7], [30], when we obtain the saddle point $q_{\rm h}$ through (60), $1-q_{\rm h}$ describes the MSE of the MMSE channel estimation, shown in (13).

Then, we have

$$\begin{aligned}&\lim_{M\to\infty}\frac{1}{M}\left[\ln\Xi_n - \ln\hat{\Xi}_n\right]\\ &= \lim_{M\to\infty}\frac{1}{M}\ln\mathbb{E}_{[\mathbf{x}_{\rm d}]_0^n}\left(\mathcal{J}_2(\tilde{Q}_{\rm h}, Q_{\rm x})\right)^N\\ &= \lim_{M\to\infty}\frac{1}{M}\ln\mathbb{E}_{[\mathbf{x}_{\rm d}]_0^n}\left[\sum_{y_{\rm d}}\mathbb{E}_{\mathbf{z}_{\rm d}}\prod_{a=0}^{n} g(\sqrt{\rho}z_{\rm d}^{(a)}, y_{\rm d}, \sigma^2)\right]^N,\end{aligned}$$

where $\mathbf{z}_{\rm d}\sim\mathcal{CN}(0, Q_z)$, $Q_z = \tilde{Q}_{\rm h}\circ Q_{\rm x}$.

Therefore,

$$\begin{aligned}&\lim_{M\to\infty}\frac{1}{M}\mathbf{H}(\mathbf{y}_{\rm d}|X_{\rm t}, Y_{\rm t}) = -\lim_{n\to 0}\frac{\partial}{\partial n}\lim_{M\to\infty}\frac{1}{M\ln 2}\\ &\ln\mathbb{E}_{[\mathbf{x}_{\rm d}]_0^n}\left[\sum_{y_{\rm d}}\mathbb{E}_{\mathbf{z}_{\rm d}}\prod_{a=0}^{n} g(\sqrt{\rho}z_{\rm d}^{(a)}, y_{\rm d}, \sigma^2)\right]^N.\end{aligned}$$

Similarly, we have

$$\begin{aligned}&\lim_{M\to\infty}\frac{1}{M}\mathbf{H}(\mathbf{y}_{\rm d}|X_{\rm t}, Y_{\rm t}, \mathbf{x}_{\rm d}) = -\lim_{n\to 0}\frac{\partial}{\partial n}\lim_{M\to\infty}\frac{1}{M\ln 2}\\ &\ln\left[\sum_{y_{\rm d}}\mathbb{E}_{\mathbf{z}_h}\prod_{a=0}^{n} g(\sqrt{\rho}z_h^{(a)}, y_{\rm d}, \sigma^2)\right]^N,\end{aligned}$$

where $\mathbf{z}_h = [z_h^{(0)}, \cdots, z_h^{(n)}]^T$ and $\mathbf{z}_h\sim\mathcal{CN}(0, \tilde{Q}_{\mathbf{h}})$.

For the equivalent channel with known $\hat{H}$ shown in (12), similarly, we have

$$\begin{aligned}&\lim_{M\to\infty}\frac{1}{M}\mathbf{H}(\hat{\mathbf{y}}_{\rm d}|\hat{H}) = -\lim_{n\to 0}\frac{\partial}{\partial n}\lim_{M\to\infty}\frac{1}{M\ln 2}\\ &\ln\mathbb{E}_{[\mathbf{x}_{\rm d}]_0^n}\left[\sum_{y_{\rm d}}\mathbb{E}_{\hat{\mathbf{z}}_{\rm d}}\prod_{a=0}^{n} g(\sqrt{\rho}\hat{z}_{\rm d}^{(a)}, y_{\rm d}, \hat{\sigma}^2)\right]^N, \quad (61)\end{aligned}$$

where $\hat{\mathbf{z}}_{\rm d} = [\hat{z}_{\rm d}^{(0)}, \cdots, \hat{z}_{\rm d}^{(n)}]^T\sim\mathcal{CN}(0, q_{\rm h}\cdot Q_{\rm x})$, $\hat{\sigma}^2 = \sigma^2+\rho(1-q_{\rm h})$. The joint distribution of $\sqrt{\rho}\hat{\mathbf{z}}_{\rm d}+\hat{\mathbf{v}}$ is the same as the joint distribution of $\sqrt{\rho}\mathbf{z}_{\rm d}+\mathbf{v}$. According to (14) and (15), we have

$$\mathbb{E}_{\mathbf{z}_{\rm d}}\prod_{a=0}^{n} g(\sqrt{\rho}z_{\rm d}^{(a)}, y_{\rm d}, \sigma^2) = \mathbb{E}_{\hat{\mathbf{z}}_{\rm d}}\prod_{a=0}^{n} g(\sqrt{\rho}\hat{z}_{\rm d}^{(a)}, y_{\rm d}, \hat{\sigma}^2),$$

and therefore

$$\lim_{M\to\infty}\frac{1}{M}\mathbf{H}(\mathbf{y}_{\rm d}|X_{\rm t}, Y_{\rm t}) = \lim_{M\to\infty}\frac{1}{M}\mathbf{H}(\hat{\mathbf{y}}_{\rm d}|\hat{H}).$$

Similarly, we have

$$\lim_{M\to\infty}\frac{1}{M}\mathbf{H}(\mathbf{y}_{\rm d}|X_{\rm t}, Y_{\rm t}, \mathbf{x}_{\rm d}) = \lim_{M\to\infty}\frac{1}{M}\mathbf{H}(\hat{\mathbf{y}}_{\rm d}|\hat{H}, \mathbf{x}_{\rm d}).$$

Therefore,

$$\lim_{M\to\infty}\frac{1}{M}\mathbf{I}(\mathbf{x}_{\rm d}; \mathbf{y}_{\rm d}|X_{\rm t}, Y_{\rm t}) = \lim_{M\to\infty}\frac{1}{M}\mathbf{I}(\mathbf{x}_{\rm d}; \hat{\mathbf{y}}_{\rm d}|\hat{H}),$$

where $\mathbf{x}_{\rm d}$ and $\hat{\mathbf{y}}_{\rm d}$ are the input and output of channel defined in (12).

Appendix B
Proof of Replica Method to Compute $\mathcal{C}_{\rm bound}$

According to Appendix A, $q_{\rm h}$ can be obtained by solving $\mathcal{F}_{\rm h}$ defined in (59).

Similarly to [21], we apply Varadhan's theorem and Gartner-Ellis theorem [31] and obtain

$$\begin{aligned}&\lim_{M\to\infty}\frac{1}{M}\ln\mathbb{E}_{[\mathbf{h}]_0^n}[\mathcal{J}_1(Q_{\rm h})]^{T_{\rm t}} = \sup_{Q_{\rm h}}\inf_{\hat{Q}_{\rm h}}\Big[\beta_{\rm t}G_1(Q_{\rm h})\\ &-\sum_{a<b}2q_{\rm h}^{ab}\hat{q}_{\rm h}^{ab} + L_1(\hat{Q}_{\rm h})\Big],\end{aligned}$$

where $q_{\rm h}^{ab}$ and $\hat{q}_{\rm h}^{ab}$ are elements of $Q_{\rm h}$ and $\hat{Q}_{\rm h}$,

$$G_1(Q_{\rm h}) = \ln\left(\mathcal{J}_1(Q_{\rm h})\right), \quad (62)$$

$$L_1(\hat{Q}_{\rm h}) = \lim_{M\to\infty} \frac{1}{M} \ln \mathbb{E}_{[\mathbf{h}]_0^n} \exp\left(\sum_{a<b} 2\hat{q}_{\rm h}^{ab} (\mathbf{h}^{(a)})^H \mathbf{h}^{(b)}\right).$$

The values of $Q_{\rm h}$ and $\hat{Q}_{\rm h}$ that achieves the extremum are called saddle point. Based on the RS assumption, we assume the off-diagonal elements of $Q_{\rm h}$ and $\hat{Q}_{\rm h}$ are the same, denoted as $q_{\rm h}$ and $\hat{q}_{\rm h}$, respectively. Then, we have

$$\beta_{\rm t} G_1(Q_{\rm h}) = -\mathcal{G}_1(n, q_{\rm h}),$$

where $\mathcal{G}_1(n,q)$ is defined in (17). Also,

$$L_1(\hat{Q}_{\rm h}) = \mathcal{L}_1(n, \hat{q}_{\rm h}),$$

where

$$\mathcal{L}_1(n,\hat{q}_{\rm h}) = -(n+1)\ln(1+\hat{q}_{\rm h}) - \ln\left(1 - \frac{\hat{q}_{\rm h}(n+1)}{1+\hat{q}_{\rm h}}\right).$$

Therefore, (59) becomes

$$\mathcal{F}_{\rm h} = \mathcal{F}_1(q_{\rm h}, \hat{q}_{\rm h}),$$

where $\mathcal{F}_1(q,\hat{q})$ is defined in (18). $(q_{\rm h}, \hat{q}_{\rm h})$ is the saddle point of $\mathcal{F}_1(q,\hat{q})$ and we have

$$\frac{\partial \mathcal{F}_1(q,\hat{q})}{\partial q} = 0, \frac{\partial \mathcal{F}_1(q,\hat{q})}{\partial \hat{q}} = 0, \tag{63}$$

at $(q_{\rm h}, \hat{q}_{\rm h})$. If there are multiple solutions, we should use the one that minimize $\mathcal{F}_1(q_{\rm h}, \hat{q}_{\rm h})$. We finish the proof of solving $q_{\rm h}$ in steps 2B)-4B) shown in the recipe. $q_{\rm h}$ depends on $\beta_{\rm t}$, and for the rest of the steps, only $q_{\rm h}$ is needed to compute $\mathcal{R}_{\rm eff}(\beta_{\rm t})$.

We use the equivalent channel shown in (11) to compute $\mathcal{R}_{\rm eff}(\beta_{\rm t})$ and we have

$$\mathcal{R}_{\rm eff}(\beta_{\rm t}) = \lim_{M\to\infty} \frac{1}{M}(\mathbf{H}(\hat{\mathbf{y}}_{\rm d}|\hat{H}) - \mathbf{H}(\hat{\mathbf{y}}_{\rm d}|\hat{H}, \mathbf{x}_{\rm d})). \tag{64}$$

Since $\sqrt{\frac{\rho}{M}}\hat{H}\mathbf{x}_{\rm d} \sim \mathcal{CN}(0, \rho q_{\rm h} \cdot I)$, we have

$$\lim_{M\to\infty} \frac{1}{M}\mathbf{H}(\hat{\mathbf{y}}_{\rm d}|\hat{H}, \mathbf{x}_{\rm d})) = \frac{\alpha}{\ln 2}\sum_{y\in\mathcal{Y}} \mathbb{E}_z\Big(g(\sqrt{\rho q_{\rm h}}z, y, \hat{\sigma}^2) \cdot \ln g(\sqrt{\rho q_{\rm h}}z, y, \hat{\sigma}^2)\Big), z \sim \mathcal{CN}(0,1).$$

According to (61), we have

$$\lim_{M\to\infty} \frac{1}{M}\mathbf{H}(\hat{\mathbf{y}}_{\rm d}|\hat{H}) = -\lim_{n\to 0}\frac{\partial}{\partial n}\lim_{M\to\infty}\frac{1}{M\ln 2}\ln \Xi_{\rm x,n}, \tag{65}$$

where

$$\Xi_{\rm x,n} = \mathbb{E}_{[\mathbf{x}_{\rm d}]_0^n}\left[\sum_{y_{\rm d}} \mathbb{E}_{\hat{\mathbf{z}}_{\rm d}} \prod_{a=0}^{n} g(\sqrt{\rho}\hat{z}_{\rm d}^{(a)}, y_{\rm d}, \hat{\sigma}^2)\right]^N = \mathbb{E}_{[\mathbf{x}_{\rm d}]_0^n}\left[\sum_{y_{\rm d}} \mathbb{E}_{\mathbf{w}} \prod_{a=0}^{n} g(\sqrt{\rho q_{\rm h}}w^{(a)}, y_{\rm d}, \hat{\sigma}^2)\right]^N,$$

with $\mathbf{w} = [w^{(0)}, \cdots, w^{(n)}]^T \sim \mathcal{CN}(0, Q_{\rm x})$. The elements of $Q_{\rm x}$ are $q_{\rm x}^{ab}$ defined as $q_{\rm x}^{ab} = \frac{(\mathbf{x}_{\rm d}^{(a)})^H \mathbf{x}_{\rm d}^{(b)}}{M}$.

We again apply Varadhan's theorem and Gartner-Ellis theorem [31] and obtain

$$\lim_{M\to\infty}\frac{\ln \Xi_{\rm x,n}}{M} = \sup_{Q_{\rm x}}\inf_{\hat{Q}_{\rm x}}\Big[\alpha G_2(Q_{\rm x}) - \sum_{a<b} 2q_{\rm x}^{ab}\hat{q}_{\rm x}^{ab} + G_3(\hat{Q}_{\rm x})\Big],$$

where

$$G_2(Q_{\rm x}) = \ln \sum_{y\in\mathcal{Y}} \mathbb{E}_{\mathbf{w}} \prod_{a=0}^{n} g(\sqrt{\rho q_{\rm h}}w^{(a)}, y, \hat{\sigma}^2) \tag{66}$$

$$G_3(\hat{Q}_{\rm x}) = \lim_{M\to\infty}\frac{1}{M}\ln \mathbb{E}_{[\mathbf{x}_{\rm d}]_0^n} \exp\left(\sum_{a<b} 2q_{\rm x}^{ab}(\mathbf{x}_{\rm d}^{(a)})^H \mathbf{x}_{\rm d}^{(b)}\right).$$

Now we apply the RS assumption by considering the off-diagonal elements of $Q_{\rm x}$ and $\hat{Q}_{\rm x}$ are the same at the saddle point, denoted as $q_{\rm x}$ and $\hat{q}_{\rm x}$. Then, we have

$$\alpha G_2(Q_{\rm x}) = -\mathcal{G}_2(n, q_{\rm x}),$$
$$G_3(\hat{Q}_{\rm x}) = -\mathcal{G}_3(n, \hat{q}_{\rm x}),$$

where $\mathcal{G}_2(n, q_{\rm x})$ and $\mathcal{G}_3(n, \hat{q}_{\rm x})$ are defined in (20) and (21). Becasue of the symmetry, only real part of $(\mathbf{x}_{\rm d}^{(a)})^H \mathbf{x}_{\rm d}^{(b)}$ is needed for computation.

Therefore, (65) becomes

$$\lim_{M\to\infty}\frac{1}{M}\mathbf{H}(\hat{\mathbf{y}}_{\rm d}|\hat{H}) = \frac{1}{\ln 2}\mathcal{F}_2(q_{\rm x}, \hat{q}_{\rm x}), \tag{67}$$

where $\mathcal{F}_2(r,\hat{r})$ is defined in (22). $(q_{\rm x}, \hat{q}_{\rm x})$ is the saddle point of $\mathcal{F}_2(r,\hat{r})$, and we have

$$\frac{\partial \mathcal{F}_2(r,\hat{r})}{\partial r} = 0, \frac{\partial \mathcal{F}_2(r,\hat{r})}{\partial \hat{r}} = 0, \tag{68}$$

at $(q_{\rm x}, \hat{q}_{\rm x})$. If there are multiple solutions, we should use the solution that minimize $\mathcal{F}_2(q_{\rm x}, \hat{q}_{\rm x})$. This proves the rest of the recipe.

References

[1] J. A. Nossek and M. T. Ivrlač, "Capacity and coding for quantized MIMO systems," in *Proceedings of the 2006 international conference on Wireless communications and mobile computing*. ACM, 2006, pp. 1387–1392.

[2] C. Mollén, J. Choi, E. G. Larsson, and R. W. Heath, "Uplink performance of wideband massive MIMO with one-bit ADCs," *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 87–100, 2017.

[3] Y. Li, C. Tao, G. Seco-Granados, A. Mezghani, A. L. Swindlehurst, and L. Liu, "Channel estimation and performance analysis of one-bit massive MIMO systems," *IEEE Trans. Signal Process*, vol. 65, no. 15, pp. 4075–4089, 2017.

[4] Y. Li, C. Tao, L. Liu, A. Mezghani, and A. L. Swindlehurst, "How much training is needed in one-bit massive MIMO systems at low SNR?" in *Global Communications Conference (GLOBECOM), 2016 IEEE*. IEEE, 2016, pp. 1–6.

[5] S.-N. Hong, S. Kim, and N. Lee, "A weighted minimum distance decoding for uplink multiuser MIMO systems with low-resolution ADCs," *IEEE Transactions on Communications*, vol. 66, no. 5, pp. 1912–1924, 2018.

[6] L. Fan, S. Jin, C.-K. Wen, and H. Zhang, "Uplink achievable rate for massive MIMO systems with low-resolution ADC," *IEEE Communications Letters*, vol. 19, no. 12, pp. 2186–2189, 2015.

[7] C.-K. Wen, C.-J. Wang, S. Jin, K.-K. Wong, and P. Ting, "Bayes-optimal joint channel-and-data estimation for massive mimo with low-precision ADCs," *IEEE Transactions on Signal Processing*, vol. 64, no. 10, pp. 2541–2556, 2016.

[8] A. K. Saxena, I. Fijalkow, A. Mezghani, and A. L. Swindlehurst, "Analysis of one-bit quantized ZF precoding for the multiuser massive MIMO downlink," in *Signals, Systems and Computers, 2016 50th Asilomar Conference on*. IEEE, 2016, pp. 758–762.

[9] Y. Li, C. Tao, A. L. Swindlehurst, A. Mezghani, and L. Liu, "Downlink achievable rate analysis in massive MIMO systems with one-bit DACs," *IEEE Communications Letters*, vol. 21, no. 7, pp. 1669–1672, 2017.

[10] S. Jacobsson, G. Durisi, M. Coldrey, and C. Studer, "Massive MU-MIMO-OFDM downlink with one-bit DACs and linear precoding," *arXiv preprint arXiv:1704.04607*, 2017.

[11] C. Kong, A. Mezghani, C. Zhong, A. L. Swindlehurst, and Z. Zhang, "Nonlinear precoding for multipair relay networks with one-bit ADCs and DACs," *IEEE Signal Processing Letters*, vol. 25, no. 2, pp. 303–307, 2018.

[12] O. B. Usman, H. Jedda, A. Mezghani, and J. A. Nossek, "MMSE precoder for massive MIMO using 1-bit quantization," in *2016 IEEE Int. Conf on Acoust, Speech and Signal Proc.*, 2016, pp. 3381–3385.

[13] K. Gao, N. Estes, B. Hochwald, J. Chisum, and J. N. Laneman, "Power-performance analysis of a simple one-bit transceiver," in *Information Theory and Applications Workshop (ITA), 2017*. IEEE, 2017, pp. 1–10.

[14] K. Gao, J. N. Laneman, and B. Hochwald, "Beamforming with multiple one-bit wireless transceivers," in *2018 Information Theory and Applications Workshop (ITA)*, Feb 2018, pp. 1–9.

[15] C. Kong, A. Mezghani, C. Zhong, A. L. Swindlehurst, and Z. Zhang, "Multipair massive MIMO relaying systems with one-bit adc s and dac s," *IEEE Transactions on Signal Processing*, vol. 66, no. 11, pp. 2984–2997, 2018.

[16] K. Gao, J. N. Laneman, and B. Hochwald, "Capacity of multiple one-bit transceivers in a rayleigh environment," in *Wireless Communications and Networking Conference (WCNC), 2018 IEEE*. IEEE, 2018, pp. 1–6.

[17] B. Hassibi and B. M. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Transactions on Information Theory*, vol. 49, no. 4, pp. 951–963, 2003.

[18] M. Mezard, G. Parisi, M. A. Virasoro, and D. J. Thouless, "Spin glass theory and beyond," *Physics Today*, vol. 41, p. 109, 1988.

[19] T. Tanaka, "Analysis of bit error probability of direct-sequence CDMA multiuser demodulators," *Advances in Neural Information Processing Systems*, pp. 315–321, 2001.

[20] ——, "A statistical-mechanics approach to large-system analysis of CDMA multiuser detectors," *IEEE Transactions on Information theory*, vol. 48, no. 11, pp. 2888–2910, 2002.

[21] ——, "Statistical learning in digital wireless communications," in *International Conference on Algorithmic Learning Theory*. Springer, 2004, pp. 464–478.

[22] D. Guo and S. Verdú, "Randomly spread CDMA: Asymptotics via statistical physics," *IEEE Transactions on Information Theory*, vol. 51, no. 6, pp. 1983–2010, 2005.

[23] ——, "Multiuser detection and statistical mechanics," in *Communications, Information and Network Security*. Springer, 2003, pp. 229–277.

[24] E. Gardner, "The space of interactions in neural network models," *Journal of physics A: Mathematical and general*, vol. 21, no. 1, p. 257, 1988.

[25] A. Engel and C. Van den Broeck, *Statistical mechanics of learning*. Cambridge University Press, 2001.

[26] A. Montanari and N. Sourlas, "The statistical mechanics of turbo codes," *The European Physical Journal B-Condensed Matter and Complex Systems*, vol. 18, no. 1, pp. 107–119, 2000.

[27] H. Nishimori and K. M. Wong, "Statistical mechanics of image restoration and error-correcting codes," *Physical Review E*, vol. 60, no. 1, p. 132, 1999.

[28] A. L. Moustakas, S. H. Simon, and A. M. Sengupta, "MIMO capacity through correlated channels in the presence of correlated interferers and noise: A (not so) large n analysis," *IEEE Transactions on Information Theory*, vol. 49, no. 10, pp. 2545–2561, 2003.

[29] E. W. Cheney and D. R. Kincaid, *Numerical mathematics and computing*. Cengage Learning, 2012.

[30] M. Opper and W. Kinzel, "Statistical mechanics of generalization," in *Models of neural networks III*. Springer, 1996, pp. 151–209.

[31] A. Dembo and O. Zeitouni, "Large deviations techniques and applications," *Applications of Mathematics*, vol. 38, 1998.