CHALMERS
UNIVERSITY OF TECHNOLOGY

Capacity Bounds under Imperfect Polarization Tracking

Citation for the original published paper (version of record):
Farsi, M., Karlsson, M., Agrell, E. (2022). Capacity Bounds under Imperfect Polarization Tracking. IEEE Transactions on Communications, 70(11): 7240-7249.
http://dx.doi.org/10.1109/TCOMM.2022.3206803

(article starts on next page)

Capacity Bounds under Imperfect Polarization Tracking

Mohammad Farsi, *Student Member, IEEE,* Magnus Karlsson, *Senior Member, IEEE,* and Erik Agrell, *Fellow, IEEE*

***Abstract*—In optical fiber communication, due to the random variation of the environment, the state of polarization (SOP) fluctuates randomly with time leading to distortion and performance degradation. The memory-less SOP fluctuations can be regarded as a two-by-two random unitary matrix. In this paper, for what we believe to be the first time, the capacity of the polarization drift channel under an average power constraint with imperfect channel knowledge is characterized. An achievable information rate (AIR) is derived when imperfect channel knowledge is available and is shown to be highly dependent on the channel estimation technique. It is also shown that a tighter lower bound can be achieved when a unitary estimation of the channel is available. However, the conventional estimation algorithms do not guarantee a unitary channel estimation. Therefore, by considering the unitary constraint of the channel, a data-aided channel estimator based on the Kabsch algorithm is proposed, and its performance is numerically evaluated in terms of AIR. Monte Carlo simulations show that Kabsch outperforms the least-square error algorithm. In particular, with complex, Gaussian inputs and eight pilot symbols per block, Kabsch improves the AIR by 0.20 to 0.30 bits/symbol throughout the range of studied signal-to-noise ratios.**

***Index Terms*—Achievable information rate, constant modulus algorithm, capacity, channel estimation, decision-directed least mean square, Kabsch algorithm, lower bound, least square error, mutual information, multimode fiber, multicore fiber, polarization-mode dispersion, state of polarization, space-division multiplexing.**

I. Introduction

THE growing demand for reliable long-distance communication makes it essential to determine the capacity limits of optical links [1]–[3]. The emergence of coherent optical communication systems enabled digital signal processing (DSP) and polarization-division multiplexing to achieve higher spectral efficiencies. However, higher data rates come with increased sensitivity to impairments such as polarization-mode dispersion (PMD) and state of polarization (SOP) fluctuations, which must be tracked dynamically in the receiver [4]. Due to the random variation of both internal and environmental impairments, the SOP fluctuates randomly with time [5]. Previous long-term measurements showed that the SOP drift might vary from days and hours in buried fibers [6], [7] to microseconds in aerial fibers [8]–[10].

M. Farsi and E. Agrell are with the Department of Electrical Engineering, Chalmers University of Technology, SE-41296 Gothenburg, Sweden (e-mail: farsim@chalmers.se; agrell@chalmers.se)

M. Karlsson is with the Department of Microtechnology and Nanoscience, Chalmers University of Technology, SE-41296 Gothenburg, Sweden (e-mail: magnus.karlsson@chalmers.se).

This work was supported by the Knut and Alice Wallenberg Foundation under grant 2018.0090.

The conventional DSP solutions for SOP tracking are replete with both blind and data-aided algorithms. The constant modulus algorithm (CMA) [11], thanks to its low complexity and tolerance to phase noise (PN), has been widely considered for blind polarization tracking [12]–[14]. Many studies have been conducted to overcome the so-called "singularity" problem of CMA [15]–[17], where the two estimated channels converge twice to the same polarization. The radially directed equalizer (RDE) and its variants were proposed to account for higher-order quadrature amplitude modulation (QAM) in [18]–[20]. The multi-modulus algorithm (MMA) [21] is also applicable to higher-order modulations. The decision-directed least mean squares (DDLMS) algorithm removed the modulation format dependence of the blind algorithms [13]. Furthermore, block-wise CMA (BW-CMA) and DDLMS (BW-DDLMS) were proposed in [22] for block-constant channels. More reliable convergence can be obtained using data-aided algorithms such as least mean square (LMS), which was adopted for SOP estimation by [13], [23], or standard least-square error (LS), which has been extensively studied for both wireless and optical applications [24]–[27]. While the LMS algorithm performs real-time continuous equalization, LS applies to block-based estimation.

While the dual-polarization (DP) channel can be represented by a unitary matrix, the majority of the available estimation techniques including LS and LMS have no unitary constraint on the estimated channel, leading to sub-optimal solutions. This makes room for algorithms that provide a unitary estimation of the DP channel. Louchet *et al.* [28] proposed the Kabsch algorithm [29] as a joint blind PN and PMD estimation method. In [30], a blind modulation-format-independent joint PN and polarization tracking algorithm was proposed, and its performance was compared with the blind Kabsch algorithm.

Depending on the estimation algorithm, the speed of the fluctuations, or the additive noise, the polarization tracking might be imperfect. This makes it relevant to ask how the capacity is affected by such an estimation error? While there are no capacity studies regarding the polarization drift channels, to the best of our knowledge, the literature is replete with fiber capacity results. The first finite capacity limit for the single-wavelength optical channel based on a sort of Gaussian Noise (GN) model was introduced in [31], [32], showing a lower bound that increases with the power until it reaches a maximum, and then it decreases due to the nonlinear impairments (NLI). Although different nonlinear models were applied, similar results were reported in [33], [34]. Since the GN model cannot truly describe the NLI [35], some modeled

the NLI as a linear time-variant distortion highly correlated over time [36], [37]. This made it possible to counteract NLI using the techniques that were conventionally used for linear impairments (e.g., PN, chromatic dispersion, PMD, and SOP [38]–[40]). For example, in [41], a tighter lower bound for single-polarization, dispersion-unmanaged systems was achieved by considering the PN.

In this paper, for the first time, we study the capacity of the polarization drift channel in the presence of imperfect knowledge of the channel (e.g., imperfect polarization tracking). Using the mismatched decoding method, we derive an AIR, which is a lower bound on the capacity, in the presence of a channel estimation error. We show that a unitary estimation of the channel leads to a tighter lower bound, which makes it reasonable to seek unitary estimators. A data-aided version of the Kabsch algorithm is proposed for DP channel estimation, for the first time to the best of our knowledge. We compare Kabsch with an LS algorithm in terms of achievable information rate (AIR) and show that Kabsch outperforms LS throughout the range of considered signal-to-noise ratios (SNRs). For instance, with only eight pilot symbols per channel and block, Kabsch improves the AIR by at least 0.2 bits per symbol compared to LS.

Notation: Column vectors are denoted by underlined letters $\underline{x}$ and matrices by uppercase roman letters X. We use boldface letters $\mathbf{x}$ for random quantities and the corresponding nonbold letters x for their realizations. Probability density functions (pdfs) are denoted as $p_{\mathbf{x}}(x)$ and conditional pdfs as $p_{\mathbf{y}|\mathbf{x}}(y|x)$, where the subscripts will sometimes be omitted if they are clear from the context. Expectation over random variables is denoted by $\mathbb{E}[\cdot]$. Sets are indicated by uppercase calligraphic letters $\mathcal{X}$. The complex zero-mean circularly symmetric Gaussian distribution of a vector is denoted by $\underline{\mathbf{x}} \sim \mathcal{CN}(\underline{0}, \Lambda_{\underline{x}})$, where $\Lambda_{\underline{x}} = \mathbb{E}[\mathbf{x}\mathbf{x}^\dagger]$ is the covariance matrix. The logarithm $\log$ refers to base 2. In the context of matrix operations, $|\cdot|$, $(\cdot)^T$, $(\cdot)^\dagger$, and $\|.\|$ represent the determinant, transpose, conjugate transpose, and Frobenius norm operators, respectively.

II. Channel Model and Mutual Information

A. Channel Model

We consider transmission over n channels in the presence of amplified spontaneous emission (ASE) noise at the receiver. The channel is assumed to be constant during a transmission block of length N, and changes randomly and independently between the blocks. The assumption that the channel does not change within a block is consistent with the fact that the SOP drifts at a much slower rate than typical transmission rates in optical links [6], [7] and is well established in optical communications literature [39], [41]. The block length N is chosen based on the application and the drift speed of the channel. The independence between random blocks is an idealized, not fully realistic, assumption, which is well established in the literature, where the SOP fluctuation is often modeled as a constant randomly chosen rotations [42], [43] or deterministic cyclic/quasi-cyclic changes [13], [44], [45]. In most optical communication systems, there is no feedback channel between the transmitter and the receiver. Therefore, the problem of input distribution optimization, corresponding to the case that the channel is known at the transmitter, is not investigated. It is assumed that PMD is negligible and all channel impairments, including nonlinearities and chromatic dispersion are ideally compensated, with the exception of polarization fluctuation and ASE noise, which are modeled by a unitary matrix and additive white Gaussian noise (AWGN), respectively. Although neglecting most real-life impairments is not fully realistic, it is the simplest case and nevertheless never before investigated. Indeed, adding more impairments would make the analysis more realistic but also more complex.

Since different blocks are independent, we just model the symbols in one transmission block in the following. The transmitted signal in each channel at time $k = 0, \dots, N-1$ is an n-dimensional random complex vector $\underline{\mathbf{s}}_k$, where $n = 2$ for DP, single-mode, single-core transmission.[1] The signal vectors $\underline{\mathbf{s}}_k$ take on values from a set $\mathcal{S}$ of zero-mean constellation points. After filtering and resampling the received signals into one sample per symbol, the vector of received samples $\underline{\mathbf{x}}_k$ can be expressed as

$$\underline{\mathbf{x}}_k = \mathrm{H}\underline{\mathbf{s}}_k + \underline{\mathbf{z}}_k, \tag{1}$$

where the $n \times n$ matrix H represents a multiple-input multiple-output (MIMO) channel and $\underline{\mathbf{z}}_k$ denotes the complex ASE noise samples at time k, which is assumed to be $\mathcal{CN}(\underline{0}, \sigma_z^2 \mathrm{I}_n)$ and independent of $\underline{\mathbf{s}}_k$. In the remainder of this paper, we will omit the time index k explicitly for notational convenience. This is possible because the input and noise are independent and identically distributed (i.i.d.) over time. We define the covariance matrix of the input vector as

$$\Lambda_{\underline{s}} = \mathbb{E}_{\underline{\mathbf{s}}}[\underline{\mathbf{s}}\underline{\mathbf{s}}^\dagger]. \tag{2}$$

In order to maintain the generality of the results, they are given for an arbitrary number of channels n whenever possible. Thus, the MIMO-AWGN channel (1) can describe a wide range of applications and impairments. However, for the purpose of DP optical channel modeling, we are particularly interested in the special case of $n = 2$ and H being unitary, denoted by $\mathrm{H_u}$.

B. Mutual Information with Perfect Knowledge of Channel at the Receiver

The conditional mutual information (MI) between two random vectors $\underline{\mathbf{s}}$, $\underline{\mathbf{x}}$, when the channel H is given, is defined as [46, Eq. (2.61)]

$$I(\underline{\mathbf{s}}; \underline{\mathbf{x}}) = \mathbb{E}_{\underline{\mathbf{s}},\underline{\mathbf{x}}}\left[\log\left(\frac{p(\underline{\mathbf{x}}|\underline{\mathbf{s}}, \mathrm{H})}{p(\underline{\mathbf{x}}|\mathrm{H})}\right)\right]. \tag{3}$$

The capacity of this channel under an average power constraint is [47]

$$C = \max_{p(\underline{\mathbf{s}})} I(\underline{\mathbf{s}}; \underline{\mathbf{x}}) \quad \text{s.t.} \quad \mathrm{tr}(\Lambda_{\underline{s}}) \leq P, \tag{4}$$

[1]The extension to space-division-multiplexed channels, for which $n > 2$, is straight-forward.

where $P > 0$ is the transmission power constraint. For the MIMO-AWGN channel in (1), the channel law given H and $\underline{s}$ is characterized by the pdf

$$p(\underline{x}|\underline{s},\mathrm{H}) = \frac{1}{\pi^n \sigma_z^{2n}} \exp\left(-\frac{\|\underline{x}-\mathrm{H}\underline{s}\|^2}{\sigma_z^2}\right). \tag{5}$$

Then, the pdf of the channel output can be calculated as

$$p(\underline{x}|\mathrm{H}) = \mathbb{E}_{\underline{\mathbf{s}}}\left[p(\underline{x}|\underline{\mathbf{s}},\mathrm{H})\right]. \tag{6}$$

Given H, the capacity-achieving distribution of the MIMO-AWGN channel law (5) is $\mathcal{CN}$ [48]. Therefore, assuming $\mathcal{CN}(\underline{0},\Lambda_{\underline{s}})$ inputs, the covariance matrix of the received samples when the channel is given, can be written as

$$\begin{aligned}\Lambda_{\underline{x}} &= \mathbb{E}_{\underline{\mathbf{x}}}\left[\underline{\mathbf{x}}\underline{\mathbf{x}}^\dagger\right] \\ &= \mathrm{H}\Lambda_{\underline{s}}\mathrm{H}^\dagger + \sigma_z^2\mathrm{I}_n.\end{aligned} \tag{7}$$

The MI of a general MIMO system for a given channel is [48]

$$I(\underline{\mathbf{s}};\underline{\mathbf{x}}) = \log\left|\mathrm{I}_n + \mathrm{H}^\dagger\mathrm{H}\mathrm{Q}\right|, \tag{8}$$

where

$$\mathrm{Q} = \Lambda_{\underline{s}}/\sigma_z^2, \tag{9}$$

and hence $\mathrm{tr}(\mathrm{Q}) \le n\eta$, where $\eta = P/(n\sigma_z^2)$ is the SNR of each channel. If the channel matrix H is confined to the set of unitary matrices (i.e., $\mathrm{H} = \mathrm{H_u}$), (8) gives

$$I(\underline{\mathbf{s}};\underline{\mathbf{x}}) = \log\left|\mathrm{I}_n + \mathrm{Q}\right|. \tag{10}$$

The capacity of a unitary MIMO-AWGN channel is the supremum of (10) for all possible Q. In general, Q needs to be optimized for each realization of the channel H, if it is known at the transmitter; however, based on (10), the MI is independent of H. Since maximizing (10) is equivalent to maximizing $|\mathrm{I}_n + \mathrm{Q}|$ and Q is positive definite, the nondiagonal elements of Q must be zero, yielding Q to be a diagonal positive definite matrix. From the well-known theorem that the geometric mean is always upper-bounded by the arithmetic mean, it is straight-forward to show that uniform power distribution at the transmitter (i.e., $\mathrm{Q} = \eta\mathrm{I}_n$) maximizes (10). Thus, the capacity of an n-dimensional unitary MIMO-AWGN channel, still assuming that the channel is perfectly known at the receiver, is

$$C = n\log(1+\eta). \tag{11}$$

The capacity of the DP channel is given by simply setting $n = 2$ in (11).

Using (3), the MI of the channel for a uniformly distributed discrete input can be expressed as

$$I(\underline{\mathbf{s}};\underline{\mathbf{x}}) = \mathbb{E}_{\underline{\mathbf{s}},\underline{\mathbf{x}}}\left[\log\left(\frac{|\mathcal{S}|\exp(-\frac{\|\underline{\mathbf{x}}-\mathrm{H}\underline{\mathbf{s}}\|^2}{\sigma_z^2})}{\sum_{\underline{s}'\in\mathcal{S}}\exp(-\frac{\|\underline{\mathbf{x}}-\mathrm{H}\underline{s}'\|^2}{\sigma_z^2})}\right)\right], \tag{12}$$

where the expectations can be estimated numerically.

III. The Mutual information in presence of Channel Estimation Error

We derived the capacity of the DP channel with the assumption of perfect channel knowledge at the receiver in (11). In this section, we derive a lower bound on the MI of the channel in the presence of an estimation error.

A. Arbitrary channel, arbitrary estimate

As already shown in [48], when the channel is known, the capacity-achieving distribution is a zero-mean $\mathcal{CN}$ input with a power constraint. Thus, keeping the capacity-achieving distribution, i.e., $\underline{\mathbf{s}} \sim \mathcal{CN}(\underline{0},\Lambda_{\underline{s}})$ seems reasonable for the imperfectly estimated channel as well.

Many mismatched decoding metrics are available in the literature that provide lower bounds on the capacity, including generalized mutual information (GMI) [49] and the well-known LM rate [50], [51]. The LM rate is proven to be tight for binary input $|\mathcal{X}| = 2$ [52]. However, the LM rate is not tight in general [53]. As a lower bound on the MI (3) between $\underline{\mathbf{s}}$ and $\underline{\mathbf{x}}$, we use the GMI definition [49, Eq. (12)]

$$I(\underline{\mathbf{s}};\underline{\mathbf{x}}) \ge I_q = \sup_{\nu\ge 0}\ \mathbb{E}_{\underline{\mathbf{x}},\underline{\mathbf{s}}}\left[\log\left(\frac{q(\underline{\mathbf{x}}|\underline{\mathbf{s}})^\nu}{\mathbb{E}_{\underline{\mathbf{s}}}[q(\underline{\mathbf{x}}|\underline{\mathbf{s}})^\nu]}\right)\right], \tag{13}$$

where $q(\cdot)$ stands for the pdf of an auxiliary channel. Note that this inequality holds for an arbitrary distribution of $q(\cdot)$. For $\nu = 1$, (13) converges to the well-known mismatched decoding inequality [54, Eq. (34)]. The mismatched channel law is here assumed to be

$$q(\underline{x}|\underline{s}) = p(\underline{x}|\underline{s},\hat{\mathrm{H}}), \tag{14}$$

which is obtained by replacing H in (5) with the estimated channel $\hat{\mathrm{H}}$. This leads to

$$\begin{aligned}\mathbb{E}_{\underline{\mathbf{s}}}[q(\underline{x}|\underline{\mathbf{s}})^\nu] &= \mathbb{E}_{\underline{\mathbf{s}}}\left[\frac{1}{\pi^{n\nu}\sigma_z^{2n\nu}}\exp\left(-\frac{\left\|\underline{x}-\hat{\mathrm{H}}\underline{\mathbf{s}}\right\|^2}{\sigma_z^2/\nu}\right)\right] \\ &= \frac{\pi^n(\sigma_z^2/\nu)^n}{\pi^{n\nu}\sigma_z^{2n\nu}}\frac{1}{\pi^n|\hat{\Lambda}_x|}\exp\left(-\underline{x}^\dagger\hat{\Lambda}_x^{-1}\underline{x}\right),\end{aligned} \tag{15}$$

where

$$\hat{\Lambda}_{\underline{x}} = \hat{\mathrm{H}}\Lambda_{\underline{s}}\hat{\mathrm{H}}^\dagger + \frac{\sigma_z^2}{\nu}\mathrm{I}_n. \tag{16}$$

Substituting (14) and (15) in (13) gives

$$I_q = \sup_{\nu\ge 0}\mathbb{E}_{\underline{\mathbf{x}},\underline{\mathbf{s}}}\left[\log\left(\frac{\frac{1}{(\sigma_z^2/\nu)^n}\exp\left(-\frac{\|\underline{x}-\hat{\mathrm{H}}\underline{s}\|^2}{\sigma_z^2/\nu}\right)}{\frac{1}{|\hat{\Lambda}_x|}\exp\left(-\underline{x}^\dagger\hat{\Lambda}_x^{-1}\underline{x}\right)}\right)\right]. \tag{17}$$

The numerator and denominator inside the $\log(\cdot)$ in (17) are, by definition, $\mathcal{CN}(\hat{\mathrm{H}}\underline{s},\sigma_z^2/\nu\mathrm{I}_n)$ and $\mathcal{CN}(\underline{0},\hat{\Lambda}_x)$, respectively. This indicates that the optimization over ν for Gaussian channels is analogous to optimizing the noise variance of the auxiliary channel [55, Example 2].

The average AIR when the estimated channel is random can be written as

$$\bar{I}_q = \mathbb{E}_{\hat{\mathbf{H}}}[I_q]. \tag{18}$$

In the following, we first derive an AIR for the MIMO-AWGN channel model with a fixed channel H and estimated channel $\hat{\mathrm{H}}$. Then, we extend the derived AIR to n-dimensional unitary channels. By assuming a unitary estimate of the channel (i.e., $\hat{\mathrm{H}}\hat{\mathrm{H}}^\dagger = \mathrm{I}_n$), a tighter AIR is derived. Finally, we use (18) to consider a random estimated channel $\hat{\mathbf{H}}$.

Theorem 1: Consider an arbitrary complex MIMO channel matrix H and a fixed (deterministic) estimation error E. Defining the estimated channel matrix $\hat{\mathrm{H}} = \mathrm{H} - \mathrm{E}$, the AIR is

$$I_q = \sup_{\nu \geq 0} \Bigg(\log \left| \mathrm{I}_n + \nu \hat{\mathrm{H}} \mathrm{Q} \hat{\mathrm{H}}^\dagger \right| - \frac{1}{\ln 2} \nu \operatorname{tr}\left(\mathrm{Q}\mathrm{E}^\dagger \mathrm{E}\right) - \frac{1}{\ln 2} \operatorname{tr}\left(\nu \mathrm{I}_n - \Lambda_{\underline{x}} \hat{\Lambda}_{\underline{x}}^{-1}\right) \Bigg), \tag{19}$$

where the definitions of $\Lambda_{\underline{x}}$, Q, and $\hat{\Lambda}_{\underline{x}}$ can be found in (7), (9), and (16), respectively.

Proof: One may rewrite (17) as

$$I_q = \sup_{\nu \geq 0} \Bigg(\log \left(\frac{|\hat{\Lambda}_x|}{(\sigma_z^2/\nu)^n} \right) + \frac{1}{\ln 2} \overbrace{\mathbb{E}_{\underline{\mathbf{x}}} \left[\underline{x}^\dagger \hat{\Lambda}_x^{-1} \underline{x} \right]}^{A} \tag{20}$$

$$- \frac{1}{\ln 2} \overbrace{\frac{1}{\sigma_z^2/\nu} \mathbb{E}_{\underline{\mathbf{s}},\underline{\mathbf{x}}} \left[\left\| \underline{x} - \hat{\mathrm{H}} \underline{s} \right\|^2 \right]}^{B} \Bigg). \tag{21}$$

Using (7), given H, the first term A can be calculated as

$$\begin{aligned} A &= \mathbb{E}_{\underline{\mathbf{x}}} \left[\underline{\mathbf{x}}^\dagger \hat{\Lambda}_{\underline{x}}^{-1} \underline{\mathbf{x}} \right] \\ &= \operatorname{tr}\left(\Lambda_{\underline{x}} \hat{\Lambda}_{\underline{x}}^{-1} \right), \end{aligned} \tag{22}$$

where the last step follows by the cyclic permutation rule of the trace. For the second term, B, when H is given, we use (5) and (14) to obtain

$$\begin{aligned} B &= \frac{1}{\sigma_z^2/\nu} \mathbb{E}_{\underline{\mathbf{s}},\underline{\mathbf{x}}} \left[\left\| \underline{x} - \hat{\mathrm{H}} \underline{s} \right\|^2 \right] && (23) \\ &= \frac{1}{\sigma_z^2/\nu} \mathbb{E}_{\underline{\mathbf{z}},\underline{\mathbf{s}}} \left[\left\| (\mathrm{H} - \hat{\mathrm{H}}) \underline{\mathbf{s}} + \underline{\mathbf{z}} \right\|^2 \right] \\ &= \frac{1}{\sigma_z^2/\nu} \left(\mathbb{E}_{\underline{\mathbf{s}}} \left[\left\| \mathrm{E} \underline{\mathbf{s}} \right\|^2 \right] + \mathbb{E}_{\underline{\mathbf{z}}} \left[\left\| \underline{\mathbf{z}} \right\|^2 \right] \right) && (24) \\ &= \frac{1}{\sigma_z^2/\nu} \operatorname{tr}\left(\sigma_z^2 \mathrm{I}_n + \mathrm{E} \Lambda_{\underline{s}} \mathrm{E}^\dagger \right) && (25) \\ &= \nu \operatorname{tr}\left(\mathrm{I}_n + \mathrm{Q}\mathrm{E}^\dagger \mathrm{E} \right), && (26) \end{aligned}$$

where (24) follows from the fact that $\underline{\mathbf{z}}$ and $\underline{\mathbf{s}}$ are independent, and the cyclic permutation rule of the trace is used in (25). Finally, substituting (22) and (26) into (21) completes the proof. ∎

For discrete inputs, the pdf of the output of the auxiliary channel is

$$q(\underline{x}) = \sum_{\underline{s} \in \mathcal{S}} P(\underline{s}) q(\underline{x}|\underline{s}), \tag{27}$$

where $q(\underline{x}|\underline{s})$ is defined in (14). Using (13), (14), and (27), the average AIR for a uniformly distributed discrete input can be expressed as

$$\bar{I}_q \geq \mathbb{E} \left[\log \left(\frac{|\mathcal{S}| \exp(-\nu \frac{\left\| \underline{\mathbf{x}} - \hat{\mathbf{H}} \underline{\mathbf{s}} \right\|^2}{\sigma_z^2})}{\sum_{\underline{s}' \in \mathcal{S}} \exp(-\nu \frac{\left\| \underline{\mathbf{x}} - \hat{\mathbf{H}} \underline{s}' \right\|^2}{\sigma_z^2})} \right) \right], \tag{28}$$

where the expectation is over $\underline{\mathbf{s}}$, $\underline{\mathbf{x}}$, and $\hat{\mathbf{H}}$, which can be estimated numerically.

B. Unitary channel, arbitrary estimate

For the sake of keeping Theorem 1 as general as possible, no assumption is made about the transmitter knowledge of the estimated channel. However, in the context of the optical communications, we are more interested in the case that the transmitter has no channel knowledge. In that case, the transmitter may model the unknown channel matrix H as random, and if this random matrix has a symmetrical distribution with respect to the components of $\underline{\mathbf{s}}$, then a uniform power distribution $\Lambda_{\underline{s}} = P\mathrm{I}_n/n$ is optimal. For example, this is the case if the channel matrix is uniformly distributed over the set of all possible unitary channel matrices.

Corollary 1: Assume that the channel is unitary (i.e., $\mathrm{H} = \mathrm{H_u}$) and that uniform power distribution takes place at the transmitter. Then the AIR of a unitary channel can be written as

$$I_q = \sup_{\nu \geq 0} \Bigg(\log \left| \mathrm{I}_n + \nu\eta \hat{\mathrm{H}} \hat{\mathrm{H}}^\dagger \right| - \frac{\nu\eta}{\ln 2} \operatorname{tr}\left(\mathrm{E}^\dagger \mathrm{E}\right) - \frac{\nu}{\ln 2} \operatorname{tr}\left(\mathrm{I}_n - (1+\eta)(\mathrm{I}_n + \nu\eta \hat{\mathrm{H}} \hat{\mathrm{H}}^\dagger)^{-1} \right) \Bigg). \tag{29}$$

Proof: The uniform power distribution $\Lambda_{\underline{s}} = P\mathrm{I}_n/n$ yields $\mathrm{Q} = \eta \mathrm{I}_n$. Knowing the channel is unitary (i.e., $\mathrm{H_u}{\mathrm{H_u}}^\dagger = \mathrm{I}_n$) and using (7) and (16), we can write $\Lambda_{\underline{x}} = \sigma_z^2(1+\eta)\mathrm{I}_n$ and $\hat{\Lambda}_{\underline{x}}^{-1} = \sigma_z^{-2}\nu(\mathrm{I}_n + \nu\eta\hat{\mathrm{H}}\hat{\mathrm{H}}^\dagger)^{-1}$. Finally, substituting Q, $\Lambda_{\underline{x}}$, and $\hat{\Lambda}_x^{-1}$ in (19) completes the proof. ∎

The optimum parameter ν in (29) depends on η, $\hat{\mathrm{H}}$, and E. It can be computed numerically for each channel realization, derived analytically in a special case (see Theorem 2 below), or approximated (see Sec. V).

Corollary 2: Assume that the channel is a fixed unitary matrix $\mathrm{H_u}$, and that the estimated channel is an arbitrary random matrix $\hat{\mathbf{H}} = \mathrm{H_u} - \mathbf{E}$. Then with a uniform power distribution, the average AIR for any $\boldsymbol{\nu} = \nu(\eta, \hat{\mathrm{H}}, \mathrm{E})$, not necessarily optimum, is

$$\bar{I}_q \geq \mathbb{E}_{\hat{\mathbf{H}}} \left[\log \left| \mathrm{I}_n + \boldsymbol{\nu}\eta \hat{\mathbf{H}} \hat{\mathbf{H}}^\dagger \right| \right] - \frac{\eta}{\ln 2} \operatorname{tr}\left(\mathbb{E}_{\mathbf{E}}[\boldsymbol{\nu} \mathbf{E}^\dagger \mathbf{E}] \right) - \frac{1}{\ln 2} \operatorname{tr}\left(\mathbb{E}_{\hat{\mathbf{H}}} \left[\boldsymbol{\nu} \mathrm{I}_n - \boldsymbol{\nu}(1+\eta)(\mathrm{I}_n + \boldsymbol{\nu}\eta \hat{\mathbf{H}} \hat{\mathbf{H}}^\dagger)^{-1} \right] \right). \tag{30}$$

Proof: By applying (18) to (29) the proof is complete. ∎

The covariance matrix of the estimation error is defined as

$$\mathrm{R_E} = \mathbb{E}_{\mathbf{E}}[\mathbf{E}^\dagger \mathbf{E}], \tag{31}$$

which is an indication of the expected squared Euclidean distance between the estimated channel and the actual channel. Note that $\mathbb{E}_{\mathbf{E}}\left[\boldsymbol{\nu}\mathbf{E}^{\dagger}\mathbf{E}\right] = \nu \mathrm{R}_{\mathrm{E}}$ for constant ν.

An important special case is when $\mathbf{E}$ has a zero-mean spherically symmetric distribution. The following corollary specializes Corollary 2 to this important case, which is valid when the channel is estimated using pilots with orthogonal rows as will be shown in Section IV-A.

Corollary 3: If $\mathbf{E}$ has a zero-mean spherically symmetric distribution, then the right-hand side of (30) is the same for any $\mathrm{H_u}$.

Proof: For any two different $\mathrm{H_u}$ and $\mathrm{H_u}'$ one can define

$$\mathbf{E}' = \mathbf{E}\mathrm{H_u}^{\dagger}\mathrm{H_u}' = \mathrm{H_u}' - \hat{\mathbf{H}}', \tag{32}$$

where $\hat{\mathbf{H}}' = \hat{\mathbf{H}}\mathrm{H_u}^{\dagger}\mathrm{H_u}'$. Since $\mathbf{E}$ has a zero-mean spherically symmetric distribution, it is invariant to rotation. Thus, for any $\mathrm{H_u}$ and $\mathrm{H_u}'$, $\mathbf{E}'$ has the same distribution as $\mathbf{E}$. We can also write

$$\begin{aligned}\hat{\mathbf{H}}'\hat{\mathbf{H}}'^{\dagger} = \hat{\mathbf{H}}\hat{\mathbf{H}}^{\dagger} &= (\mathrm{H_u} - \mathbf{E})(\mathrm{H_u} - \mathbf{E})^{\dagger} \\ &= (\mathrm{I}_n - \mathbf{E}\mathrm{H_u}^{\dagger})(\mathrm{I}_n - \mathbf{E}\mathrm{H_u}^{\dagger})^{\dagger},\end{aligned} \tag{33}$$

and with the same logic, for any $\mathrm{H_u}$, $\mathbf{E}\mathrm{H_u}^{\dagger}$ has the same distribution as $\mathbf{E}$ and $\hat{\mathbf{H}}\hat{\mathbf{H}}^{\dagger}$ has the same distribution as $(\mathrm{I}_n - \mathbf{E})(\mathrm{I}_n - \mathbf{E})^{\dagger}$. Since $\mathbf{E}$ is independent of the channel realization H_u, (30) yields the same AIR independently of $\mathrm{H_u}$. ∎

C. Arbitrary channel, unitary estimate

We have not made any assumption on the estimation technique, so the derived lower bounds hold for an arbitrary estimator. It can be seen that (30) highly depends on the choice of estimation technique, so one can tighten the bound by choosing a suitable estimator.

Corollary 4: For a complex unitary channel H_u, a unitary estimated channel $\hat{\mathbf{H}} = \mathrm{H}_u - \mathbf{E}$, and a uniform power distribution, an AIR is

$$\begin{aligned}\bar{I}_q \geq\ & n\mathbb{E}_{\mathbf{E}}\left[\log(1+\boldsymbol{\nu}\eta)\right] \\ & - \frac{\eta}{\ln 2}\left(\mathrm{tr}(\mathbb{E}_{\mathbf{E}}[\boldsymbol{\nu}\mathbf{E}^{\dagger}\mathbf{E}]) + n\mathbb{E}_{\mathbf{E}}\left[\frac{\boldsymbol{\nu}(\boldsymbol{\nu}-1)}{1+\boldsymbol{\nu}\eta}\right]\right).\end{aligned} \tag{34}$$

Proof: By applying $\hat{\mathbf{H}}\hat{\mathbf{H}}^{\dagger} = \mathrm{I}_n$ to (30) and simplifying the obtained expression, the proof is complete. ∎

Note that (34) is independent of H, meaning that the AIR is the same for any unitary channel. Interestingly, the unitary estimation of the channel leads to a simpler bound.

In general, optimizing ν for each realization of the estimated channel is a computationally demanding procedure. However, when both the actual and the estimated channels are unitary, the optimal ν value can be analytically obtained as a function of η and E.

Theorem 2: Assume a complex unitary channel with uniform power distribution and define $\kappa = \mathrm{tr}(\mathrm{E}^{\dagger}\mathrm{E})/n$. Then, for each realization of a unitary estimated channel, the supremum in Corollary 1 is obtained for

$$\nu = \nu^* = \begin{cases} \frac{\sqrt{(4\kappa+1)\eta^2+4(\kappa+1)\eta+4}+\eta(1-2\kappa)-2}{2\eta(\kappa\eta+1)} & 0 \leq \kappa \leq 2 \\ 0 & \kappa \geq 2 \end{cases} \tag{35}$$

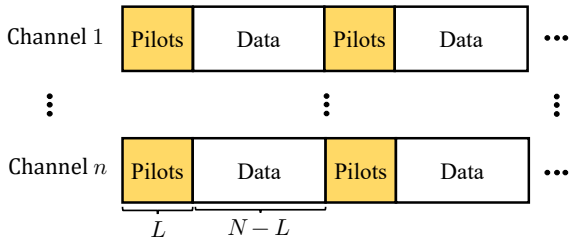

Fig. 1: Transmission block model

Proof: See Appendix. ∎

IV. Channel Estimation

In this section, first, the well-established LS estimation algorithm is presented. Then, a unitary estimation method is proposed for unitary channels. As illustrated in Fig. 1, to make data-aided channel estimation possible, the first L symbols of the transmission block are forming the $n \times L$ matrix of pilot symbols $\mathrm{D} = [\underline{s}_0, ..., \underline{s}_{L-1}]$, which are assumed to be known at both transmitter and receiver. When the total transmission power is constrained, the optimal pilot assignment D should have the following properties [56]:

- $L \geq n$,
- $\mathrm{DD}^{\dagger} = \frac{PL}{n}\mathrm{I}_n$,

where the first condition states that in each row of the channel matrix there are n unknown complex parameters that must be estimated, and to ensure the uniqueness of the estimation, at least n pilots are needed in each row of the pilot matrix. The second condition states that any training matrix with orthogonal rows of the same norm $\sqrt{PL/n}$ is optimal [57], [58]. Note that infinitely many pilot sequences satisfy the above conditions. The complex matrix $\mathbf{X} = [\underline{\mathbf{x}}_0, ..., \underline{\mathbf{x}}_{L-1}]$ of received symbols is

$$\mathbf{X} = \mathbf{H}_{\mathrm{u}}\mathrm{D} + \mathbf{Z}, \tag{36}$$

where $\mathbf{Z} = [\underline{\mathbf{z}}_0, ..., \underline{\mathbf{z}}_{L-1}]$ is an $n \times L$ matrix of i.i.d. noise samples which are assumed to be $\mathcal{CN}(\underline{0}, \sigma_z^2\mathrm{I}_n)$, and $\mathbf{H}_{\mathrm{u}}$ is an $n \times n$ random unitary channel matrix, which is assumed to remain constant during a transmission block.

For the channel estimation problem, we define the number of degrees of freedom (DOF) as the number of independent real values that are needed to be estimated and denoted by Ω.

A. LS Algorithm

Conventional optical transmission systems often use LMS for a real-time estimate of the channel, because LMS tracks the channel change with each received symbol. However, in this paper it is assumed that the channel is constant during a transmission block and the LS method is well adapted to block based transmissions. The LS estimator estimates the channel by minimizing the squared error between the desired and received signal. For a MIMO-AWGN channel, the LS optimization problem can be expressed as [59]

$$\min_{\hat{\mathbf{H}}} \left\| \mathbf{X} - \hat{\mathbf{H}}\mathrm{D} \right\|^2. \tag{37}$$

The LS optimization problem (37) can be translated to finding $\Omega = 2n^2$ independent real values. Knowing D and $\mathbf{X}$, the solution of (37) is [59], [60, Ch. 8]

$$\hat{\mathbf{H}}_{\text{LS}} = \mathbf{X}\mathrm{D}^\dagger(\mathrm{D}\mathrm{D}^\dagger)^{-1}. \tag{38}$$

However, the pilots are chosen in such a way that $\mathrm{D}\mathrm{D}^\dagger = \frac{PL}{n}\mathrm{I}_n$. Therefore, we can write

$$\hat{\mathbf{H}}_{\text{LS}} = \frac{n}{PL}\mathbf{X}\mathrm{D}^\dagger = \mathrm{H}_{\text{u}} + \frac{n}{PL}\mathbf{Z}\mathrm{D}^\dagger. \tag{39}$$

Thus, the estimation error matrix $\mathbf{E} = -\frac{n}{PL}\mathbf{Z}\mathrm{D}^\dagger$ has a zero-mean spherically symmetric distribution, and hence Corollary 3 is applicable. This is because multiplying $\mathbf{Z}$ and a matrix with orthogonal rows $\mathrm{D}^\dagger$ does not change the distribution. It can be shown that for the LS algorithm [59]

$$\mathrm{R}_{\text{E}} = \mathbb{E}_{\mathbf{E}}\left[\mathbf{E}^\dagger\mathbf{E}\right] = \frac{n\sigma_z^2}{PL}\mathrm{I}_n = \frac{1}{\eta L}\mathrm{I}_n \tag{40}$$

showing that the estimation error of the LS algorithm is inversely proportional to the SNR and the pilot length L. The implementation complexity of the LS algorithm is discussed in Section IV-C.

B. Kabsch Algorithm

The problem with both blind estimation algorithms (e.g., CMA, RDE, and MMA) and pilot-aided estimation algorithms (e.g., LS and LMS) is that their optimization problems are not adopted for unitary channel estimation, making their solutions suboptimal if $\mathbf{H}$ is known to be unitary. Thus, in this part, we apply the unitary constraint of the channel to the estimation problem of (37) and write

$$\min_{\hat{\mathbf{H}}} \left\| \mathbf{X} - \hat{\mathbf{H}}\mathrm{D} \right\|^2 \quad \text{s.t.} \quad \hat{\mathbf{H}}\hat{\mathbf{H}}^\dagger = \mathrm{I}_n. \tag{41}$$

The optimal solution to this problem is given by the Kabsch algorithm [29] as

$$\hat{\mathbf{H}}_{\text{Kabsch}} = \mathbf{U}\mathbf{V}^\dagger, \tag{42}$$

where $\mathbf{U}\Sigma\mathbf{V}^\dagger$ is the singular value decomposition function of $\mathbf{X}\mathrm{D}^\dagger$. Since the channel is estimated from the pilots, unlike the conventional blind estimators, the Kabsch algorithm is not influenced by the choice of modulation format for the actual data transmission, which may be different from that of the pilot symbols. The Kabsch algorithm was proposed for optical communication by Louchet *et al.* [28] as a blind polarization tracking algorithm, where decision-directed symbols were used instead of pilots.

In contrast to LS, no analytical result is known for R_{E} of the Kabsch algorithm. Although we cannot analytically prove it, we can make an intuitive prediction by considering that for a unitary estimation problem (41) $\Omega = n^2$ while for a general estimation problem (37) $\Omega = 2n^2$. Since the channel estimation problem is equivalent to finding Ω independent real-valued quantities, one can predict that the estimation error of Kabsch would be half of LS. More interestingly, for special unitary channels where $|\mathrm{H}_{\text{u}}| = 1$, the DOF is $n^2 - 1$ and the gain by the unitary algorithm can be even higher; however, this gain vanishes for large n.

Note that in the case of the DP channel, the singular value decomposition is deployed only on a two-by-two matrix, making it less computationally complex than for higher n. The computational complexity of the Kabsch algorithm is analyzed in Section IV-C.

C. Implementation Complexity

The implementation complexity of the two considered algorithms can be assessed by comparing their scalability in the dimension n and pilot length L. The LS algorithm has only one matrix multiplication $\mathbf{X}\mathbf{D}^\dagger$ (39), which scales as $O(n^2L)$. However, the Kabsch algorithm, in addition to computing $\mathbf{X}\mathbf{D}^\dagger$, needs to calculate one extra $n \times n$ matrix multiplication (42) and the singular value decomposition. Therefore, the Kabsch algorithm scales as $O(2n^3 + n^2L)$; however, as mentioned in Sec. IV, $L \geq n$, and thus the Kabsch algorithm scales as $O(2n^3 + n^2L) = O(n^2L)$, which surprisingly is the same scaling as with LS. Moreover, for the DP channel ($n = 2$), the pilot length determines L and the implementation complexity of both algorithms.

V. Numerical Results

In this section, through Monte Carlo trials, the AIR of the DP channel (i.e., $n = 2$) for the estimation algorithms detailed in Section IV is computed. While AIRs are derived for a fixed channel matrix H, the estimated channel $\hat{\mathbf{H}}$ is dependent on each realization of the channel. For comparison, the BW-DDLMS algorithm [22] is also investigated. Due to unreliable convergence of BW-DDLMS, the BW-CMA algorithm [22] is used in a pre-convergence phase after which we switch to BW-DDLMS. For both BW-CMA and BW-DDLMS, the block length is set to 1000 symbols, the step size is fixed to 10^{-3}, and the number of iterations is set to 5 and 20, respectively. Refer to [22, Section III] for a more detailed implementation description. A deterministic sequence of L quadrature phase-shift keying (QPSK) symbols is selected to satisfy the pilot conditions detailed in Section IV.

Numerical results verify that the estimation error of the Kabsch algorithm completely follows our prediction in Sec. IV-B. Thus, to perform a fair comparison between the unitary and nonunitary estimators, we define a new parameter called estimation error per DOF as $\mathcal{E}^2 = \operatorname{tr}\mathrm{R}_{\text{E}}/(n\Omega)$. Note that for a general nonunitary estimator $\Omega = 2n^2$ and for a general unitary estimator $\Omega = n^2$.

Using $\mathcal{CN}$ inputs, Fig. 2 shows the Monte-Carlo averaged AIR $\bar{I}_q$ of the unitary DP channel as a function of η and $\mathcal{E}^2$ according to (30) and (34) for $n = 2$. Each Monte Carlo AIR value is generated as follows.

- For the dashed curves, a single unitary H_{u} is picked, multiple $\mathbf{E}$ are generated from $\mathcal{CN}(\underline{0}, \mathrm{R}_{\text{E}})$ where $\mathrm{R}_{\text{E}} = 2n^2\mathcal{E}^2\mathrm{I}_n$ and $\hat{\mathbf{H}} = \mathrm{H}_{\text{u}} - \mathbf{E}$ is formed. Finally, (30) is averaged over $\mathbf{E}$. Note that based on Corollary 3, for an estimation error $\mathbf{E}$ with a spherically symmetric distribution, (30) is independent of H_{u}. Since optimizing ν for each $\mathbf{E}$ realization is computationally demanding, ν is approximated with ν^* (35), which is a suboptimal value for (30).

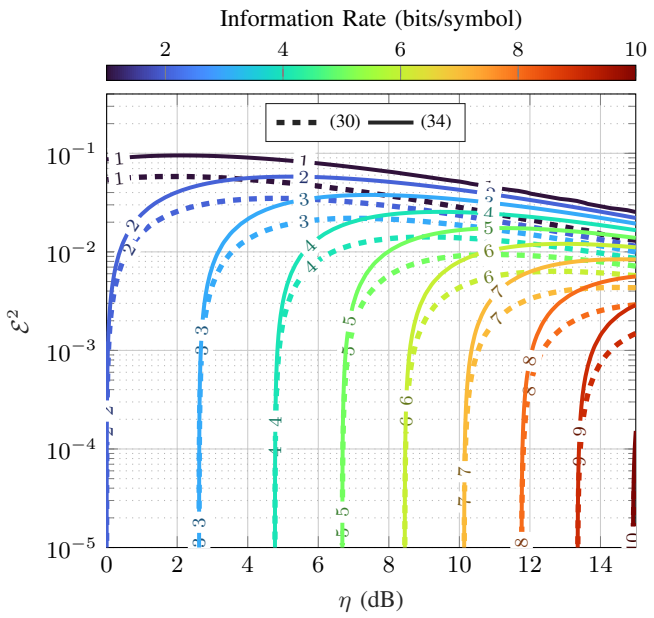

Fig. 2: The average AIR $\bar{I}_q$ of the unitary DP channel as a function of SNR η and estimation error per DOF $\mathcal{E}^2$ when $\nu = \nu^*$. Normally, a high SNR is associated with a low $\mathcal{E}^2$ and vice versa, depending on the applied estimation algorithms.

- For the solid curves, multiple $\mathbf{E}$ are generated from $\mathcal{CN}(\underline{0}, \mathrm{R_E})$ where $\mathrm{R_E} = n^2\mathcal{E}^2\mathrm{I}_n$. For each $\mathbf{E}$ realization, the ν parameter is optimized (i.e., $\nu = \nu^*$) using (35); then, (34) is averaged over $\mathbf{E}$.

The results illustrate that for each estimation error per DOF $\mathcal{E}^2$, there exists a specific η which maximizes the average AIR. This can be justified according to the right-hand sides of (30) and (34), where the first term increases in a logarithmic manner with respect to SNR, but the second term decreases linearly with SNR. Therefore there is an optimum SNR that maximizes the AIR.

With $\mathcal{CN}$ inputs in use, Fig. 3(a) presents the AIR of LS and Kabsch according to (30) and (34), respectively. The solid red line indicates the case when the receiver perfectly knows the channel. It can be seen that choosing $\nu = \nu^*$ improves the AIR with respect to the constant $\nu = 1$ (i.e., the mismatched decoding lower bound). Although ν^* is suboptimal for the LS algorithm, it still offers a noticeable improvement. This is important because ν^* is computed analytically and comes with almost no extra computational complexity. It can also be concluded that with $L = 8$, Kabsch surpasses LS throughout the range of considered SNRs. More specifically, at 4 and 14 dB SNRs (see insets), Kabsch has at least 0.20 and 0.30 bits per symbol higher AIR, respectively. Additionally, it is clear from the results that LS is upper-bounded by Kabsch, and increasing the SNR cannot fill the gap. This behavior can be justified because, unlike LS, Kabsch guarantees a unitary estimation of the channel leading to a lower estimation error. Moreover, as the SNR increases, the gap between Kabsch and LS and the actual channel MI is almost constant. Since the theoretical gap is $(\eta/\ln 2)\,\mathrm{tr}(\mathrm{R_E})$ according to Corollary 4, we conclude that the error covariance $\mathrm{R_E}$ of Kabsch is inversely proportional to the SNR, i.e., $\mathrm{R_E} \propto 1/\eta$. Unlike Fig. 2, the AIR bounds are monotonically increasing with SNR which is due to the fact that the estimation error is decreasing with SNR.

A comparison between LS, Kabsch, and BW-DDLMS with DP-16-QAM inputs and $\nu = 1$ is provided in Fig. 3(b). The MI when the channel is perfectly known at the receiver is marked by the dash-dotted red line. Evidently, Kabsch outperforms LS throughout the considered range of SNRs. The results also support the fact that Kabsch upper-bounds LS for various inputs, which completely agrees with Fig. 2, where the unitary estimation of the channel leads to a higher AIR. It can also be seen that BW-DDLMS outperforms both LS and Kabsch in the entire considered SNR range. However, this performance comes with significantly higher complexity than the LS and Kabsch algorithms. The BW-DDLMS algorithm is carried out on a block of 1000 symbols for 15 iterations, while the data-aided algorithms are executed on only 8 symbols in each block. Besides, by setting $L = 32$, both LS and Kabsch will outperform BW-DDLMS.

Fig. 4 displays the information gap between the AIRs and the capacity of the DP channel, when (a) $\mathcal{CN}$ inputs with $\nu = 1$ and $\nu = \nu^*$ and (b) DP-16-QAM inputs are used and $\nu = 1$. It can be seen that setting $\nu = \nu^*$ gives a lower information gap for both algorithms. The overall dominance of Kabsch throughout the range of considered SNRs can be easily verified. Evidently, it is beneficial to use higher pilot numbers at low SNRs. Given the fact that $\mathrm{R_E} \propto 1/(L\eta)$ for LS (40) and that Kabsch in Fig. 4 follows the same trend with respect to L, we can empirically conclude that $\mathrm{R_E}$ of Kabsch is also inversely proportional to the pilot length and the SNR. It is also interesting to see that for Gaussian inputs, unlike the LS algorithm, the information gap of the Kabsch algorithm is independent of SNR (see Fig. 4(a)), which is explained by Corollary 4 and $\mathrm{R_E} \propto 1/(L\eta)$. On the contrary, for discrete inputs, the information gap decreases at higher SNRs (see blue dashed lines in Fig. 4(b)). This is because the mutual information I and the average AIR $\bar{I}_q$ both converge to the same asymptotic value at high SNR η, as shown in Fig. 3(b).

It is beneficial to use Kabsch to limit the rate loss due to the pilots. For example, a reasonable performance is achieved by only a block of $L = 16$ pilot symbols, which is relatively small compared to the transmission block size in optical communication systems, implying that the rate loss is negligible. For instance, for a system operating at a rate of 28 Gbaud, even if the channel remains constant for one microsecond (i.e., the SOP drift time is one microsecond), it corresponds to a block length N of at least $28,000$ symbols and the rate loss due to 16 pilots is negligible.

VI. Conclusion

The capacity of unitary MIMO-AWGN channels and specifically DP was investigated, with applications to dual-polarization transmission in optical fibers. An AIR with imperfect channel knowledge was derived and showed that the AIR is highly dependent on the estimation algorithm. In the

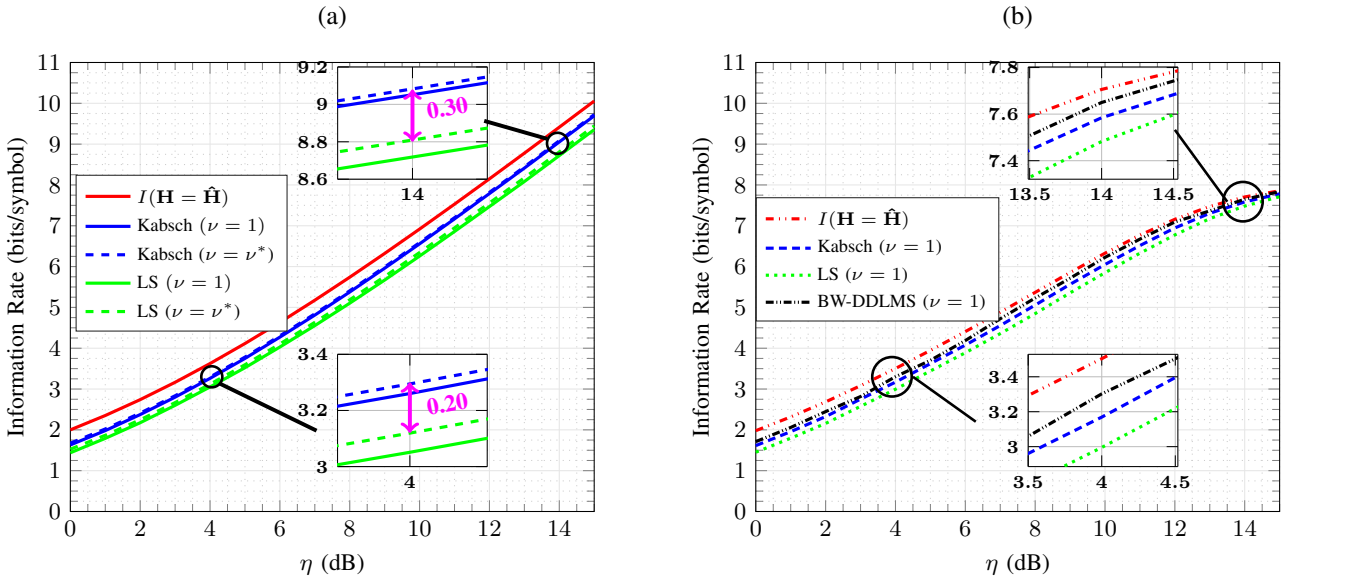

Fig. 3: The average AIR $\bar{I}_q$ when $L=8$ DP pilot symbols are used. (a) Using $\mathcal{CN}$ inputs where for LS and Kabsch, $\bar{I}_q$ is according to (30) and (34), respectively. (b) Using uniformly distributed DP-16-QAM inputs where $\bar{I}_q$ for LS, Kabsch, and BW-DDLMS [22] is according to (28).

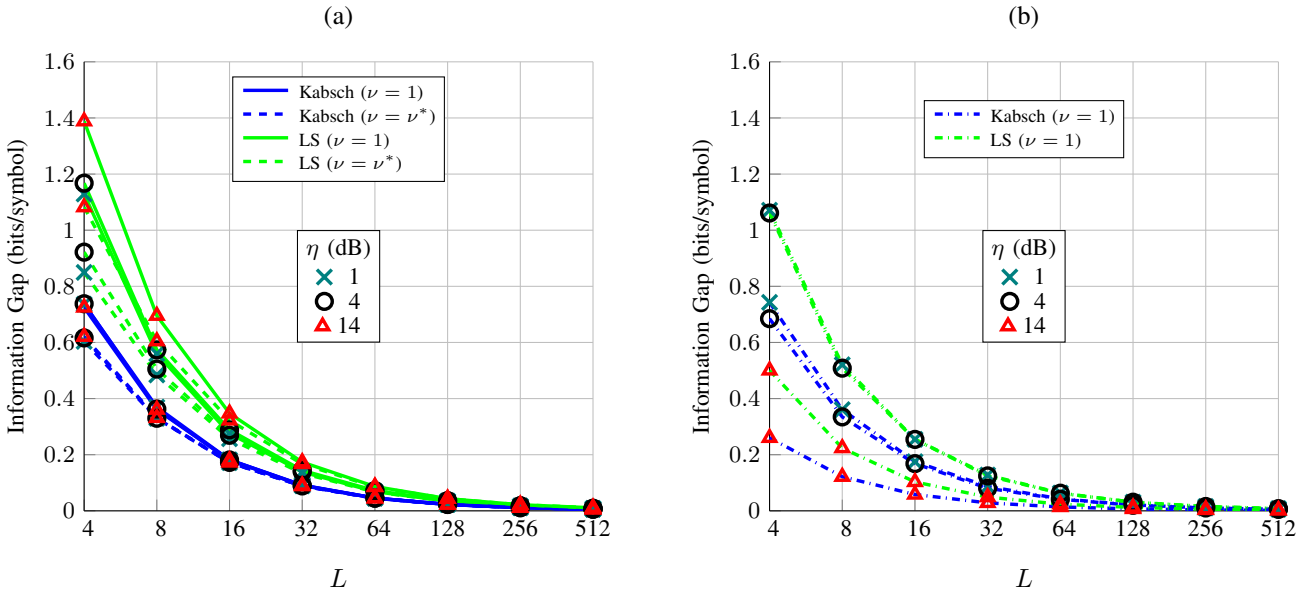

Fig. 4: The average information gap $I-\bar{I}_q$ for a range of pilot lengths. (a) $\mathcal{CN}$ inputs where $\bar{I}_q$ for LS and Kabsch algorithms is according to (30) and (34), respectively. (b) Uniformly distributed DP-16-QAM inputs where $\bar{I}_q$ for both LS and Kabsch is according to (28).

case of unitary channels, higher AIRs are obtained with a unitary estimation of the channel where the optimum ν value is obtained with a closed-form expression. The bounds are derived for any $n \geq 2$ dimensions, meaning that the results can be applied to other optical channels. In particular, Theorem 1 can be directly applied to space-division multiplexed channels which are impaired with polarization- and mode-dependent loss. Although it is shown that the unitary estimation of the channel yields a higher AIR, the majority of the presently deployed estimation algorithms in the optical communications literature are designed for general nonunitary channels. Therefore, the data-aided Kabsch algorithm is proposed to ensure a unitary estimate of the channel. Numerical results showed that for various input distributions, Kabsch outperforms LS in terms of AIR. Also, like LS, Kabsch can perform very well with only a few pilot symbols, making the transmission rate loss due to the pilots negligible.

As a future work, considering more real-world impairments (e.g., PMD, chromatic dispersion, and nonlinearities) may lead to more realistic bounds.

APPENDIX
PROOF OF THEOREM 2

Applying $\hat{\mathrm{H}}\hat{\mathrm{H}}^{\dagger} = \mathrm{I}_2$ to (29) gives

$$I_q = \sup_{\nu \geq 0} n \log(1+\nu\eta) - \frac{n\nu\eta}{\ln 2}\left(\kappa + \frac{\nu-1}{1+\nu\eta}\right) \tag{43}$$

$$= -\frac{n\eta}{\ln 2} \inf_{\nu \geq 0} \tilde{I}_q(\nu), \tag{44}$$

where

$$\tilde{I}_q(\nu) = -\frac{\ln 2}{\eta}\log(1+\nu\eta) + \kappa\nu + \frac{\nu^2-\nu}{1+\nu\eta}. \tag{45}$$

The first derivative of (45) is

$$\frac{\partial \tilde{I}_q(\nu)}{\partial \nu} = \kappa + \frac{(\nu\eta+2)(\nu-1)}{(1+\nu\eta)^2} \tag{46}$$

and the second derivative is

$$\frac{\partial^2 \tilde{I}_q(\nu)}{\partial \nu^2} = \frac{\eta^2\nu + 3\eta + 2}{(1+\nu\eta)^3} > 0 \qquad \forall \nu \geq 0. \tag{47}$$

Thus, (46) is monotonically increasing with ν, and hence it is minimum at $\nu = 0$, yielding

$$\frac{\partial \tilde{I}_q(\nu)}{\partial \nu} \geq \kappa - 2. \tag{48}$$

For $\kappa > 2$, (46) is always positive, and hence (45) is monotonically increasing with ν. Thus, the optimal ν is zero ($\nu^* = 0$) yielding $I_q = 0$.

For $0 \leq \kappa \leq 2$, we equate (46) to zero which yields

$$\overbrace{(\kappa\eta^2+\eta)}^{a}\nu^2 + \overbrace{(2\kappa\eta - \eta + 2)}^{b}\nu + \overbrace{(\kappa-2)}^{c} = 0, \tag{49}$$

where $a > 0$ and since

$$b^2 - 4ac = (4\kappa+1)\eta^2 + 4(\kappa+1)\eta + 4 > 4, \tag{50}$$

both roots of (49) are real. Since $c \leq 0$, (49) has only one nonnegative root, which is

$$\nu^* = \frac{-b + \sqrt{b^2-4ac}}{2a} \tag{51}$$

and since the second derivative (47) is positive, ν^* is indeed a minimum of (45). Therefore,

$$\nu^* = \begin{cases} \frac{\sqrt{(4\kappa+1)\eta^2+4(\kappa+1)\eta+4}+\eta(1-2\kappa)-2}{2\eta(\kappa\eta+1)} & 0 \leq \kappa \leq 2 \\ 0 & \kappa > 2 \end{cases} \tag{52}$$

and the proof is completed.

VII. ACKNOWLEDGMENT

The authors are grateful to the Editor Dr. Gianluigi Liva for helpful suggestions.

REFERENCES

[1] D. Monroe, "Optical fibers getting full," *Commun. ACM*, vol. 59, no. 10, pp. 10–12, Sep. 2016.

[2] E. Agrell, A. Alvarado, and F. R. Kschischang, "Implications of information theory in optical fiber communications," *Philos. Trans. R. Soc. A*, vol. 374, Mar. 2016.

[3] R.-J. Essiambre, G. J. Foschini, P. J. Winzer, G. Kramer, and B. Goebel, "Capacity limits of optical fiber networks," *J. Lightwave Technol.*, vol. 28, pp. 662–701, Feb. 2010.

[4] S. J. Savory, "Digital coherent optical receivers: Algorithms and subsystems," *IEEE J. Sel. Topics Quantum Electron.*, vol. 16, no. 5, pp. 1164–1179, Sep.-Oct. 2010.

[5] C. D. Poole and C. R. Giles, "Polarization-dependent pulse compression and broadening due to polarization dispersion in dispersion-shifted fiber," *Opt. Lett.*, vol. 13, no. 2, pp. 155–157, Feb. 1988.

[6] M. Karlsson, J. Brentel, and P. A. Andrekson, "Long-term measurement of PMD and polarization drift in installed fibers," *J. Lightwave Technol.*, vol. 18, no. 7, pp. 941–951, Jul. 2000.

[7] C. T. Allen, P. K. Kondamuri, D. L. Richards, and D. C. Hague, "Measured temporal and spectral PMD characteristics and their implications for network-level mitigation approaches," *J. Lightwave Technol.*, vol. 21, no. 1, pp. 79–86, Jan. 2003.

[8] D. Waddy, P. Lu, L. Chen, and X. Bao, "Fast state of polarization changes in aerial fiber under different climatic conditions," *IEEE Photon. Technol. Lett.*, vol. 13, no. 9, pp. 1035–1037, Sep. 2001.

[9] P. M. Krummrich, D. Ronnenberg, W. Schairer, D. Wienold, F. Jenau, and M. Herrmann, "Demanding response time requirements on coherent receivers due to fast polarization rotations caused by lightning events," *Opt. Express*, vol. 24, no. 11, pp. 12 442–12 457, May 2016.

[10] Z. Zhang, X. Bao, Q. Yu, and L. Chen, "Fast state of polarization and PMD drift in submarine fibers," *IEEE Photon. Technol. Lett.*, vol. 18, no. 9, pp. 1034–1036, May 2006.

[11] D. Godard, "Self-recovering equalization and carrier tracking in two-dimensional data communication systems," *IEEE Trans. on Commun.*, vol. 28, no. 11, pp. 1867–1875, Nov. 1980.

[12] S. J. Savory, G. Gavioli, R. I. Killey, and P. Bayvel, "Transmission of 42.8 Gbit/s polarization multiplexed NRZ-QPSK over 6400 km of standard fiber with no optical dispersion compensation," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, 2007, p. OTuA1.

[13] S. J. Savory, "Digital filters for coherent optical receivers," *Opt. Express*, vol. 16, no. 2, pp. 804–817, Jan. 2008.

[14] K. Kikuchi, "Polarization-demultiplexing algorithm in the digital coherent receiver," in *Proc. IEEE/LEOS Summer Topical Meetings*, 2008, pp. 101–102.

[15] ——, "Performance analyses of polarization demultiplexing based on constant-modulus algorithm in digital coherent optical receivers," *Opt. Express*, vol. 19, no. 10, pp. 9868–9880, May 2011.

[16] L. Liu, Z. Tao, W. Yan, S. Oda, T. Hoshida, and J. C. Rasmussen, "Initial tap setup of constant modulus algorithm for polarization demultiplexing in optical coherent receivers," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, Mar. 2009, p. OMT2.

[17] M. S. Faruk, Y. Mori, C. Zhang, and K. Kikuchi, "Proper polarization demultiplexing in coherent optical receiver using constant modulus algorithm with training mode," in *Proc. OptoElectron. Commun. Conf. (OECC)*, 2010, pp. 768–769.

[18] M. Ready and R. Gooch, "Blind equalization based on radius directed adaptation," in *Proc. Int. Conf. Acoust. Speech Sig. Process. (ICASSP)*, vol. 3, Apr. 1990, pp. 1699–1702.

[19] I. Fatadin, D. Ives, and S. J. Savory, "Blind equalization and carrier phase recovery in a 16-QAM optical coherent system," *J. Lightwave Technol.*, vol. 27, no. 15, pp. 3042–3049, Aug. 2009.

[20] D. Lavery, M. Paskov, R. Maher, S. J. Savory, and P. Bayvel, "Modified radius directed equaliser for high order QAM," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, Sep. 2015, p. We4D5.

[21] J. Yang, J.-J. Werner, and G. Dumont, "The multimodulus blind equalization and its generalized algorithms," *IEEE J. Sel. Areas Commun.*, vol. 20, no. 5, pp. 997–1015, Jun. 2002.

[22] M. Selmi, C. Gosset, M. Noelle, P. Ciblat, and Y. Jaouen, "Block-wise digital signal processing for POLMUX QAM/PSK optical coherent systems," *J. Lightwave Technol.*, vol. 29, no. 20, pp. 3070–3082, 2011.

[23] C. R. S. Fludger, T. Duthel, D. van den Borne, C. Schulien, E.-D. Schmidt, T. Wuth, J. Geyer, E. De Man, G.-D. Khoe, and H. de Waardt, "Coherent equalization and POLMUX-RZ-DQPSK for robust 100-GE transmission," *J. Lightwave Technol.*, vol. 26, no. 1, pp. 64–72, Jan. 2008.

[24] I. Barhumi, G. Leus, and M. Moonen, "Optimal training design for MIMO OFDM systems in mobile wireless channels," *IEEE Trans. Signal Process.*, vol. 51, no. 6, pp. 1615–1624, June 2003.

[25] E. Ip and J. M. Kahn, "Digital equalization of chromatic dispersion and polarization mode dispersion," *J. Lightwave Technol.*, vol. 25, no. 8, pp. 2033–2043, Aug. 2007.

[26] W. Shieh, "Maximum-likelihood phase and channel estimation for coherent optical OFDM," *IEEE Photon. Technol. Letters*, vol. 20, no. 8, pp. 605–607, Apr. 2008.

[27] M. Kuschnerov, M. Chouayakh, K. Piyawanno, B. Spinnler, E. de Man, P. Kainzmaier, M. S. Alfiad, A. Napoli, and B. Lankl, "Data-aided versus blind single-carrier coherent receivers," *IEEE Photon. Journal*, vol. 2, no. 3, pp. 387–403, June 2010.

[28] H. Louchet, K. Kuzmin, and A. Richter, "Joint carrier-phase and polarization rotation recovery for arbitrary signal constellations," *IEEE Photon. Technol. Lett.*, vol. 26, no. 9, pp. 922–924, May 2014.

[29] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, vol. 32, no. 5, pp. 922–923, Sep. 1976.

[30] C. B. Czegledi, E. Agrell, M. Karlsson, and P. Johannisson, "Modulation format independent joint polarization and phase tracking for coherent receivers," *J. Lightwave Technol.*, vol. 34, no. 14, pp. 3354–3364, Jul. 2016.

[31] A. Splett, C. Kurtzke, and K. Petermann, "Ultimate transmission capacity of amplified optical fiber communication systems taking into account fiber nonlinearities," in *Proc. Eur. Conf. Opt. Commun. (ECOC)*, 1994, pp. 41–44.

[32] C. Kurtzke, "Kapazitätsgrenzen digitaler optischer Übertragungssysteme," Doctoral Thesis (in German), Technische Universität Berlin, Berlin, Germany, 1995. [Online]. Available: http://doi.org/10.14279/depositonce-5080

[33] A. G. Green, P. B. Littlewood, P. P. Mitra, and L. G. L. Wegener, "Schrödinger equation with a spatially and temporally random potential: Effects of cross-phase modulation in optical communication," *Phys. Rev. E*, vol. 66, p. 046627, Oct. 2002.

[34] P. P. Mitra and J. B. Stark, "Nonlinear limits to the information capacity of optical fiber communications," *Nature*, vol. 411, pp. 1027–1030, June 2001.

[35] E. Agrell, A. Alvarado, G. Durisi, and M. Karlsson, "Capacity of a nonlinear optical channel with finite memory," *J. Lightwave Technol.*, vol. 32, no. 16, pp. 2862–2876, Aug. 2014.

[36] M. Secondini and E. Forestieri, "Analytical fiber-optic channel model in the presence of cross-phase modulation," *IEEE Photon. Technol. Lett.*, vol. 24, no. 22, pp. 2016–2019, Nov. 2012.

[37] M. Secondini, E. Forestieri, and G. Prati, "Achievable information rate in nonlinear WDM fiber-optic systems with arbitrary modulation formats and dispersion maps," *J. Lightwave Technol.*, vol. 31, no. 23, pp. 3839–3852, Dec. 2013.

[38] M. Secondini and E. Forestieri, "On XPM mitigation in WDM fiber-optic systems," *IEEE Photon. Technol. Lett.*, vol. 26, no. 22, pp. 2252–2255, Nov. 2014.

[39] R. Dar, M. Feder, A. Mecozzi, and M. Shtaif, "Time varying ISI model for nonlinear interference noise," in *Proc. Opt. Fiber Commun. Conf. (OFC)*, Mar. 2014, p. W2A.62.

[40] ——, "Inter-channel nonlinear interference noise in WDM systems: Modeling and mitigation," *J. Lightwave Technol.*, vol. 33, no. 5, pp. 1044–1053, Mar. 2015.

[41] R. Dar, M. Shtaif, and M. Feder, "New bounds on the capacity of the nonlinear fiber-optic channel," *Opt. Lett.*, vol. 39, no. 2, pp. 398–401, Jan. 2014.

[42] K. Kikuchi, "Polarization-demultiplexing algorithm in the digital coherent receiver," in *Proc. IEEE/LEOS Summer Topical Meetings*, 2008, p. MC2.2.

[43] I. Roudas, A. Vgenis, C. S. Petrou, D. Toumpakaris, J. Hurley, M. Sauer, J. Downie, Y. Mauro, and S. Raghavan, "Optimal polarization demultiplexing for coherent optical communications systems," *J. Lightwave Technol.*, vol. 28, no. 7, pp. 1121–1134, 2009.

[44] F. Heismann and K. Tokuda, "Polarization-independent electro-optic depolarizer," *Opt. Lett.*, vol. 20, no. 9, pp. 1008–1010, 1995.

[45] N. J. Muga and A. N. Pinto, "Adaptive 3-D Stokes space-based polarization demultiplexing algorithm," *J. Lightwave Technol.*, vol. 32, no. 19, pp. 3290–3298, 2014.

[46] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Hoboken, NJ, USA: Wiley, Jul. 2006.

[47] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423/623–656, Jul., Oct. 1948.

[48] E. Telatar, "Capacity of multi-antenna Gaussian channels," *Eur. Trans. Telecommun.*, vol. 10, no. 6, pp. 585–595, Nov.–Dec. 1999.

[49] A. Ganti, A. Lapidoth, and I. E. Telatar, "Mismatched decoding revisited: General alphabets, channels with memory, and the wide-band limit," *IEEE Trans. Info. Theory*, vol. 46, no. 7, pp. 2315–2328, 2000.

[50] I. Csiszár and J. Korner, "Graph decomposition: A new key to coding theorems," *IEEE Trans. Info. Theory*, vol. 27, no. 1, pp. 5–12, 1981.

[51] J. Y. N. Hui, "Fundamental issues of multiple accessing," Ph.D. dissertation, Mass. Inst. Technol. (MIT), Cambridge, MA, 1983.

[52] V. B. Balakirsky, "A converse coding theorem for mismatched decoding at the output of binary-input memoryless channels," *IEEE Trans. Info. Theory*, vol. 41, no. 6, pp. 1889–1902, 1995.

[53] I. Csiszar and P. Narayan, "Channel capacity for a given decoding metric," *IEEE Trans. Info. Theory*, vol. 41, no. 1, pp. 35–43, 1995.

[54] D.-M. Arnold, H.-A. Loeliger, P. O. Vontobel, A. Kavcic, and W. Zeng, "Simulation-based computation of information rates for channels with memory," *IEEE Trans. Info. Theory*, vol. 52, no. 8, pp. 3498–3508, 2006.

[55] L. Schmalen, A. Alvarado, and R. Rios-Müller, "Performance prediction of nonbinary forward error correction in optical transmission experiments," *J. Lightwave Technol.*, vol. 35, no. 4, pp. 1015–1027, 2017.

[56] B. Hassibi and B. Hochwald, "How much training is needed in multiple-antenna wireless links?" *IEEE Trans. Inform. Theory*, vol. 49, no. 4, pp. 951–963, Apr. 2003.

[57] T. L. Marzetta, "Blast training: Estimating channel characteristics for high capacity space-time wireless," in *Proc. Annual Allerton Conf. Commun. Control and Computing*, vol. 37. Citeseer, 1999, pp. 958–966.

[58] X. Ma, G. B. Giannakis, and S. Ohno, "Optimal training for block transmissions over doubly selective wireless fading channels," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1351–1366, 2003.

[59] M. Biguesh and A. Gershman, "Training-based MIMO channel estimation: a study of estimator tradeoffs and optimal training signals," *IEEE Trans. on Signal Process.*, vol. 54, no. 3, pp. 884–893, Mar. 2006.

[60] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory*. Englewood Cliffs: NJ: Prentice-Hall, 1993.