Gradient of Mutual Information in Linear Vector Gaussian Channels in the Presence of Input Noise

Fraser K. Coutts, John Thompson, and Bernard Mulgrew
Institute for Digital Communications, University of Edinburgh, Edinburgh, EH9 3FG, UK
fraser.coutts@ed.ac.uk

***Abstract*—This paper considers a general linear vector Gaussian channel with arbitrary signalling in the presence of Gaussian or Gaussian mixture input noise — i.e., noise added to a desired signal prior to its measurement. Generalising the fundamental relationship unveiled by Guo and extended by Palomar, we show for this scenario that the gradient of the mutual information between a desired signal — or its discrete class label — and a measured output with respect to the measurement matrix can be expressed in a novel form without a requirement for the approximations made in previous papers. We demonstrate that the derived expressions can outperform approximate gradient terms when integrated within a gradient ascent multi-objective optimisation approach.**

I. Introduction

In recent years, the fundamental relationships between information theory and estimation theory have been investigated [1]–[4]. The interplay between mutual information (MI) and minimum mean square error (MMSE) — denoted I-MMSE — established by these works has provided new insights into various applications [5]–[7]. In [1] and [4], Guo *et al.* established an explicit relationship between MMSE and the derivative of the input-output MI for additive Gaussian and non-Gaussian noise channels, respectively. Palomar *et al.* then generalised the I-MMSE relationship to linear vector Gaussian channels [2]. Recent research in [3] reviews a number of other advances, and employs a functional approach to study signal models with both additive and multiplicative noise.

To reduce computational complexity and increase interpretability when processing high dimensional data, one typically employs dimensionality reduction. While linear methods based on random projections have gained significant attention recently under the guise of compressive sensing [8] (CS), random projections may not be the best choice if we know the statistical properties of the underlying signal [9]. In the information-theoretic (IT) approach, the projection matrix is designed by maximising the MI between the projected signal and the source signal or its class label [10]–[12]. Intuitively, as MI increases, the recovery of the source signal or label information improves; indeed, the Bayes classification error is bounded by the MI [12]. Given that the MI is typically not easy to calculate numerically, IT algorithms typically seek an approximation to the Shannon MI. For example, in some studies [10], [11], the quadratic MI — which can be calculated analytically using quadratic Rényi entropy — is used instead. Without compromising or simplifying the objective function, work in [13] and [14] has demonstrated the use of Shannon MI optimisation for linear feature design in both signal recovery and classification, respectively, for an arbitrary source distribution. In [15], Wang *et al.* combine these aspects to formulate an algorithm capable of balancing signal recovery and classification.

The majority of CS approaches tackle scenarios in which a source is compressively sampled in the presence of measurement noise — i.e., noise added *after* the act of measurement. The main signal model considered in this paper instead corresponds to an instance of CS with input — or "folded" [16] — noise, which some argue is a more realistic setup for CS [17]. Within the confines of an input noise model — where the I-MMSE relationships of [1], [2] do not apply — Gu *et al.* establish in [18]–[20] an approximation to the gradient of MI with respect to a channel matrix by exploiting a first-order Taylor expansion of the entropy. To achieve such an approximation, these works place constraints on the mean of the output signal and on the characteristics of the source — which is assigned a Gaussian mixture (GM) distribution. In general, a single Gaussian model does not provide a sufficiently accurate description of source signals [21]; instead, the distribution of a collection of signals can be approximated by a mixture of several Gaussians. In CS scenarios, such GM models (GMMs) have been shown to be effective [22] and in some cases superior to sparse signal models [21].

In this paper, we show for a linear vector channel with Gaussian or GM input noise that the gradient of the MI between a desired signal with arbitrary probability distribution — or its discrete class label — and a measured output with respect to the measurement matrix can be expressed in a novel form without a requirement for the approximations made in related work [18]–[20]. To illustrate their utility, we demonstrate that the derived expressions can be integrated within a gradient ascent multi-objective optimisation approach that extends the work of [15]. Specifically, the method in [15] functions only for a channel with Gaussian or Poisson measurement noise, while our approach applies to channels with both GM input noise and Gaussian measurement noise. The Bayesian inference model of [23] was used in the approach of [15]; since this model only considers measurement noise, we utilise a novel extension of this model that incorporates GM input noise. We benchmark our derived gradient terms against approximations to the gradient of MI obtained via the approach of [18].

Below, Sec. II and Sec. III establish the signal model considered in this paper and some key theoretical results. Sec. IV provides a practical demonstration of some of the

derived expressions, and conclusions are drawn in Sec. V.

Notation: Straight bold lowercase and uppercase symbols denote vectors and matrices, respectively. Italicised uppercase letters such as $\boldsymbol{Y}$ and C denote random vectors and variables; their realisations are represented using a lowercase equivalent, such as $\boldsymbol{y}$ or c. We use the notation $\mathbb{E}_{\boldsymbol{x},\boldsymbol{y},k}[f(\boldsymbol{x},\boldsymbol{y},k)] = \sum_k s_k \iint p_{\boldsymbol{x}|k}(\boldsymbol{x}|k) p_{\boldsymbol{y}|\boldsymbol{x},k}(\boldsymbol{y}|\boldsymbol{x},k) f(\boldsymbol{x},\boldsymbol{y},k)\, d\boldsymbol{y}\, d\boldsymbol{x}$ for an arbitrary function $f(\boldsymbol{x},\boldsymbol{y},k)$ and discrete probability s_k dependent on index k. Following typical notation, $\mathbb{E}[\boldsymbol{x}|\boldsymbol{y}]$ denotes the expectation of $\boldsymbol{x}$ over the distribution $p_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}|\boldsymbol{y})$. Throughout this work, $\mathbf{I}_n$ is an $n \times n$ identity matrix and $[\boldsymbol{\Phi}]_{ij}$ is the (i,j)th element of a matrix $\boldsymbol{\Phi}$.

II. Signal Model

In this work, we consider a real-valued signal model. The theoretical results of Sec. III can be extended to complex-valued signals with minor modifications; the details of this are therefore omitted in the pursuit of brevity. The real-valued CS model with measurement noise is considered in the majority of related works [1]–[3]. This is typically conveyed as follows:

$$\boldsymbol{Y} = \boldsymbol{\Phi}\boldsymbol{X} + \boldsymbol{W}\,. \tag{1}$$

Here, we instead utilise the following input ("folding" [16]) noise model considered in some other related works [18], [20]:

$$\boldsymbol{Y} = \boldsymbol{\Phi}(\boldsymbol{X} + \boldsymbol{N}) + \boldsymbol{W}\,. \tag{2}$$

Following the compressive sampling protocol, we have measurements $\boldsymbol{Y} \in \mathbb{R}^m$ obtained from some desired signal $\boldsymbol{X} \in \mathbb{R}^n$ via a compressive measurement matrix $\boldsymbol{\Phi} \in \mathbb{R}^{m \times n}$, with $m \ll n$. The noise term $\boldsymbol{N} \in \mathbb{R}^n$, which is independent of $\boldsymbol{X}$, is a GM defined by

$$\boldsymbol{N} \sim \sum\nolimits_{k=1}^{K} s_k \mathcal{N}(\boldsymbol{n}; \boldsymbol{\mu}^{(k)}, \boldsymbol{\Gamma}^{(k)})\,, \tag{3}$$

with $\boldsymbol{\mu}^{(k)} \in \mathbb{R}^n$, $\boldsymbol{\Gamma}^{(k)} \in \mathbb{R}^{n \times n}$, and $\sum_{k=1}^{K} s_k = 1$. The vector $\boldsymbol{W} \sim \mathcal{N}(\boldsymbol{w}; \boldsymbol{\nu}, \boldsymbol{\Lambda})$ represents arbitrary additive Gaussian noise. Given the above, the distribution of $\boldsymbol{Y}$ given the observation of $\boldsymbol{x}$ and the knowledge that $\boldsymbol{n}$ has been generated according to the kth Gaussian component in the distribution of $\boldsymbol{N}$ is

$$p_{\boldsymbol{y}|\boldsymbol{x},k}(\boldsymbol{y}|\boldsymbol{x},k) = \frac{e^{-\frac{1}{2}(\boldsymbol{y}-\bar{\boldsymbol{y}}^{(k)})^{\mathrm{T}} \boldsymbol{\Sigma}^{(k),-1} (\boldsymbol{y}-\bar{\boldsymbol{y}}^{(k)})}}{(2\pi)^{\frac{m}{2}} \det\{\boldsymbol{\Sigma}^{(k)}\}^{\frac{1}{2}}}\,, \tag{4}$$

with mean vector $\bar{\boldsymbol{y}}^{(k)} = \boldsymbol{\Phi}(\boldsymbol{x} + \boldsymbol{\mu}^{(k)}) + \boldsymbol{\nu}$ and covariance $\boldsymbol{\Sigma}^{(k)} = \boldsymbol{\Phi}\boldsymbol{\Gamma}^{(k)}\boldsymbol{\Phi}^{\mathrm{T}} + \boldsymbol{\Lambda}$. It is assumed that $\boldsymbol{\Sigma}^{(k)}$ is invertible. One can therefore obtain $p_{\boldsymbol{y}|\boldsymbol{x}}(\boldsymbol{y}|\boldsymbol{x}) = \sum_{k=1}^{K} s_k p_{\boldsymbol{y}|\boldsymbol{x},k}(\boldsymbol{y}|\boldsymbol{x},k)$.

Suppose that an instance of $\boldsymbol{X}$ is generated by one of J underlying classes, with each class, $c = 1 \ldots J$, occurring with probability z_c. If the data distribution for class c is $p_{\boldsymbol{x}|c}(\boldsymbol{x}|c)$, the joint density is $p_{\boldsymbol{x},c}(\boldsymbol{x},c) = z_c p_{\boldsymbol{x}|c}(\boldsymbol{x}|c)$, and the global signal density is $p_{\boldsymbol{x}}(\boldsymbol{x}) = \sum_{c=1}^{J} z_c p_{\boldsymbol{x}|c}(\boldsymbol{x}|c)$. For now, we make no assumption on the form of $p_{\boldsymbol{x}|c}(\boldsymbol{x}|c)$.

III. Key Theoretical Results

In this section, we generalise the gradient of MI results of [2] to scenarios incorporating a linear vector channel with Gaussian or GM input noise and Gaussian measurement noise. We state new expressions for the gradient of the MI between a desired signal with arbitrary probability distribution — or its discrete class label — and a measured output with respect to the measurement matrix.

Theorem 1 (Gradient of $I(\boldsymbol{X};\boldsymbol{Y})$ with Gaussian input noise)**.** With $I(\boldsymbol{X};\boldsymbol{Y})$ defined as the Shannon MI between $\boldsymbol{X}$ and $\boldsymbol{Y}$, for the signal model of (2) with $\boldsymbol{N}$ modelled by a single Gaussian distribution such that $K = 1$, $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^{(1)}$, and $\boldsymbol{\Sigma} = \boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Phi}^{\mathrm{T}} + \boldsymbol{\Lambda}$, the gradient of $I(\boldsymbol{X};\boldsymbol{Y})$ with respect to $\boldsymbol{\Phi}$ is

$$\nabla_{\boldsymbol{\Phi}} I(\boldsymbol{X};\boldsymbol{Y}) = \boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}\mathbf{E}\Big(\mathbf{I}_n - \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}\boldsymbol{\Gamma}\Big)\,, \tag{5}$$

with MMSE matrix $\mathbf{E} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}])\,(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}])^{\mathrm{T}}]$.

Proof. One can obtain (5) via the derivations in [2]. The first term is simply the well-researched gradient of $I(\boldsymbol{X};\boldsymbol{Y})$ with respect to $\boldsymbol{\Phi}$ for the case of the signal model in (1) ((21) in [2]), and the second term is — via the chain rule — the product of the derivative of $I(\boldsymbol{X};\boldsymbol{Y})$ with respect to the covariance matrix $\boldsymbol{\Sigma}$ ((27) in [2]) and the derivative of $\boldsymbol{\Sigma}$ with respect to $\boldsymbol{\Phi}$. ■

Theorem 2 (Gradient of $I(\boldsymbol{X};\boldsymbol{Y})$ with GM input noise)**.** The gradient of $I(\boldsymbol{X};\boldsymbol{Y})$ with respect to $\boldsymbol{\Phi}$ for the signal model of (2) with $\boldsymbol{N}$ modelled by a GM distribution is

$$\nabla_{\boldsymbol{\Phi}} I(\boldsymbol{X};\boldsymbol{Y}) = \boldsymbol{\Lambda}^{-1}\boldsymbol{\Phi}\mathbf{E}_{\boldsymbol{z},\boldsymbol{x}}\,, \tag{6}$$

with $\boldsymbol{Z} = \boldsymbol{X} + \boldsymbol{N}$ such that $\boldsymbol{X} \to \boldsymbol{Z} \to \boldsymbol{Y}$, $\mathbf{E}_{\boldsymbol{z},\boldsymbol{x}} = \mathbb{E}_{\boldsymbol{x},\boldsymbol{y}}\big[\,(\mathbb{E}[\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}] - \mathbb{E}[\boldsymbol{z}|\boldsymbol{y}])(\mathbb{E}[\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}] - \mathbb{E}[\boldsymbol{z}|\boldsymbol{y}])^{\mathrm{T}}\,\big]$,

$$\mathbb{E}[\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}] = \mathop{\mathbb{E}}_{k|\boldsymbol{x},\boldsymbol{y}}\Big[\,\Big(\boldsymbol{\Gamma}^{(k),-1} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Phi}\Big)^{-1} \times \Big(\boldsymbol{\Gamma}^{(k),-1}\big(\boldsymbol{x} + \boldsymbol{\mu}^{(k)}\big) + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{y} - \boldsymbol{\nu})\Big)\Big]\,, \tag{7}$$

and

$$\mathbb{E}[\boldsymbol{z}|\boldsymbol{y}] = \mathop{\mathbb{E}}_{\boldsymbol{x}|\boldsymbol{y}}\big[\,\mathbb{E}[\boldsymbol{z}|\boldsymbol{x},\boldsymbol{y}]\,\big] = \mathop{\mathbb{E}}_{k|\boldsymbol{y}}\Big[\,\Big(\boldsymbol{\Gamma}^{(k),-1} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Phi}\Big)^{-1} \times \Big(\boldsymbol{\Gamma}^{(k),-1}\big(\mathbb{E}[\boldsymbol{x}|\boldsymbol{y},k] + \boldsymbol{\mu}^{(k)}\big) + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{y} - \boldsymbol{\nu})\Big)\Big]\,. \tag{8}$$

Proof. The gradient of $I(\boldsymbol{X};\boldsymbol{Y})$ with respect to the measurement matrix $\boldsymbol{\Phi}$ is $\nabla_{\boldsymbol{\Phi}} I(\boldsymbol{X};\boldsymbol{Y}) = \nabla_{\boldsymbol{\Phi}} h(\boldsymbol{Y}) - \nabla_{\boldsymbol{\Phi}} h(\boldsymbol{Y}|\boldsymbol{X})$. To obtain (6), we formulate individual expressions for the gradient of the entropy of $\boldsymbol{Y}$, $\nabla_{\boldsymbol{\Phi}} h(\boldsymbol{Y})$, and the gradient of the conditional entropy, $\nabla_{\boldsymbol{\Phi}} h(\boldsymbol{Y}|\boldsymbol{X})$. In the interest of conserving space, the resulting proof will be described in a subsequent publication. ■

Remark 1. Note that if $K = 1$, $\boldsymbol{\Gamma} = \boldsymbol{\Gamma}^{(1)}$, and $\boldsymbol{\Sigma} = \boldsymbol{\Phi}\boldsymbol{\Gamma}\boldsymbol{\Phi}^{\mathrm{T}} + \boldsymbol{\Lambda}$, Theorem 2 reduces to the result of Theorem 1.

Theorem 3 (Gradient of $I(C;\boldsymbol{Y})$ with GM input noise)**.** With $I(C;\boldsymbol{Y})$ defined as the Shannon MI between the class of $\boldsymbol{X}$ — denoted by random variable C — and $\boldsymbol{Y}$, the gradient of $I(C;\boldsymbol{Y})$ with respect to $\boldsymbol{\Phi}$ for the signal model of (2) is

$$\nabla_{\boldsymbol{\Phi}} I(C;\boldsymbol{Y}) = \boldsymbol{\Lambda}^{-1}\boldsymbol{\Phi}\mathbf{E}_{\boldsymbol{z},c}\,. \tag{9}$$

Here, $\mathbf{E}_{\boldsymbol{z},c} = \mathbb{E}_{c,\boldsymbol{y}}\big[\,(\mathbb{E}[\boldsymbol{z}|c,\boldsymbol{y}] - \mathbb{E}[\boldsymbol{z}|\boldsymbol{y}])(\mathbb{E}[\boldsymbol{z}|c,\boldsymbol{y}] - \mathbb{E}[\boldsymbol{z}|\boldsymbol{y}])^{\mathrm{T}}\,\big]$,

$$\mathbb{E}[\boldsymbol{z}|c,\boldsymbol{y}] = \mathop{\mathbb{E}}_{k|c,\boldsymbol{y}}\Big[\,\Big(\boldsymbol{\Gamma}^{(k),-1} + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{\Phi}\Big)^{-1} \times \Big(\boldsymbol{\Gamma}^{(k),-1}\big(\mathbb{E}[\boldsymbol{x}|c,\boldsymbol{y},k] + \boldsymbol{\mu}^{(k)}\big) + \boldsymbol{\Phi}^{\mathrm{T}}\boldsymbol{\Lambda}^{-1}(\boldsymbol{y} - \boldsymbol{\nu})\Big)\Big]\,, \tag{10}$$

and $\mathbb{E}[\boldsymbol{z}|\boldsymbol{y}]$ is as defined in (8).

Proof. The gradient of the MI between C and $\boldsymbol{Y}$ can be defined as $\nabla_{\boldsymbol{\Phi}} I(C;\boldsymbol{Y}) = \nabla_{\boldsymbol{\Phi}} h(\boldsymbol{Y}) - \nabla_{\boldsymbol{\Phi}} h(\boldsymbol{Y}|C)$. An expression for $\nabla_{\boldsymbol{\Phi}} h(\boldsymbol{Y})$ is obtained as part of the extended proof of Theorem 2; thus, we require only an expression for $\nabla_{\boldsymbol{\Phi}} h(\boldsymbol{Y}|C)$ to evaluate $\nabla_{\boldsymbol{\Phi}} I(C;\boldsymbol{Y})$. The derivation of this — which, for brevity, is omitted here — uses the expression for $\nabla_{\boldsymbol{\Phi}} h(\boldsymbol{Y}|\boldsymbol{X})$ from Theorem 2 and the knowledge that $C \to \boldsymbol{X} \to \boldsymbol{Z} \to \boldsymbol{Y}$ forms a Markov chain. ■

IV. Practical Demonstration

A. Scenario

For a practical demonstration of the utility of the derived expressions, we will consider an application involving the signal model in (2) and the result given in Theorem 2. Within the context of this model, we opt for both $\boldsymbol{X}$ and $\boldsymbol{N}$ to be represented by GMMs. Their distributions are defined as

$$\boldsymbol{X} \sim \sum\nolimits_{c=1}^{J} z_c \sum\nolimits_{o=1}^{O} \pi_{c,o} \mathcal{N}(\boldsymbol{x}; \boldsymbol{\chi}_{c,o}, \boldsymbol{\Omega}_{c,o}) \,, \tag{11}$$

$$\boldsymbol{N} \sim \sum\nolimits_{\ell=1}^{L} r_\ell \sum\nolimits_{g=1}^{G} \upsilon_{\ell,g} \mathcal{N}(\boldsymbol{n}; \boldsymbol{\eta}_{\ell,g}, \boldsymbol{\Theta}_{\ell,g}) \,. \tag{12}$$

Here, the probability distributions of classes $c = 1 \dots J$ and $\ell = 1 \dots L$ of $\boldsymbol{X}$ and $\boldsymbol{N}$ are characterised by GMs with O and G components, respectively.

If we want to identify the $\boldsymbol{\Phi}$ that maximises $I(\boldsymbol{X};\boldsymbol{Y})$, we can reparameterise $\boldsymbol{N}$ according to (3) and use the relevant $\nabla_{\boldsymbol{\Phi}} I(\boldsymbol{X};\boldsymbol{Y})$ from (6) in a gradient ascent scenario akin to that of [15]. Analogously, we could maximise $I(\boldsymbol{N};\boldsymbol{Y})$ by treating $\boldsymbol{X}$ as a noise signal. Here, we consider the case that $I(\boldsymbol{X};\boldsymbol{Y})$ is to be maximised, while $I(\boldsymbol{N};\boldsymbol{Y})$ is to be minimised. Such a scenario may occur when both a desired and an undesired signal — with known or identifiable probability distributions — are simultaneously measured via matrix $\boldsymbol{\Phi}$ in the presence of measurement noise $\boldsymbol{W}$. The undesired signal encapsulated by $\boldsymbol{N}$ might, for example, contain sensitive information. The objective function to be maximised in this case would be

$$F(\boldsymbol{\Phi}, \beta) = I(\boldsymbol{X};\boldsymbol{Y}) - \beta I(\boldsymbol{N};\boldsymbol{Y}) \,, \tag{13}$$

where $\beta \in \mathbb{R}$ controls the relative importance of the "privacy term" $I(\boldsymbol{N};\boldsymbol{Y})$. This problem is reminiscent of the one in [15]; though in this case, we distinguish between two independent sources instead of balancing the reconstruction and classification of a single source. If desired, one could add a term containing $I(C;\boldsymbol{Y})$ to (13) to account for the MI between the underlying classes of $\boldsymbol{X}$ and the output.

An iterative gradient ascent algorithm can attempt to identify the $\boldsymbol{\Phi}$ that maximises $F(\boldsymbol{\Phi}, \beta)$ by setting $\boldsymbol{\Phi} \leftarrow \boldsymbol{\Phi} + \delta \nabla_{\boldsymbol{\Phi}} F(\boldsymbol{\Phi}, \beta)$ at each iteration. The step size $\delta > 0$ controls the rate of change of $\boldsymbol{\Phi}$. To obtain $\nabla_{\boldsymbol{\Phi}} I(\boldsymbol{X};\boldsymbol{Y})$, we reparameterise the probability distribution of $\boldsymbol{N}$ according to (3) with

$$K = LG \,, \qquad s_k = r_{\ell'} \upsilon_{\ell',g'} \,, \quad \boldsymbol{\mu}^{(k)} = \boldsymbol{\eta}_{\ell',g'} \,,$$

$$\boldsymbol{\Gamma}^{(k)} = \boldsymbol{\Theta}_{\ell',g'} \,, \quad \ell' = \left\lceil \tfrac{k}{G} \right\rceil \,, \qquad g' = ((k-1) \bmod G) + 1,$$

where $\lceil \cdot \rceil$ is the ceiling function. We can then evaluate (6) via Monte Carlo (MC) integration and utilise the inference model detailed in Sec. IV-B.

To obtain $\nabla_{\boldsymbol{\Phi}} I(\boldsymbol{N};\boldsymbol{Y})$, we redefine the probability distribution of $\boldsymbol{X}$ as a GM with JO components and use an appropriately modified version of the inference model below.

B. Inference Model

Under the chosen signal model in (2), the general Bayesian inference model is a novel extension of the model in [23] that incorporates input noise. It is constructed with the following modifications to the existing model:

$$p_{\boldsymbol{x}|\boldsymbol{y}}(\boldsymbol{x}|\boldsymbol{y}) = \sum\nolimits_{c=1}^{J} \tilde{z}_c \, p_{\boldsymbol{x}|\boldsymbol{y},c}(\boldsymbol{x}|\boldsymbol{y}, c) \,, \tag{14}$$

$$\tilde{z}_c = p_{c|\boldsymbol{y}}(c|\boldsymbol{y}) = \frac{z_c \, p_{\boldsymbol{y}|c}(\boldsymbol{y}|c)}{p_{\boldsymbol{y}}(\boldsymbol{y})} = \frac{z_c \, p_{\boldsymbol{y}|c}(\boldsymbol{y}|c)}{\sum_{c'=1}^{J} z_{c'} \, p_{\boldsymbol{y}|c}(\boldsymbol{y}|c')} \,, \tag{15}$$

$$p_{\boldsymbol{y}|c,k,o}(\boldsymbol{y}|c,k,o) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\Phi}(\boldsymbol{\chi}_{c,o} + \boldsymbol{\mu}^{(k)}) + \boldsymbol{\nu}, \boldsymbol{\Sigma}^{(k)} + \boldsymbol{\Phi}\boldsymbol{\Omega}_{c,o}\boldsymbol{\Phi}^{\mathrm{T}}), \tag{16}$$

$$p_{\boldsymbol{y}|c}(\boldsymbol{y}|c) = \sum\nolimits_{o=1}^{O} \sum\nolimits_{k=1}^{K} \pi_{c,o} \, s_k \, p_{\boldsymbol{y}|c,k,o}(\boldsymbol{y}|c,k,o) \,, \tag{17}$$

$$p_{\boldsymbol{x}|\boldsymbol{y},c}(\boldsymbol{x}|\boldsymbol{y},c) = \sum\nolimits_{o=1}^{O} \sum\nolimits_{k=1}^{K} \tilde{\pi}_{c,o}^{(k)} \mathcal{N}(\boldsymbol{x}; \tilde{\boldsymbol{\chi}}_{c,o}^{(k)}, \tilde{\boldsymbol{\Omega}}_{c,o}^{(k)}) \,, \tag{18}$$

$$\tilde{\boldsymbol{\Omega}}_{c,o}^{(k)} = \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Sigma}^{(k),-1} \boldsymbol{\Phi} + \boldsymbol{\Omega}_{c,o}^{-1} \right)^{-1} \,, \tag{19}$$

$$\tilde{\pi}_{c,o}^{(k)} = \frac{\pi_{c,o} \, s_k \, p_{\boldsymbol{y}|c,k,o}(\boldsymbol{y}|c,k,o)}{p_{\boldsymbol{y}|c}(\boldsymbol{y}|c)}, \tag{20}$$

$$\tilde{\boldsymbol{\chi}}_{c,o}^{(k)} = \tilde{\boldsymbol{\Omega}}_{c,o}^{(k)} \left(\boldsymbol{\Phi}^{\mathrm{T}} \boldsymbol{\Sigma}^{(k),-1} \left(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{\mu}^{(k)} - \boldsymbol{\nu} \right) + \boldsymbol{\Omega}_{c,o}^{-1} \boldsymbol{\chi}_{c,o} \right) . \tag{21}$$

An estimate of $\boldsymbol{x}$ given knowledge of $\boldsymbol{y}$ can be obtained via

$$\hat{\boldsymbol{x}} = \mathbb{E}[\boldsymbol{x}|\boldsymbol{y}] = \sum\nolimits_{c=1}^{J} \tilde{z}_c \int \boldsymbol{x} p_{\boldsymbol{x}|\boldsymbol{y},c}(\boldsymbol{x}|\boldsymbol{y}, c) \, d\boldsymbol{x} \tag{22}$$

and the most likely class given knowledge of $\boldsymbol{y}$ is

$$\hat{c} = \max_c \tilde{z}_c = \max_c \left\{ z_c p_{\boldsymbol{y}|c}(\boldsymbol{y}|c) \right\} \,. \tag{23}$$

Using the above model, closed-form expressions can also be found for

$$\mathbb{E}[\boldsymbol{x}|c,\boldsymbol{y},k] = \frac{\sum_{o=1}^{O} \pi_{c,o} \, p_{\boldsymbol{y}|c,k,o}(\boldsymbol{y}|c,k,o) \tilde{\boldsymbol{\chi}}_{c,o}^{(k)}}{\sum_{o'=1}^{O} \pi_{c,o'} \, p_{\boldsymbol{y}|c,k,o}(\boldsymbol{y}|c,k,o')} \tag{24}$$

and

$$\mathbb{E}[\boldsymbol{x}|\boldsymbol{y},k] = \frac{\sum_{c=1}^{J} \sum_{o=1}^{O} z_c \, \pi_{c,o} \, p_{\boldsymbol{y}|c,k,o}(\boldsymbol{y}|c,k,o) \tilde{\boldsymbol{\chi}}_{c,o}^{(k)}}{\sum_{c'=1}^{J} \sum_{o'=1}^{O} z_{c'} \, \pi_{c',o'} \, p_{\boldsymbol{y}|c,k,o}(\boldsymbol{y}|c',k,o')} \,. \tag{25}$$

A similar inference model can also be derived for the classification and reconstruction of $\boldsymbol{N}$.

C. Gradient-Based Numerical Solution

To attempt to identify the $\boldsymbol{\Phi}$ that maximises $F(\boldsymbol{\Phi}, \beta)$, we adapt the iterative gradient ascent approach of [15]. Note that $F(\boldsymbol{\Phi}, \beta)$ — and indeed the objective function in [15] — is not, in general, a convex or concave function of $\boldsymbol{\Phi}$; thus, finding a global-optimal solution is not guaranteed. During iterations, we constrain the energy of $\boldsymbol{\Phi}$ such that $\mathrm{tr}\{\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}}\} = m$. In agreement with [15], during testing, an orthonormality ($\boldsymbol{\Phi}\boldsymbol{\Phi}^{\mathrm{T}} = \mathbf{I}_m$) constraint offered similar performance. Updating $\boldsymbol{\Phi}$ without a constraint also resulted in comparable performance; however, a greater degree of care

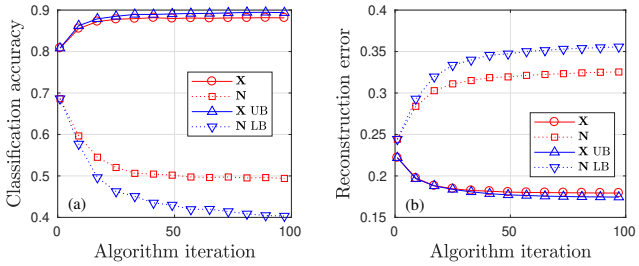

Fig. 1. (a) Classification accuracy and (b) reconstruction error versus algorithm iteration for sources $\boldsymbol{X}$ and $\boldsymbol{N}$ when using novel gradient terms.

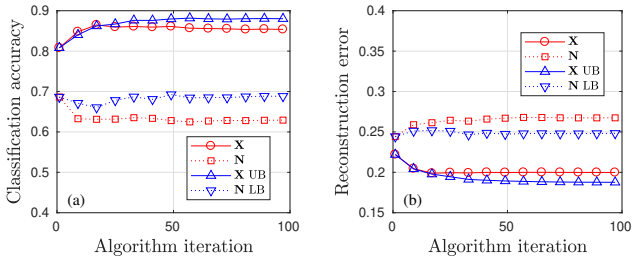

Fig. 2. (a) Classification accuracy and (b) reconstruction error versus algorithm iteration when using approximation [18] to gradient terms.

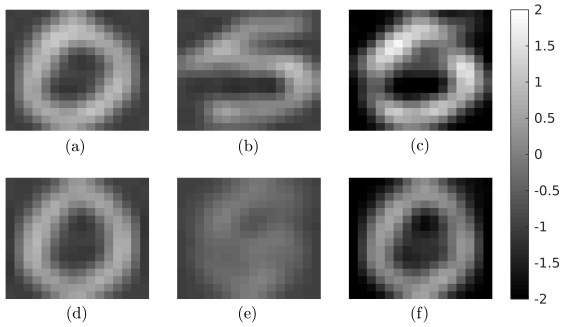

Fig. 3. Original digits: (a) $\boldsymbol{x}$, (b) $\boldsymbol{n}$, (c) $\boldsymbol{x}+\boldsymbol{n}$. Reconstructed digits following design of $\boldsymbol{\Phi}$: (d) $\hat{\boldsymbol{x}}$, (e) $\hat{\boldsymbol{n}}$, and (f) $\hat{\boldsymbol{x}}+\hat{\boldsymbol{n}}$.

was required when selecting step size δ. Note that without an active constraint, the effect of $\boldsymbol{W}$ on the output measurement can be negated by scaling the elements of $\boldsymbol{\Phi}$ towards infinity.

The gradient ascent approach operates over a total of ρ iterations and can be summarised as follows:

- Draw $[\boldsymbol{\Phi}]_{ij}$ from $\mathcal{N}(0,1)$; normalise s.t. $\text{tr}\{\boldsymbol{\Phi}\boldsymbol{\Phi}^{\text{T}}\}=m$.
- While the number of iterations is below ρ:
  1) Draw S samples of $\boldsymbol{X}$, $\boldsymbol{N}$, and $\boldsymbol{W}$; evaluate (2).
  2) Compute (19), (20), and (21); evaluate (25).
  3) Compute $\nabla_{\boldsymbol{\Phi}} I(\boldsymbol{X};\boldsymbol{Y})$ via (6) using MC integration.
  4) Update the inference model for $\boldsymbol{N}$; compute $\nabla_{\boldsymbol{\Phi}} I(\boldsymbol{N};\boldsymbol{Y})$.
  5) $\boldsymbol{\Phi} \leftarrow \boldsymbol{\Phi} + \delta \nabla_{\boldsymbol{\Phi}} F(\boldsymbol{\Phi},\beta)$; normalise s.t. $\text{tr}\{\boldsymbol{\Phi}\boldsymbol{\Phi}^{\text{T}}\} = m$.

D. Simulation Results for Real Data

Real data was acquired from the *USPS* dataset, which contains 16×16 grayscale images of handwritten digits 0 to 9. To aid the fitting of GMMs to the images, the two-dimensional discrete cosine transform was applied to each image and the result was vectorised such that $n=256$. Signals $\boldsymbol{X},\boldsymbol{N}\in\mathbb{R}^n$ were chosen to represent digits 0 to 4 and 5 to 9, respectively. The sum $\boldsymbol{x}+\boldsymbol{n}$ is therefore the superposition of two digits. Using the expectation-maximisation algorithm [24] and 500 training examples of each digit, the distributions of $\boldsymbol{X}$ and $\boldsymbol{N}$ were fitted such that each of the $J=L=5$ classes had $O=G=3$ mixture components. For demonstration purposes, samples of measurement noise $\boldsymbol{W}\in\mathbb{R}^m$ were drawn according to $\boldsymbol{W}\sim\mathcal{N}(\boldsymbol{w};\boldsymbol{0},10^{-6}\mathbf{I}_m)$. Values of $m=8$ and $\delta=0.01$ were used and $S=500$ Monte Carlo draws were utilised to evaluate $\nabla_{\boldsymbol{\Phi}}F(\boldsymbol{\Phi},\beta)$ with $\beta=1$. At each iteration, 100 test samples of $\boldsymbol{X}$ and $\boldsymbol{N}$ were used to compute the classification accuracy (measured in the range $[0,1]$) and reconstruction error. The latter was measured via the metrics

$$\epsilon_{\boldsymbol{x}} = \mathbb{E}\left[\frac{\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2^2}\right], \quad \epsilon_{\boldsymbol{n}} = \mathbb{E}\left[\frac{\|\hat{\boldsymbol{n}}-\boldsymbol{n}\|_2^2}{\|\boldsymbol{n}\|_2^2}\right], \tag{26}$$

where $\hat{\boldsymbol{n}}=\mathbb{E}[\boldsymbol{n}|\boldsymbol{y}]$. Results were averaged over 100 instances of the simulation scenario.

Fig. 1 shows, in red, the proposed method's performance over 100 iterations. In blue are upper- and lower-bounds for $\boldsymbol{X}$ and $\boldsymbol{N}$; these were obtained by optimising for $I(\boldsymbol{X};\boldsymbol{Y})$ only and $I(\boldsymbol{N};\boldsymbol{Y})$ only, respectively. Clearly, by optimising over two terms simultaneously, some performance has been lost. However, the results do confirm that the method is working as expected; i.e., information relating to $\boldsymbol{X}$ has been retained while information relating to $\boldsymbol{N}$ has been reduced. It is clear that increasing $I(\boldsymbol{X};\boldsymbol{Y})$ decreases reconstruction error for $\boldsymbol{X}$ while increasing the accuracy at which the underlying classes can be estimated. Note that a similar relationship was observed in [15]. For $\boldsymbol{N}$, decreasing $I(\boldsymbol{N};\boldsymbol{Y})$ increases reconstruction error while decreasing classification accuracy.

For comparison purposes, we used an identical problem formulation to that defined above with alternative expressions for $\nabla_{\boldsymbol{\Phi}} I(\boldsymbol{X};\boldsymbol{Y})$ and $\nabla_{\boldsymbol{\Phi}} I(\boldsymbol{N};\boldsymbol{Y})$ derived from the approximate gradient of MI solution presented in [18]. Results for this method are shown in Fig. 2. When compared with the proposed gradient term, the approximate gradient operates similarly when optimising for $\boldsymbol{X}$, but offers poorer performance when minimising the information regarding $\boldsymbol{N}$. Indeed, the approximate gradient term is not capable of solely minimising $I(\boldsymbol{N};\boldsymbol{Y})$ and the intended lower-bound for $\boldsymbol{N}$ is ignored when optimising for both $\boldsymbol{X}$ and $\boldsymbol{N}$.

After 100 iterations, the measurement matrix $\boldsymbol{\Phi}$ generated by the proposed approach was utilised to facilitate the reconstruction of the digits in Fig. 3. It can be seen that information regarding $\boldsymbol{X}$ has been preserved, while knowledge of the structure of $\boldsymbol{N}$ has been reduced.

If we wish to reduce computational complexity, we might choose to approximate the distributions of $\boldsymbol{X}$ and $\boldsymbol{N}$ with fewer GM components. Conversely, to improve performance, we might choose to increase the number of components. Similarly, increasing m should increase performance at the expense of higher complexity. Fig. 4 demonstrates the impact of changing the number of components and measurements for both the proposed and approximate gradient terms. Each data point is the final $\epsilon_{\boldsymbol{x}}$ obtained after $\rho=100$ iterations. For reference purposes, results from using a random, untrained, measurement matrix are also provided; for this, elements of $\boldsymbol{\Phi}$ were generated such that $[\boldsymbol{\Phi}]_{ij}\sim\mathcal{N}(0,1)$ and $\boldsymbol{\Phi}$ was then

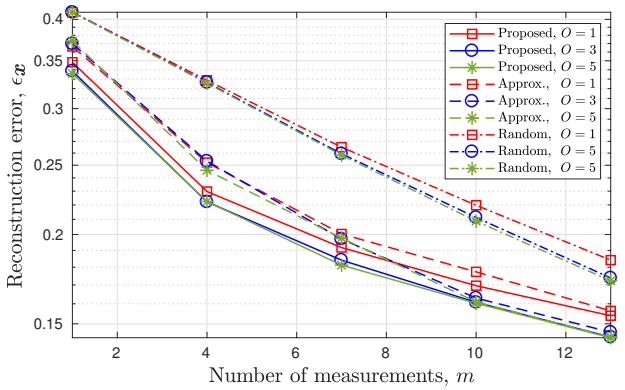

Fig. 4. Reconstruction error $\epsilon_{\boldsymbol{x}}$ versus number of measurements when using proposed and approximate [18] gradient terms with $G = O \in \{1, 3, 5\}$ GM components for each class. Results for random $\boldsymbol{\Phi}$ are also shown.

normalised to satisfy $\text{tr}\{\boldsymbol{\Phi}\boldsymbol{\Phi}^{\text{T}}\} = m$.

Fig. 4 confirms that increasing m and O generally increases performance, though the increase from $O = 3$ to $O = 5$ is small — indicating that increasing O beyond this point is unnecessary. The figure also demonstrates that using a random matrix instead of one obtained via the proposed gradient-based numerical solution results in universally poorer performance. The proposed gradient term outperforms the approximation offered by [18] for low m; however, as m increases, the two approaches begin to exhibit similar performance. As computation of the approximate gradient terms does not require MC integration, and since computational complexity increases with m, the approximate gradient terms become more appealing when m is large; thus, they are a viable alternative to the proposed gradient terms.

V. Conclusions

In this paper, we have shown that for a vector Gaussian channel with GM input noise, the gradient of the MI between a desired signal — or its discrete class label — and a measured output with respect to the measurement matrix can be expressed without a requirement for the approximations made in related work. By approximating the distributions of image data using GMMs, we have demonstrated that the proposed gradient terms can be integrated within a gradient ascent multi-objective optimisation approach. Furthermore, simulation results have demonstrated that the measurement matrix generated following the insertion of approximate gradient terms into the same approach offers poorer performance; however, this disparity diminishes as the number of measurements, m, is increased. When maximising $I(\boldsymbol{X}; \boldsymbol{Y})$, trained measurement matrices have been shown to give superior performance to those with randomly generated elements.

Acknowledgement

This work was supported by the Engineering and Physical Sciences Research Council of the UK (EPSRC) Grant number EP/S000631/1 and the UK MOD University Defence Research Collaboration (UDRC) in Signal Processing. The authors would like to thank Prof. Mike Davies, Prof. Mathini Sellathurai, and Dr João F. C. Mota for their valuable input.

References

[1] D. Guo, S. Shamai, and S. Verdú, "Mutual information and minimum mean-square error in Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 51, no. 4, pp. 1261–1282, April 2005.

[2] D. P. Palomar and S. Verdú, "Gradient of mutual information in linear vector Gaussian channels," *IEEE Trans. Inf. Theory*, vol. 52, no. 1, pp. 141–154, Jan. 2006.

[3] M. Sedighizad and B. Seyfe, "Gradients of the fundamental information measures: Theory and applications," *Signal Processing*, vol. 162, pp. 296–306, 2019.

[4] D. Guo, S. Shamai, and S. Verdú, "Additive non-Gaussian noise channels: mutual information and conditional mean estimation," in *Proc. Int. Symp. Inf. Theory*, Sep. 2005, pp. 719–723.

[5] A. Lozano, A. M. Tulino, and S. Verdú, "Optimum power allocation for parallel Gaussian channels with arbitrary input distributions," *IEEE Trans. Information Theory*, vol. 52, no. 7, pp. 3033–3051, July 2006.

[6] F. Pérez-Cruz, M. R. Rodrigues, and S. Verdú, "Generalized mercury/waterfilling for multiple-input multiple-output channels," in *45th Allerton Conf. Commun., Control, and Computing*, 2007.

[7] S. S. Christensen, R. Agarwal, E. De Carvalho, and J. M. Cioffi, "Weighted sum-rate maximization using weighted MMSE for MIMO-BC beamforming design," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 4792–4799, Dec. 2008.

[8] E. Candes and M. Wakin, "An introduction to compressive sampling," *IEEE Signal Process. Mag.*, vol. 25, no. 2, pp. 21–30, March 2008.

[9] J. Duarte-Carvajalino and G. Sapiro, "Learning to sense sparse signals: Simultaneous sensing matrix and sparsifying dictionary optimization," *IEEE Trans. Image Process.*, vol. 18, no. 7, pp. 1395–1408, July 2009.

[10] K. Torkkola, "Learning discriminative feature transforms to low dimensions in low dimensions," in *Adv. Neural Inf. Process. Syst.*, 2001, pp. 969–976.

[11] K. E. Hild, D. Erdogmus, K. Torkkola, and J. C. Principe, "Feature extraction using information-theoretic learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 9, pp. 1385–1392, Sep. 2006.

[12] Z. Nenadic, "Information discriminant analysis: Feature extraction with an information-theoretic objective," *IEEE Trans. Pattern Anal. Mach. Int.*, vol. 29, no. 8, pp. 1394–1407, Aug. 2007.

[13] W. Carson, M. Chen, M. Rodrigues, R. Calderbank, and L. Carin, "Communications-inspired projection design with application to compressive sensing," *SIAM J. Imag. Sci.*, vol. 5, no. 4, pp. 1185–1212, 2012.

[14] M. Chen, W. Carson, M. Rodrigues, R. Calderbank, and L. Carin, "Communications inspired linear discriminant analysis," *Proc. Int. Conf. Machine Learning*, pp. 1507–1514, 2012.

[15] L. Wang *et al.*, "Information-theoretic compressive measurement design," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1150–1164, Jun. 2017.

[16] E. Arias-Castro and Y. C. Eldar, "Noise folding in compressed sensing," *IEEE Sig. Process. Lett.*, vol. 18, no. 8, pp. 478–481, Aug. 2011.

[17] K. Krishnamurthy, R. Willett, and M. Raginsky, "Target detection performance bounds in compressive imaging," *EURASIP J. Adv. Signal Process.*, vol. 2012, no. 1, p. 205, Sep. 2012.

[18] Y. Gu and N. A. Goodman, "Compressed sensing kernel design for radar range profiling," in *2013 IEEE Radar Conf.*, April 2013.

[19] Y. Gu, N. A. Goodman, and A. Ashok, "Radar target profiling and recognition based on TSI-optimized compressive sensing kernel," *IEEE Trans. Signal Process.*, vol. 62, no. 12, pp. 3194–3207, June 2014.

[20] Y. Gu and N. Goodman, "Information-theoretic compressive sensing kernel optimization and Bayesian Cramér-Rao bound for time delay estimation," *IEEE Trans. Signal Process.*, vol. 65, no. 17, pp. 4525–4537, Sep. 2017.

[21] G. Yu and G. Sapiro, "Statistical compressed sensing of Gaussian mixture models," *IEEE Trans. Signal Process.*, vol. 59, no. 12, pp. 5842–5858, Dec. 2011.

[22] F. Renna, R. Calderbank, L. Carin, and M. Rodrigues, "Reconstruction of signals drawn from a Gaussian mixture via noisy compressive measurements," *IEEE Trans. Signal Process.*, vol. 62, no. 9, pp. 2265–2277, May 2014.

[23] M. Chen *et al.*, "Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds," *IEEE Trans. Signal Process.*, vol. 58, no. 12, pp. 6140–6155, Dec. 2010.

[24] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statist. Soc.. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.