A Bregman Matrix and the Gradient of Mutual Information for Vector Poisson and Gaussian Channels

Liming Wang, David Carlson, Miguel R.D. Rodrigues, Robert Calderbank and Lawrence Carin

Abstract

A generalization of Bregman divergence is developed and utilized to unify vector Poisson and Gaussian channel models, from the perspective of the gradient of mutual information. The gradient is with respect to the measurement matrix in a compressive-sensing setting, and mutual information is considered for signal recovery and classification. Existing gradient-of-mutual-information results for *scalar* Poisson models are recovered as special cases, as are known results for the vector Gaussian model. The Bregman-divergence generalization yields a Bregman *matrix*, and this matrix induces numerous matrix-valued metrics. The metrics associated with the Bregman matrix are detailed, as are its other properties. The Bregman matrix is also utilized to connect the relative entropy and mismatched minimum mean squared error. Two applications are considered: compressive sensing with a Poisson measurement model, and compressive topic modeling for analysis of a document corpora (word-count data). In both of these settings we use the developed theory to optimize the compressive measurement matrix, for signal recovery and classification.

Index Terms

Vector Poisson channels, vector Gaussian channels, gradient of mutual information, Bregman divergence, Bregman matrix, minimum mean squared error (MMSE).

I. Introduction

There is increasing interest in exploring connections between various quantities in information theory and estimation theory. Specifically, *mutual information* and *conditional mean estimation* have been

L. Wang, D. Carlson, R. Calderbank and L. Carin are with the Department of Electrical and Computer Engineering, Duke University, Durham, NC 27708 USA (email: {liming.w, david.carlson, robert.calderbank, lcarin}@duke.edu).

M. R. D. Rodrigues is with Department of Electronic and Electrical Engineering, University College London, London, U.K. (e-mail: m.rodrigues@ucl.ac.uk).

discovered to possess close interrelationships. Guo, Shamai and Verdú [1] have expressed the derivative of mutual information in a scalar Gaussian channel via the (nonlinear) *minimum mean-squared error* (MMSE). The connections have also been extended from the scalar Gaussian to the scalar Poisson channel model, the latter utilized to model optical communication systems [2]. Palomar and Verdú [3] have expressed the gradient of mutual information in a vector Gaussian channel in terms of the MMSE matrix. It has also been found that the relative entropy can be represented in terms of the mismatched MMSE estimates [4], [5]. Recently, parallel results for scalar binomial and negative binomial channels have been established [6], [7]. Inspired by Kabanov's result [8], [9], it has been demonstrated that for certain channels, exploring the gradient of mutual information can be more tractable than evaluating the mutual information itself, while simultaneously being of practical ($e.g.$, gradient-descent) interest. Further, it has also been shown that the derivative of mutual information with respect to key system parameters also relates to the conditional mean estimates in other channel settings beyond the Gaussian and Poisson models [10].

This paper pursues this overarching theme for *vector* Poisson channels, and provides a unification of the gradient of mutual information for the vector Poisson and Gaussian channel models. In [11] the author provides a general review of developments for communication theory in Poisson channels. The filtering and detection problems for Poisson channels have been considered in [12], [13]. The capacity of a Poisson channel under various circumstances has been investigated in [14], [15], [16], [17], [18].

One of the goals of this paper is to generalize the gradient of mutual information from scalar to vector Poisson channel models. This generalization is relevant not only theoretically, but also from the practical perspective, in view of numerous applications of the vector Poisson channel model in X-ray [19] and document classification systems (based on word counts) [20]. In those applications, the Poisson channel matrix plays an essential role for dimensionality reduction, and it can be manipulated such that the measured signal maximally conveys information about the underlying input signal. Mutual information is often employed as an information-loss measure, and the gradient provides a means to optimize the mutual information with respect to specific system parameters (via gradient descent methods). The mutual information is considered from the perspectives of both signal recovery and classification (the latter associated with feature design [21]).

We also encapsulate under a unified framework the gradient of mutual information results for scalar

Gaussian channels, scalar Poisson channel and their vector counterparts. This encapsulation is inspired by recent results that express the derivative of mutual information in scalar Poisson channels as the average value of the Bregman divergence associated with a particular function, between the input and the conditional mean estimate of the input [22]. By constructing a *generalization* of the classical Bregman divergence, we extend these results from the scalar to the vector case. This generalization yields a Bregman matrix, that appears to be new to the best of our knowledge. The gradients of mutual information for the vector Poisson model and the vector Gaussian model, as well as the scalar counterparts, are then also expressed – akin to [22] – in terms of the average value of the Bregman matrix associated with a particular (vector) function; this loss function is between the input vector and the conditional mean estimate of the input vector.

We also study various properties of the Bregman matrix. Specifically, the properties of the matrix are shown to mimic those of the classical Bregman divergence, which include non-negativity, linearity, convexity, duality and optimality. Equipped with these properties, various results relying on the classical Bregman divergence may be extended to the multi-dimensional case. Additionally, it has been shown that re-expressing results via a Bregman divergence can often lead to enhancements to the speed of various optimization algorithms [23].

We demonstrate applications of our theoretical results in the context of Poisson compressive sensing, and compressive document classification (compressive "topic modeling"). In the former problem, the proposed results are utilized to design the sensing matrix, such that the compressive measurement maximally preserves the information contained in the source. Offline design of the compressive measurement matrix is considered, as well as sequential online design. In the document-modeling problem, the vector Poisson channel model is employed as a dimensionality-reduction methodology ($i.e.$, feature design) on document topics, and document classification is performed directly on the compressive data (this simplifies the clustering analysis, by easing computational requirements). In this context, rather than characterizing documents in terms of counts of single words, the proposed theory is used to characterize documents in terms of counts of *groups* of words ($i.e.$, we perform learning of key words). Compared to randomly selecting the sensing matrix [24], the proposed method is shown to yield superior performance.

The remainder of the paper is organized as follows. Section II provides the definition of the vector Poisson channel, and reviews its vector Gaussian counterpart. We present gradients of mutual information

in Section III, in the context of both signal recovery and classification. In Section IV we introduce the Bregman matrix, and in Section V we use this matrix to unify gradient-of-mutual-information results for Gaussian and Poisson channels; properties of the Bregman matrix are also investigated. In Section VI we present two examples as verifications and applications of the proposed theoretical results. Concluding remarks are provided in Section VII.

Notation: Upper case letters denote random variables. Instantiations of the random variables are denoted by lower case letters. We use – except for the scaling matrix and the scaling factor – identical notation for the scalar Poisson channel and the vector Poisson channel. The context defines whether we are dealing with scalar or vector quantities. P_X denotes the probability measure induced by X, and $p_X(x)$ denotes the probability density function of X in case X admits one. The superscript $(\cdot)^T$ denotes the matrix transpose. Φ_{ij} and X_i denote the ij-th entry of a matrix Φ and i-th entry of a vector X, respectively. X_i^c denotes the collection of all but i-th entries of vector X, i.e., $X_i^c = \cup_{j\neq i}\{X_j\}$. $\mathrm{Tr}(\cdot)$ denotes the matrix trace. $\mathrm{Pois}(X; z)$ denotes a standard Poisson distribution with parameter z for random variable X. Product sign $\times$ is explicitly expressed only if the equation does not fit in one line, or it is needed to clearly mark a separation between two terms. We use the symbol $\mathbb{E}[\cdot\,|\,\cdot\,]$ to denote the conditional expectation of the first argument conditioned on the second argument. $\langle\cdot,\cdot\rangle$ is defined as the canonical inner product in $\mathbb{R}^k$, i.e., $\langle x, y\rangle = x^T y$ for $x, y \in \mathbb{R}^k$.

II. The Vector Poisson and Gaussian Channel Models

A. Vector Poisson channel

The vector Poisson channel model is defined as

$$\mathrm{Pois}(Y; \Phi X + \lambda) = P_{Y|X}\left(Y|X\right) = \prod_{i=1}^{m} P_{Y_i|X}\left(Y_i|X\right) = \prod_{i=1}^{m} \mathrm{Pois}\left(Y_i; (\Phi X)_i + \lambda_i\right) \tag{1}$$

where the random vector $X = (X_1, X_2, \ldots, X_n) \in \mathbb{R}_+^n$ represents the channel input, the random vector $Y = (Y_1, Y_2, \ldots, Y_m) \in \mathbb{Z}_+^m$ represents the channel output, the matrix $\Phi \in \mathbb{R}_+^{m\times n}$ represents a linear transformation whose role is to entangle the different inputs, and the vector $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_m) \in \mathbb{R}_+^m$ represents the dark current.

The vector Poisson channel model associated with arbitrary m and n is a generalization of the standard

scalar Poisson model associated with $m = n = 1$, as given by [2], [22]:

$$P_{Y|X}(Y|X) = \text{Pois}(Y; \phi X + \lambda) \tag{2}$$

where the scalar random variables $X \in \mathbb{R}_+$ and $Y \in \mathbb{Z}_+$ are associated with the input and output of the scalar channel, respectively, $\phi \in \mathbb{R}_+$ is a scaling factor, and $\lambda \in \mathbb{R}_+$ is the dark current.

The goal is to design Φ with the objective of maximizing the mutual information between X and Y. Toward that end, we consider the gradient of mutual information with respect to Φ:

$$\nabla_\Phi I(X;Y) = [\nabla_\Phi I(X;Y)_{ij}] \tag{3}$$

where $\nabla_\Phi I(X;Y)_{ij}$ represents the (i,j)-th entry of the matrix $\nabla_\Phi I(X;Y)$. As a parallel result to the scalar case presented in [2], we also consider the gradient with respect to the vector dark current

$$\nabla_\lambda I(X;Y) = [\nabla_\lambda I(X;Y)_i] \tag{4}$$

where $\nabla_\lambda I(X;Y)_i$ represents the i-th entry of the vector $\nabla_\lambda I(X;Y)$.

It may occur that the distribution of the signal X is constituted by a mixture of components, $i.e.$,

$$P_X(X) = \sum_{i=1}^{L} \pi_i P_{X|C=i}(X|C=i), \tag{5}$$

where π is a probability mass function supported on $C \in \{1, 2, \ldots, L\}$. This is the setting of interest for L-class classification problems. There may be more interest in recovering C than in recovering X, and in that setting one is interested in the mutual information between the class label C and the output Y. For that case we seek

$$\nabla_\Phi I(C;Y) = [\nabla_\Phi I(C;Y)_{ij}]\,. \tag{6}$$

The mutual information $I(X;Y)$ is termed the mutual information for signal recovery, while $I(C;Y)$ is called the mutual information for classification. The choice of mutual information as the metric is motivated by theoretical properties of mutual information, specifically that the MMSE and Bayesian classification error can be bounded via the mutual information [25], [26], [27].

B. Vector Gaussian channel

Below we will develop a theory specifically for the gradient of mutual information for vector Poisson signal models, and make connections to existing results for the special case of a scalar Poisson model. In addition, we will unify the vector Poisson and vector Gaussian channel models under a new theory, employing a new Bregman matrix. We therefore briefly review the Gaussian channel and its various gradient results.

The vector Gaussian channel model is given by:

$$Y|X \sim \mathcal{N}(\Phi X, \Lambda^{-1}), \tag{7}$$

where $\mathcal{N}(\cdot,\cdot)$ denotes the multivariate Gaussian distribution with corresponding mean vector and covariance matrix, $X \in \mathbb{R}^n$ represents the vector-valued channel input, $Y \in \mathbb{R}^m$ represents the vector-valued channel output, $\Phi \in \mathbb{R}^{m\times n}$ represents the channel matrix, and Λ^{-1} is a covariance matrix associated with the zero-mean Gaussian noise. Note that in both the vector Poisson and vector Gaussian models ΦX is the mean observation; the Gaussian model has the additional parameter of the covariance matrix Λ^{-1} (for the scalar Poisson case the mean and variance are equal). In the Gaussian case X and Φ can have both positive and negative components, whereas in the Poisson case both are non-negative.

It has been established that the gradient of mutual information between the input and the output of the vector Gaussian channel model in (7), with respect to the channel matrix, obeys the relationship [3]:

$$\nabla_\Phi I(X;Y) = \Lambda\Phi E, \tag{8}$$

where

$$E = \mathbb{E}\left[(X - \mathbb{E}(X|Y))(X - \mathbb{E}(X|Y))^T\right] \tag{9}$$

denotes the MMSE matrix.

The gradient of mutual information between the class label and the output for the vector Gaussian channel, with respect to the channel matrix, is [21]

$$\nabla_\Phi I(C;Y) = \Lambda\Phi\tilde{E}, \tag{10}$$

where

$$\tilde{E} = \mathbb{E}\left[(\mathbb{E}(X|Y,C) - \mathbb{E}(X|Y))(\mathbb{E}(X|Y,C) - \mathbb{E}(X|Y))^T\right] \tag{11}$$

denotes the equivalent MMSE matrix.

The above gradient results for the Gaussian channel are valid for any $P(X)$ consistent with regularity conditions [3].

III. Gradient of Mutual Information for Vector Poisson Channels

A. Gradient of Mutual Information for Signal Recovery

We now present the gradient of mutual information with respect to the measurement matrix Φ and with respect to the dark current λ for vector Poisson channel models. In order to take full generality of the input distribution into consideration, we utilize the Radon-Nikodym derivatives to represent the probability measures of interests. Consider random variables $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$. Let $f^{\theta}_{Y|X}$ be the Radon-Nikodym derivative of probability measure $P^{\theta}_{Y|X}$ with respect to an arbitrary measure Q_Y, provided that $P^{\theta}_{Y|X}$ is absolutely continuous with respect to Q_Y, *i.e.*, $P^{\theta}_{Y|X} \ll Q_Y$. $\theta \in \mathbb{R}$ is a parameter. f^{θ}_{Y} is the Radon-Nikodym derivative of the probability measure P^{θ}_{Y} with respect to Q_Y provided that $P^{\theta}_{Y} \ll Q_Y$. Note that in the continuous or discrete case, $f^{\theta}_{Y|X}$ and f^{θ}_{Y} are simply probability density or mass functions with Q_Y chosen to be the Lebesgue measure or the counting measure, respectively. We note that similar notation is also used for the signal classification case, except that we may also need to condition both on X and C. Some results of the paper require the assumption of the regularity conditions (RC), which are listed in Appendix A. We will assume all four regularity conditions RC1–RC4 whenever necessary in the proof and the statement of the results. Recall [28] that for a function $f(x,\theta) : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}$ with a Lebesgue measure μ on $\mathbb{R}^n$, we have $\frac{\partial}{\partial\theta}\int f(x,\theta)d\mu(x) = \int \frac{\partial}{\partial\theta} f(x,\theta)d\mu(x)$, if $|\frac{\partial}{\partial\theta} f(x,\theta)| \leq g(x)$, $\forall x, \theta$, where $g \in L^1(\mu)$. Hence, in light of this criterion, it is straightforward to verify that the RC are valid for many common distributions of X.

Theorem 1. *Consider the vector Poisson channel model in* (1)*. The gradient of mutual information*

between the input and output of the channel, with respect to the matrix Φ, is given by:

$$\begin{aligned}[\nabla_{\Phi} I(X;Y)_{ij}] =& \big[\mathbb{E}\left[X_j \log((\Phi X)_i + \lambda_i)\right] \\ & - \mathbb{E}\left[\mathbb{E}[X_j|Y] \log \mathbb{E}[(\Phi X)_i + \lambda_i|Y]\right]\big] \qquad (12) \\ =& \mathbb{E}\left[X_j \log\left(\frac{(\Phi X)_i + \lambda_i}{\mathbb{E}[(\Phi X)_i + \lambda_i|Y]}\right)\right], \qquad (13)\end{aligned}$$

and with respect to the dark current is given by:

$$\begin{aligned}[\nabla_{\lambda} I(X;Y)_{i}] =& \big[\mathbb{E}[\log((\Phi X)_i + \lambda_i)] \\ & - \mathbb{E}[\log \mathbb{E}[(\Phi X)_i + \lambda_i|Y]]\big]. \qquad (14)\end{aligned}$$

irrespective of the input distribution $P_X(X)$, provided that the regularity conditions in Appendix A hold.

Proof. See Appendix B. □

It is clear that Theorem 1 represents a multi-dimensional generalization of Theorems 1 and 2 in [2]. The scalar result follows immediately from the vector counterpart by taking $m = n = 1$.

Corollary 1. *Consider the scalar Poisson channel model in* (2)*. The derivative of mutual information between the input and output of the channel with respect to ϕ is given by:*

$$\begin{aligned}\frac{\partial}{\partial \phi} I(X;Y) =& \mathbb{E}\left[X \log((\phi X) + \lambda)\right] \\ & - \mathbb{E}\left[\mathbb{E}[X|Y] \log \mathbb{E}[\phi X + \lambda|Y]\right] \qquad (15) \\ =& \mathbb{E}\left[X \log\left(\frac{\phi X + \lambda}{\mathbb{E}[\phi X + \lambda|Y]}\right)\right], \qquad (16)\end{aligned}$$

and with respect to the dark current is given by:

$$\begin{aligned}\frac{\partial}{\partial \lambda} I(X;Y) =& \mathbb{E}[\log(\phi X + \lambda)] \\ & - \mathbb{E}[\log \mathbb{E}[\phi X + \lambda|Y]]. \qquad (17)\end{aligned}$$

irrespective of the input distribution $P_X(X)$, provided that the regularity conditions in Appendix A hold.

It is also of interest to note that the gradient of mutual information for vector Poisson channels appears to admit an interpretation akin to that of the gradient of mutual information for vector Gaussian channels

in (8) and (9) (see also [3]). In particular, both gradient results can be expressed in terms of the average of a multi-dimensional measure of the error between the input vector and the conditional mean estimate of the input vector under appropriate loss functions. This interpretation can be made precise – as well as unified – by constructing a generalized notion of Bregman divergence that encapsulates the classical one; the new form is a Bregman *matrix*. We consider this in Sections IV and V.

B. Gradient of Mutual Information for Classification

Theorem 2. *Consider the vector Poisson channel model in* (1) *and signal model in* (5)*. The gradient with respect to* Φ *of the mutual information between the class label and output of the channel is*

$$[\nabla_\Phi I(C;Y)_{ij}] = \mathbb{E}\left[\mathbb{E}[X_j|Y,C]\log\frac{\mathbb{E}[(\Phi X)_i+\lambda_i|Y,C]}{\mathbb{E}[(\Phi X)_i+\lambda_i|Y]}\right], \tag{18}$$

and with respect to the dark current is given by

$$(\nabla_\lambda I(C;Y))_i = \mathbb{E}\left[\log\frac{\mathbb{E}[(\Phi X)_i+\lambda_i|Y,C]}{\mathbb{E}[(\Phi X)_i+\lambda_i|Y]}\right]. \tag{19}$$

irrespective of the input distribution $P_{X|C}(X|C)$*, provided that the regularity conditions in Appendix A hold.*

Proof. See Appendix C. □

IV. Generalization of Bregman Divergence: the Bregman Matrix

A. Preliminaries

The classical Bregman divergence was originally constructed to determine common points of convex sets [29]. It was discovered later that the Bregman divergence induces numerous well-known metrics and has a bijection to the exponential family [30].

Definition 1 (Classical Bregman Divergence [29])**.** *Let* $F:\Omega\to\mathbb{R}_+$ *be a continuously-differentiable real-valued and strictly convex function defined on a closed convex set* $\Omega\subset\mathbb{R}^k$*. The Bregman divergence between* $x,y\in\Omega$ *is defined as*

$$D_F(x,y) := F(x) - F(y) - \langle\nabla F(y), x-y\rangle. \tag{20}$$

Note that different choices of the function F induce different metrics. For example, Euclidean distance, Kullback-Leibler divergence, Mahalanobis distance and many other widely-used distances are specializations of the Bregman divergence, associated with different choices of the function F [30].

There exist several generalizations of the classical Bregman divergence, including the extension to functional spaces [31] and a sub-modular extension [32]. However, such generalizations aim to extend the domain rather than the range of the Bregman divergence. This renders such generalizations unsuitable to problems where the "error" term is multi-dimensional rather than uni-dimensional, $e.g.$, the MMSE matrix in (9).

We now construct a generalization that extends the range of a Bregman divergence from scalar to matrix spaces (viewed as multi-dimensional vector spaces), to address the issue. We refer to this as the Bregman matrix. We start by reviewing several notions that are useful for the definition of the Bregman matrix.

Definition 2 (Generalized Inequality [33])**.** *Let $F : \Omega \to \mathbb{R}^{m\times n}$ be a continuously-differentiable function, where $\Omega \subset \mathbb{R}^l$ is a convex subset. Let $K \subset \mathbb{R}^{m\times n}$ be a proper cone, i.e., K is convex, closed, with non-empty interior and pointed. We define a partial ordering $\preceq_K$ on $\mathbb{R}^{m\times n}$ as follows. $\forall x, y \in K$, we have*

$$x \preceq_K y \Longleftrightarrow y - x \in K, \tag{21}$$

$$x \prec_K y \Longleftrightarrow y - x \in \mathrm{int}(K), \tag{22}$$

where $\mathrm{int}(\cdot)$ denotes the interior of the set. We write $x \succeq_K y$ and $x \succ_K y$ if $y \preceq_K x$ and $y \prec_K x$, respectively.

We define F to be K-convex if and only if:

$$F(\theta x + (1-\theta)y) \preceq_K \theta F(x) + (1-\theta)F(y) \tag{23}$$

for $\forall x, y \in \Omega$ and $\theta \in [0, 1]$.

We define F to be strictly K-convex if and only if:

$$F(\theta x + (1-\theta)y) \prec_K \theta F(x) + (1-\theta)F(y) \tag{24}$$

for $\forall x, y \in \Omega$ with $x \neq y$ and $\theta \in (0, 1)$.

Definition 3 (Fréchet Derivative [28]). *Let V and Z be Banach spaces with norms $\|\cdot\|_V$ and $\|\cdot\|_Z$, respectively, and let $U \subset V$ be open. $F : U \to Z$ is called* Fréchet differentiable *at $x \in U$ if there exists a bounded linear operator $DF(x)(\cdot) : V \to Z$ such that*

$$\lim_{\|h\|_V \to 0} \frac{\|F(x+h) - F(x) - DF(x)(h)\|_Z}{\|h\|_V} = 0. \tag{25}$$

$DF(x)(\cdot)$ is called the Fréchet derivative *of F at x.*

Note that the Fréchet derivative corresponds to the usual derivative of matrix calculus for finite dimensional vector spaces. However, by employing the Fréchet derivative, it is also possible to make extensions from finite to infinite dimensional spaces, such as L^p spaces.

B. Definition, Interpretation and Properties

We are now in a position to offer a definition of the Bregman matrix.

Definition 4. *Let $K \subset \mathbb{R}^{m\times n}$ be a proper cone and Ω be an open convex subset in a Banach space W. $F : \Omega \to \mathbb{R}^{m\times n}$ is a Fréchet-differentiable strictly K-convex function. The Bregman matrix $D_F(x, y)$ associated with the function F between $x, y \in \Omega$ is defined as follows:*

$$D_F(x, y) := F(x) - F(y) - DF(y)(x - y), \tag{26}$$

where $DF(y)(\cdot)$ is the Fréchet derivative of F at y.

This notion of a Bregman matrix is able to incorporate various previous extensions depending on the choices of the proper cone K and the Banach space W. For example, if we choose K to be the first quadrant (all coordinates are non-negative), we have the entry-wise convexity extension. In this case, the Bregman matrix is essentially equivalent to a matrix-valued function with each entry being a classical Bregman divergence and the Fréchet derivative becomes the ordinary Jacobian. If we choose K to be the space of positive-definite bounded linear operators, we have the positive definiteness extension. By choosing W to be an L^p space, then the definition is similar to that in [31]. If we choose W to be the space of real matrices and K be the collection of positive real numbers, it generalizes the results in [34].

Rather than being viewed as a simple loss function, the Bregman matrix admits an interesting geometric interpretation since it can be understood as a matrix-valued pseudo Riemannian metric (or a tensor metric)

[35]. To see this, let us consider two points $z \in \Omega$ and $z + dz \in \Omega$ with $\|dz\|$ being close to 0. Consider the first-order functional Taylor's expansion for $F(x)$ at z

$$F(x) = F(z) + DF(z)(x - z) + \frac{1}{2}[D^2F(z)(x - z)](x - z) + R_n(x), \tag{27}$$

where $R_n(x)$ is the residue term with $R_n(x) = o(\|x - z\|^2)$ and $D^2F : \Omega \to L(\Omega, L(\Omega, \mathbb{R}^{m\times n}))$ is the second-order Fréchet derivative [36] which corresponds to the Fréchet derivative of DF, and $L(\cdot, \cdot)$ denotes the collection of all bounded linear operators from the first argument space to the second argument space. When Ω is in a finite dimension vector space and $m = n = 1$, D^2F is the ordinary Hessian matrix.

Let us now calculate the Bregman matrix between z and $z + dz$,

$$\begin{aligned} D_F(z, z + dz) &= F(z) - F(z + dz) - DF(z + dz)(z - z - dz) && (28)\\ &= F(z) - F(z) - DF(z)(dz) - \frac{1}{2}[D^2F(z)(dz)](dz) \\ &+ DF(z)(dz) + [D^2F(z)(dz)](dz) + o(\|dz\|^2) && (29)\\ &\approx \frac{1}{2}[D^2F(z)(dz)](dz), && (30) \end{aligned}$$

where (29) uses the Taylor's expansion of F and DF. This formula indicates that the Bregman matrix can be infinitesimally viewed as a matrix-valued pseudo metric whose metric matrix is imposed by D^2F. If $m = n = 1$ and Ω is a finite dimensional vector space, then D^2F becomes the Hessian matrix and we have

$$\begin{aligned} D_F(z, z + dz) &\approx \frac{1}{2}[D^2F(z)(dz)](dz) && (31)\\ &= \frac{1}{2}(dz)^T D^2F(z)(dz), && (32) \end{aligned}$$

where $D^2F(z)$ serves as the Riemannian metric. This interpretation for classical Bregman divergence has also been recognized in [37]. Therefore, the Bregman matrix $D_F(\cdot, \cdot)$ can be viewed locally as a matrix-valued pseudo Riemannian metric function whose geometry information is induced solely by F [38].

The Bregman matrix also inherits various properties akin to the properties of the classical Bregman divergence, that has led to its wide utilization in optimization and computer vision problems [23], [39].

Theorem 3. *Let $K \subset \mathbb{R}^{m\times n}$ be a proper cone and Ω be an open convex subset in a Banach space W. $F : \Omega \to \mathbb{R}^{m\times n}$ and $G : \Omega \to \mathbb{R}^{m\times n}$ are Fréchet-differentiable strictly K-convex functions. Then the Bregman matrix $D_F(x, y)$ between $x, y \in \Omega$ associated with the function F exhibits the properties:*

1) $D_F(x, y) \succeq_K \mathbf{0}$*, where* $\mathbf{0}$ *is the zero matrix.*
2) $D_{c_1F+c_2G}(x, y) = c_1D_F(x, y) + c_2D_G(x, y)$*, where the constants* c_1, $c_2 > 0$.
3) $D_F(\cdot, y)$ *is K-convex for any* $y \in \Omega$.

Proof. See Appendix D. □

The Bregman matrix also exhibits a duality property similar to the duality property of the classical Bregman divergence when we choose some proper cone K, that may be useful for many optimization problems [39], [40].

Theorem 4. *Let $F : \Omega \to \mathbb{R}^{m\times n}$ be a strictly K-convex function, where $\Omega \subset \mathbb{R}^k$ is a convex subset. Choose K to be the first quadrant space $\mathbb{R}^{m\times n}_+$ (space formed by matrices with all entries positive). Let $(F^\star, x^\star, y^\star)$ be the Legendre transform of (F, x, y). Then, we have that:*

$$D_F(x, y) = D_{F^\star}(y^\star, x^\star). \tag{33}$$

Proof. See Appendix D. □

Via this theorem, it is possible to simplify the calculation of the Bregman divergence in scenarios where the dual form is easier to calculate than the original form. Mirror-descent methods, which have been shown to be computationally efficient for many optimization problems [39], [41], leverage this idea.

The Bregman matrix also exhibits another property akin to that of the classical Bregman divergence. In particular, it has been shown that for a metric that can be expressed in terms of the classical Bregman divergence, the optimal error relates to the conditional mean estimator of the input [42]. Similarly, it can also be shown that for a metric that can be expressed in terms of a Bregman matrix, the optimal error also relates to the conditional mean estimator of the input. However, this generalization from the scalar to the vector case requires the partial order interpretation of the minimization.

Theorem 5. *Consider a probability space $(\mathcal{S}, s, \mu)$, where s is the σ-algebra of $\mathcal{S}$ and μ is a probability measure on s. Let $F : \Omega \to \mathbb{R}^{m\times n}$ be strictly K-convex as before and Ω is a convex subset in a Banach*

space W. Let $X : \mathcal{S} \to \Omega$ be a random variable with $\mathbb{E}\left[\|X\|_2\right] < \infty$ and $\mathbb{E}\left[\|F(X)\|_2\right] < \infty$. Let $s_1 \subset s$ be a sub σ-algebra. Then, for any s_1-measurable random variable Y, we have that:

$$\arg\min_{Y} \mathbb{E}_{X,Y}\left[D_F(X,Y)\right] = \mathbb{E}\left[X|s_1\right], \tag{34}$$

where the minimization is interpreted in the partial ordering sense, i.e., if $\exists Y'$ such that $\mathbb{E}_{X,Y'}[D_F(X,Y')] \preceq_K \mathbb{E}[D_F(X,\mathbb{E}\left[X|s_1\right])]$, then $Y' = \mathbb{E}\left[X|s_1\right]$.

Proof. See Appendix D. □

Various properties of the Bregman matrix enable the possibility to extend previous methods based on the classical Bregman divergence to multi-dimensional cases. Here we illustrate an application of the Bregman matrix in mirror-descent methods. We first briefly review the motivation of mirror-descent methods. The regularized gradient method can be viewed as an approximation of a given function $f(x)$ at x_t by a quadratic function $f_t(x)$ as follows.

$$f_t(x) := f(x_t) + \langle \nabla_x f(x_t), x - x_t)\rangle + \frac{1}{2}(x - x_t)^T(x - x_t), \tag{35}$$

where the last term is the regularization. The general idea of mirror-descent methods is to replace that term by the Bregman divergence. For example, given a Bregman matrix $D_F(x,y)$, we can derive the following mirror descent function:

$$f_t(x) := f(x_t) + \langle \nabla_x f(x_t), x - x_t)\rangle + \mathrm{Tr}\left((D_F(x,x_t))(D_F(x,x_t)^T\right), \tag{36}$$

where the concrete choice of the associated function F depends on the nature of specific problems. It often occurs that the Bregman matrix is difficult to calculate directly in practice. Hence, rather than calculating the Bregman matrix itself, one may work directly on its dual form by Theorem 4, provided that it is easier to calculate the dual form. This idea for mirror-descent methods has been shown to be very computationally efficient and has been successfully implemented in many very large-scale optimization problems [39], [41].

V. Unification: A Bregman Matrix Perspective

Using the Bregman matrix, the gradient of mutual information for both vector Gaussian and Poisson channel models can be formulated into a unified framework. The interpretation of the gradient results for

vector Poisson and vector Gaussian channels, $i.e.$, as the average of a multi-dimensional generalization of the error between the input vector and the conditional mean estimate of the input vector, under appropriate loss functions, together with the properties of the Bregman matrix, pave the way to the unification of the various theorems.

Theorem 6. *Assume that the distribution of the input X satisfies the regularity conditions in Appendix A. The gradient of mutual information with respect to Φ for the vector Poisson channel model in* (1) *can be represented as follows:*

$$\nabla_{\Phi} I(X;Y) = \mathbb{E}\left[D_F(X, \mathbb{E}[X|Y])\right], \tag{37}$$

where $D_F(\cdot,\cdot)$ is a Bregman matrix associated with a strictly K-convex function

$$F(x) = x(\log(\Phi x + \lambda))^T - [x, \ldots, x] + [\mathbf{1}, \ldots, \mathbf{1}], \tag{38}$$

where $\mathbf{1} = [1, \ldots, 1]$.

The gradient of mutual information with respect to Φ for the vector Gaussian channel model in (7) *can be represented as follows:*

$$\nabla_{\Phi} I(X;Y) = \mathbb{E}\left[D_F(X, \mathbb{E}[X|Y])\right], \tag{39}$$

where $D_F(\cdot,\cdot)$ is a Bregman matrix associated with a strictly K-convex function

$$F(x) = \Lambda \Phi x x^T. \tag{40}$$

Proof. See Appendix E. □

Atar and Weissman [22] have also recognized that the derivative of mutual information with respect to the scaling for the scalar Poisson channel could also be represented in terms of a (classical) Bregman divergence. Such a result, applicable to the scalar Poisson channel as well as a result applicable to the scalar Gaussian channel, can be seen to be corollaries to Theorem 6; this is in view of the fact that the classical Bregman divergence is a specialization of the generalized one.

Corollary 2. *Assume that the distribution of input X satisfies the regularity conditions in Appendix A. The derivative of mutual information with respect to the scaling factor for the scalar Poisson channel*

model is given by:

$$\frac{\partial}{\partial \phi} I(X;Y) = \mathbb{E}\left[D_F(X, \mathbb{E}[X|Y])\right], \tag{41}$$

where $F(x) = x\log(\phi x + \lambda) - x + 1$.

Proof. By Theorem 6, $F(x) = x\log(\phi x + \lambda) - x + 1$. Hence $DF(x) = \log(\phi x + \lambda)$. It is straightforward to verify that $\mathbb{E}\left[D_F(X, \mathbb{E}[X|Y])\right]$ induces the scalar gradient result. □

Corollary 3. *Assume that the distribution of the input X satisfies the regularity conditions in Appendix A. The derivative of mutual information with respect to the scaling factor for the scalar Gaussian channel model is given by:*

$$\frac{\partial}{\partial \phi} I(X;Y) = \mathbb{E}\left[D_F(X, \mathbb{E}[X|Y])\right], \tag{42}$$

where $F(x) = \sigma^{-2}\phi x^2$. σ^2 is the variance of the noise.

Proof. By Theorem 6, $F(x) = \sigma^{-2}\phi x^2$. Equation (42) follows from a simple calculation, and the result from [3] that $\frac{\partial}{\partial \phi} I(X;Y) = \sigma^{-2}\phi \mathbb{E}[(X - \mathbb{E}(X|Y))^2]$ □

Similarly, the gradient of mutual information for classification under the vector Poisson and Gaussian channels can be incorporated into one framework, as the expected Bregman matrix between two conditional estimates.

Theorem 7. *Assume that the distribution of the input X satisfies the regularity conditions in Appendix A. We also assume the Gaussian channel model and Poisson channel model, as in* (1) *and* (7)*, respectively. Then the gradient of mutual information for classification $I(C;Y)$, with respect to Φ for the vector Poisson channel, can be represented as follows:*

$$\nabla_\Phi I(C;Y) = \mathbb{E}\left[D_F(\mathbb{E}[X|Y,C], \mathbb{E}[X|Y])\right], \tag{43}$$

where $D_F(\cdot,\cdot)$ is a Bregman matrix associated with a strictly K-convex function

$$F(x) = x(\log(\Phi x + \lambda))^T - [x, \ldots, x] + [\mathbf{1}, \ldots, \mathbf{1}], \tag{44}$$

where $\mathbf{1} = [1, \ldots, 1]$.

For the case of the vector Gaussian channel model, the gradient of mutual information for classification

$I(C;Y)$, with respect to Φ can be represented as follows:

$$\nabla_{\Phi} I(C;Y) = \mathbb{E}\left[D_F(\mathbb{E}[X|Y,C], \mathbb{E}[X|Y])\right], \tag{45}$$

where the associated function $F(x)$ is given by

$$F(x) = \Lambda \Phi x x^T. \tag{46}$$

Proof. See Appendix E. □

The Bregman matrix not only unifies various gradient results, it also serves as an interconnection between the mismatched estimation and relative entropy. It is found [4] that the relative entropy between two distributions can be calculated as an integral of the difference between mismatched mean square estimation errors for the Gaussian channel, and similar result holds for the scalar Poisson channel [22], where the relative entropy is represented as an integral of the difference between two classical Bregman divergences. The result for the vector Gaussian channel can also be represented in terms of the Bregman matrix, which is summarized in the following theorem.

Theorem 8. *Let P and Q be two arbitrary distributions for the random variable $X \in \mathbb{R}^n$. $N \sim \mathcal{N}(0, I)$ is a standard multivariate Gaussian random variable with zero mean and identity covariance matrix. Then,*

$$D(P\|Q) = \frac{1}{2}\mathbb{E}_P\left[\int_0^{\infty} Tr\left[D_F(\mathbb{E}_P[X|\sqrt{\gamma}X + N], \mathbb{E}_Q[X|\sqrt{\gamma}X + N])\right] \mathrm{d}\gamma\right], \tag{47}$$

where the associated function $F(x) = xx^T$. $\mathbb{E}_P(\cdot)$ and $\mathbb{E}_Q(\cdot)$ denote the expectations with the distribution of X being P and Q respectively.

Proof. See Appendix E. □

As discussed in the previous section, the Bregman matrix can be interpreted as a matrix-valued pseudo Riemannian metric. From this perspective, we now present a result which leverages the results in Theorem 6 and the Riemannian metric interpretation of the Bregman matrix. Let us assume that the family of feasible channel parameters Φ forms a manifold M_{Φ}. It follows from a classical argument via the implicit function theorem [43] that the function graph $\{(\Phi, I(X;Y))\}_{\Phi \in M_{\Phi}}$ is also a manifold whose dimension agrees with M_{Φ}, and we denote this mutual information manifold as M. We have the following theorem

where we assume that we have vectorized Φ, $D_F(X, \mathbb{E}[X|Y])$ and $\nabla_\Phi I(X;Y)$ from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^{mn\times 1}$.

Theorem 9. *Consider the mutual information manifold M with Riemannian metric $\mathfrak{I}$ for either Gaussian or Poisson channel, where $\mathfrak{I}$ is the identity matrix and the channel parameter manifold M_Φ is equipped with Riemannian metric $\mathfrak{g}$. There exists a Riemannian isometry between $(M_\Phi, \mathfrak{g})$ and $(M, \mathfrak{I})$, if we choose matrix $\mathfrak{g}$ to be the Riemannian metric defined by $\mathfrak{g} = diag\{[\mathbb{E}[D_F(X, \mathbb{E}[X|Y])]]^2\}$, where $[\cdot]^2$ denotes term-wise square of the argument vector and F is selected accordingly for either Gaussian or Poisson channel as in Theorem 6.*

Proof. See Appendix E. □

This theorem suggests that one may work directly with $(M_\Phi, \mathfrak{g})$ to investigate the properties of the mutual information under a proper choice of $\mathfrak{g}$. Meanwhile, we point out here that since the Riemannian geometry information is essentially controlled by function F, F is supposed to vary at each individual Φ as a Riemannian metric.

VI. Applications

We demonstrate numerical experiments on applications of the gradient results to the projection design for Poisson compressive sensing, on both synthetic and real datasets. One key assumption for our approach is that the distribution of input X is known. However, in practice, the distribution of X may not be directly available. This issue can be generally addressed in two different ways. The first approach is to learn the distribution based on training datasets, provided it is available. This can be readily carried out via the EM algorithm or Bayesian inference. The other way is to adaptively design Φ via a sequence of measurements which sequentially refines the estimate to the input distribution. In the following subsections, we present examples leveraging these two approaches.

A. Synthetic Data

We first apply the above theoretical results to synthetic data. Assume the signal $X \in \mathbb{R}_+^n$ has a log-normal mixture distribution and that the system model is

$$Y \sim \text{Pois}(Y; \Phi X), \qquad X = \exp(Z), \qquad Z|C \sim \mathcal{N}(Z; \mu_c, \Sigma_c), \qquad C \sim \sum_{k=1}^{L} \pi_k \delta_k, \tag{48}$$

where $X_i = \exp(Z_i)$ with X_i and Z_i denoting respectively the ith components of X and Z, $\pi_k > 0$, $\sum_{k=1}^{L} \pi_k = 1$, and δ_k is a unit point measure concentrated at k. The ith component of Y is drawn $Y_i \sim \text{Pois}[Y_i; (\Phi X)_i]$, where $(\Phi X)_i$ denotes the ith component of the vector ΦX. This model may also be expressed $Y|Z \sim \text{Pois}(Y; \Phi \exp(Z))$ with $Z \sim \sum_{k=1}^{L} \pi_k \mathcal{N}(Z; \mu_k, \Sigma_k)$. Therefore, X is modeled as a log-normal mixture model. In (48) we explicitly express the draw of class label C, as it is needed when interested in $I(Y; C)$.

We first consider the signal recovery problem, in which we wish to recover X based upon Y; in this case we seek the projection matrix Φ that maximizes $I(X; Y)$. We also consider the classification problem, for which we design Φ to maximize $I(Y; C)$. As we mentioned before, the mutual information in those cases does not possess an explicit formula except under very few special input distributions. In order to optimize the mutual information, we must resort to the gradient descent method and it is performed on Φ in both cases, with an added total-energy constraint, on $\text{Tr}(\Phi^T \Phi)$. Theorem 1 is employed to express gradients for maximization of $I(X; Y)$, while Theorem 2 is employed when maximizing $I(Y; C)$. Explicit formulas for optimizing the mutual information are available provided that the posterior density is known. As we will present later, the posterior density can be approximated by the Laplace method.

It is assumed that a single random vector $X \in \mathbb{R}_+^n$ is drawn, and $Y_i \sim \text{Pois}(Y_i; \phi_i^T X)$, where ϕ_i^T is a vector defined by the ith row of Φ. In "offline" design of Φ, all rows of Φ are designed at once, and $p(X)$ is defined by the model (48). In "online" design, each ϕ_1, ϕ_2, ... is constituted sequentially, with $p(X|\{Y_j\}_{j=1,...,i})$ employed when computing the mutual information for specification of ϕ_{i+1}. In the below experiments, we consider both offline and online design of Φ, with the expectation that online design will be better, since it adapts Φ to the signal X under test (with the added computational cost of sequential design).

Each of the Gaussian mixture components $\mathcal{N}(Z; \mu_k, \Sigma_k)$ may be viewed as a separate model for Z, with a total of L such models. The probabilities $\{\pi_k\}$ represent our prior belief about which model is responsible for any particular data sample. Since the prior for Z associated with each model is Gaussian, it is reasonable to also approximate the posterior of the model for Z as being Gaussian as well (this is done for each of the L models, and the cumulative model is a GMM). Considering each of the L models separately, the mean of the approximate posterior Gaussian is taken as the mode of the true posterior (maximum *a posterior*, or MAP, solution), and the covariance is taken as the Hessian of the model

parameters about the mode. This is termed the Laplace approximation [44], and this is implemented L times, once for each of the models (mixture components).

After acquiring $\{Y_j\}_{j=1,\ldots,i}$, our posterior with respect to X considers all L models (model averaging), and is represented

$$p(X|\{Y_j\}_{j=1,i}) = \sum_{k=1}^{L} p(k|\{Y_j\}_{j=1,\ldots,i}) p(X|k, \{Y_j\}_{j=1,\ldots,i}) \tag{49}$$

$$= \sum_{k=1}^{L} p(k|\{Y_j\}_{j=1,\ldots,i}) p(Z = \log X|k, \{Y_j\}_{j=1,\ldots,i}), \tag{50}$$

where $p(Z = \log X|k, \{Y_j\}_{j=1,\ldots,i})$ is manifested via the aforementioned Laplace approximation, denoted $\mathcal{N}(Z; \tilde{\mu}_k, \tilde{\Sigma}_k)$; $\tilde{\mu}_k$ and $\tilde{\Sigma}_k$ are respectively the Laplace-updated Gaussian mean and covariance matrix for mixture component (model) k. The prior belief about the probability of model k is $p(k) = \pi_k$, and the posterior is

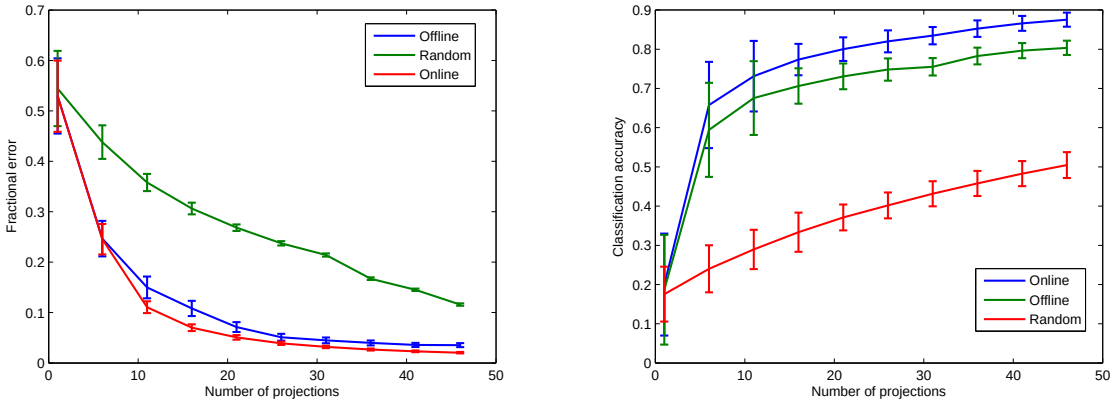
$$p(k|\{Y_j\}) = \frac{\pi_k \int_Z dZ p(\{Y_j\}|Z) p(Z|k)}{p(\{Y_j\})} = \frac{\pi_k \int_Z dZ \text{ Pois}(\{Y_j\}; \Phi e^Z) \mathcal{N}(Z; \tilde{\mu}_k, \tilde{\Sigma}_k)}{\sum_{k'=1}^{L} \pi_{k'} \int_{Z'} dZ' \text{ Pois}(\{Y_j\}; \Phi e^{Z'}) \mathcal{N}(Z'; \tilde{\mu}_{k'}, \tilde{\Sigma}_{k'})} \tag{51}$$

where $\{Y_j\}$ is here meant to concisely represent $\{Y_j\}_{j=1,\ldots,i}$. The integration with respect to Z in (51) is readily performed numerically with Monte Carlo integration.

The expressions (50)-(51) are used for online design, and they are also used to express our estimate of X based on a set of measurements (regardless of how Φ was designed). In the classification case, (51) is used to provide our estimate of which class C was responsible for the observed data Y.

Numerous examples of this type have been successfully tested with the analysis framework, one of which we elucidate here. We consider $L = 3$ mixture components, with $\pi_1 = 0.5$, $\pi_2 = 0.3$ and $\pi_3 = 0.2$. It is assumed that $n = 100$, and the respective n-dimensional means are $\mu_1 = (-1, \ldots, -1)$, $\mu_2 = (0, \ldots, 0)$, and $\mu_3 = (1, \ldots, 1)$. The covariances are set as $\Sigma_1 = \Sigma_2 = \Sigma_3 = AA^T + \sigma^2 I$, where I is the $n \times n$ identity matrix, σ^2 is a small variance that allows the covariance to be full rank, and $A \in \mathbb{R}^{100\times 50}$. Each entry of A was drawn i.i.d. from normal distribution $\mathcal{N}(0, 0.04)$. We employ 500 Monte Carlo samples to calculate the gradient and 500 iterations for the gradient descent. In terms of convergence speed, we find that our algorithm converges well after a few hundreds iterations.

In Fig. 1, we illustrate the means and variances of the fractional error and classification accuracy for the three methods with increasing number of projections, with results based on 100 independent simulations.



(a) (b)

Figure 1. Mean and standard deviation of the fractional error and classification accuracy for compressive Poisson measurements. The results are based on 100 runs. (a) Fractional error for signal recovery, (b) Average classification accuracy with increasing number of projections.

The fractional error is defined as $\frac{\|\hat{X}-X\|_2^2}{\|X\|_2^2}$. The energy constraint was $\text{Tr}(\Phi\Phi^T) = 1$, and in the case of random design, we draw each entry $\Phi_{ij} \sim \text{Gamma}(0.1, 0.1)$ to maintain the positivity and normalize Φ such that $\text{Tr}(\Phi\Phi^T) = 1$. From Figure 1 note that the designed Φ perform significantly better than random design, for both signal recovery and classification. Moreover, the online designed Φ performs better than its offline-designed counterpart, although the difference is not substantial.

B. Document Classification

In this example $W \in \mathbb{Z}_+^n$ represents counts of the occurrences of each of n words in a document. It is assumed that there are L classes of documents, and π_k represents the *a priori* probability of document class k. Class k is characterized by an n-dimensional probability vector over words, β_k, where $\sum_{w=1}^n \beta_{kw} = 1$, $\beta_{kw} \geq 0$, and β_{kw} represents the probability of word w in document class k. The draw of words for a given document in class k is assumed modeled $W \sim \text{Pois}(W; X)$, where $X = \gamma\beta_k$, with $\gamma \in \mathbb{R}_+$. Consequently, the total number of words $|W|$ associated with the document is assumed drawn $|W| \sim \text{Pois}(|W|; \gamma)$. Using a Poisson factor model [45], one may infer a set $\{\beta_k\}_{k=1,C}$ characteristic of a corpus. We henceforth assume that the set of probability vectors $\{\beta_k\}_{k=1,C}$ is known (learned based on training data as in [45]).

The number of words n in the dictionary $\mathcal{D}$ may be large. Rather than counting the number of times each of the n words are separately manifested, we may more efficiently count the number of times words in *subsets* of $\mathcal{D}$ are manifested (each subset of words acts like key words associated with a

given topic). Specifically, consider a compressive measurement for a document as $Y|X \sim \text{Pois}(Y; \Phi X)$, where $\Phi \in \{0,1\}^{m\times n}$, with $m \ll n$. Let $\phi_i^T \in \{0,1\}^n$ represent the ith row of Φ, with Y_i the ith component of Y. Then $Y_i|X \sim \text{Pois}(Y_i; \phi_i^T X)$ is equal in distribution to $Y_i = \sum_{j=1}^n \phi_{ij} W_j$, where $W_j|X \sim \text{Pois}(W_j; X_j)$, and $\phi_{ij} \in \{0,1\}$ is the jth component of vector ϕ_i. Therefore, Y_i represents the number of times words in the *set* defined by the non-zero elements of ϕ_i are manifested in the document; Y represents the number of times words are manifested within m distinct sets, with the sets defined by the non-zero elements in the rows of Φ.

Note that we use a *binary* Φ because the compressive measurements may be manifested by simply summing the number of words in the document associated with each of the m subsets of words. Hence, these compressive measurements may be constituted directly based on the observed count of words in a given document. We may also theoretically allow $\Phi \in \mathbb{R}_+^{m\times n}$, but we cannot usefully apply this result to observed documents.

For matrix $\Phi \in \{0,1\}^{m\times n}$, the overall compressive document measurement is represented

$$Y \sim \text{Pois}(Y; \Phi X) \ , \quad X = \gamma \beta_c \ , \quad \gamma \sim \text{Gamma}(a_0, b_0) \ , \quad C \sim \sum_{k=1}^{L} \pi_k \delta_k \tag{52}$$

where it is assumed that $\{\beta_k\}_{k=1,\ldots,L}$ are known. Computational methods like those discussed in [45] are used to infer γ and C based upon a compressive measurement Y. The goal is to design Φ with the goal of maximizing $I(X;Y)$ or $I(C;Y)$.

We use Theorems 1 and 2 to design a binary Φ. To do this, instead of directly optimizing Φ, we put a logistic link on each value $\Phi_{ij} = \text{logit}(M_{ij})$. We can state the gradient with respect to Φ as:

$$[\nabla_M I(X;Y)_{ij}] = [\nabla_\Phi I(X;Y)_{ij}][\nabla_M \Phi_{ij}] \tag{53}$$

To calculate the designed M, we first initialize the matrix at random. We use Monte Carlo integration to estimate the gradient and used a standard of 1000 gradient steps when the matrix had clearly converged. The step size was set to be 1/10 of the maximum of the absolute value of $\nabla_M I(X;Y)$. Finally, we threshold Φ at 0.5 to get the final binary Φ. We employ 500 Monte Carlo samples to calculate the gradient and 500 iterations for the gradient descent. In this experiment, our algorithm converges well after a few hundreds iterations.

To classify the documents, we use the maximum *a posteriori* (MAP) estimate, with our predicted class

$$c_i^\star = \arg\max_j p(c_i = j|y_i) \tag{54}$$

$$p(c_i = j|y_i) \propto \frac{\pi_j p(\gamma_i) p(y_i|c_i, \gamma_i)}{p(\gamma_i|c_i, y_i)}|_{\gamma_i = \gamma_i^*} \tag{55}$$

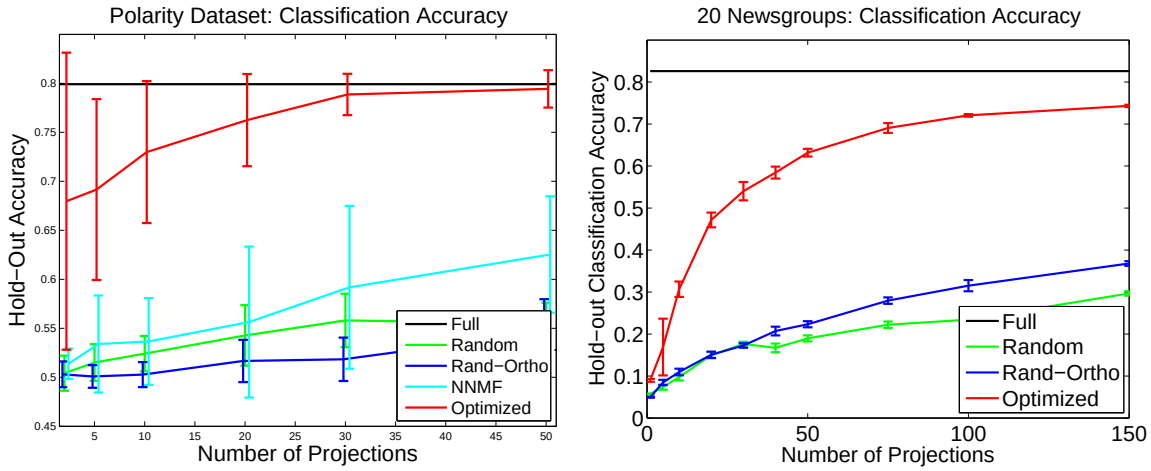
$$p(\gamma_i|c_i, y_i) = \text{Gamma}(a_0 + \sum_{j=1}^{m} y_{ij}, b_0 + ||\Phi d_{c_i}||_1) \tag{56}$$

where $\gamma_i^\star$ can be any positive real number.

We demonstrate designed projections for classification on the Polarity dataset [46] and the 20 Newsgroups corpus[1]. The Polarity dataset has $n = 8828$ unique words and two possible classes ($i.e.$, $L = 2$, corresponding to positive and negative sentiment), and the Newsgroup data has $n = 8052$ unique words with $L = 20$ different newsgroups. When learning the class-dependent $\{\beta_1, \ldots, \beta_L\}$, we placed the prior Dir$(0.1, \ldots, 0.1)$ for each β_k, and the components γ_i had a prior Gamma$(0.1, 0.1)$ [45]. To process and test the measurement design, we split into 10 groups of 60% training and 40% testing. We learn $\{\beta_1, \ldots, \beta_L\}$ on the training data, and use this along with the prior on γ to learn the measurement matrix via gradient descent. Classification versus number of projections for random projections and designed projections are shown in Figure 2. The random design was constituted by using a drawing each entry in the binary matrix from a Bernoulli random variable with $p = .05$. The results were robust to setting the p in the Bernoulli random variable between $.01$ and $.1$, and performance degraded outside that range. When we compare to the random-orthogonal projection matrix, we enforce that each row of the sensing matrix is orthogonal to all other rows. To do this, we draw each column in the matrix Φ from a multinomial distribution with probability $(\frac{1}{m}, \ldots, \frac{1}{m})$. This gives that each column has exactly one non-zero entry, and will give orthogonal factors. When considering non-negative matrix factorization (NNMF) [47], a NNMF is performed on the training count matrix by using the algorithm in [47]. (Performance of the heuristic NNMF projection matrices is also dependent on the algorithm used. The NNMF algorithm in [48] was also attempted, but the classification results were dominated by that of [47].) After getting the principal non-negative factors, we threshold the non-negative factors so that 5% of the factors are non-zero to get a design matrix. The results were robust to perturbations in the threshold value to set between 1% and 10% of the values to be non-zero. In Figure 2(a), we show the results on the Polarity

[1]http://qwone.com/~jason/20Newsgroups/

dataset. We obtain nearly identical performance to the fully observed case (no compression), after only $m = 50$ designed projections. Note that the designed case dominates the performance of the random cases and the heuristic design of non-negative matrix factorization (NNMF). In Figure 2(b), we show the results on the 20 Newsgroups dataset. The designed projections dominate the random projections and have similar performance to the fully observed count vectors with 150 projections. In this case, the NNMF greatly outperformed the purely random methods, but was once again significantly improved upon by the designed case.



(a) (b)

Figure 2. Error-bar plots of MAP classification accuracy with increasing number of projections. Random denotes a random binary matrix with 1% non-zero values. Rand-Ortho denotes a random binary matrix restricted to an orthogonal matrix with one non-zero entry per column. NNMF denotes using non-negative matrix factorization on the normalized counts to heuristically design the projections. Optimized denotes the design methods discussed in section VI-B. (a) Results on the polarity dataset. (b) Results on the 20 Newsgroups corpus.

Note that in the examples considered, the topic labels were given by the dataset, and our goal was classification. Comparison to performance based on using all the words (non-compressive measurements) is therefore the appropriate reference. In other applications one must learn the characteristics of the topics based upon a corpus ($i.e.$, one must learn which topic labels are appropriate). For that one may use one of the many types of topic models that have been developed in the literature [49], [50]. That was beyond the scope of this study, and was unnecessary for the datasets considered for demonstration of the proposed methods.

VII. Conclusions

The relationship between the mutual information and conditional mean estimator for vector Poisson channels has been examined. It has been shown that the gradients of mutual information with respect to the scaling matrix and dark current for vector Poisson channels can be formulated into a relatively simple form. By revealing the gradient of mutual information, it is possible to use gradient descent type optimization algorithms to solve problems in application areas associated with Poisson vector data ($e.g.$, word counts in documents). The results of this paper may be used for optimal design of compressive measurement matrices for compressive sensing with Poisson data models.

The Bregman matrix has been proposed to extend the range of the classical Bregman divergence to the multi-dimensional case. Several theoretical properties of the Bregman matrix have been analyzed, such as non-negativity, linearity, convexity and duality, which make it possible to extend many previous algorithms based on the classical Bregman divergence. We establish the connection between the Bregman matrix and the gradient of mutual information for both Gaussian and Poisson channels. The relative entropy and the mismatched MMSE can also be connected in terms of the Bregman matrix.

Acknowledgement

The research reported here was supported in part by the Defense Advanced Research Projects Agency (DARPA), under the KeCom program managed by Dr. Mark Neifeld.

Appendix A
Regularity Conditions

In this paper, we assume the following four regularity conditions (RC) on the interchangeability of integration and differentiation.

RC1:

$$\frac{\partial}{\partial\theta}\mathbb{E}_{Q_Y}\left[f^{\theta}_{Y|X}\right]=\mathbb{E}_{Q_Y}\left[\frac{\partial}{\partial\theta}f^{\theta}_{Y|X}\right], \tag{57}$$

RC2:

$$\frac{\partial}{\partial\theta}\mathbb{E}_{P_X}\left[f^{\theta}_{Y|X}\right]=\mathbb{E}_{P_X}\left[\frac{\partial}{\partial\theta}f^{\theta}_{Y|X}\right], \tag{58}$$

RC3:

$$\frac{\partial}{\partial\theta}\mathbb{E}_{P_XQ_Y}\left[f^{\theta}_{Y|X}\log f^{\theta}_{Y|X}\right]=\mathbb{E}_{P_XQ_Y}\left[\frac{\partial}{\partial\theta}\left(f^{\theta}_{Y|X}\log f^{\theta}_{Y|X}\right)\right]. \tag{59}$$

RC4:

$$\frac{\partial}{\partial\theta}\mathbb{E}_{Q_Y}\left[f^\theta_{Y|X}\log f^\theta_{Y|X}\right]=\mathbb{E}_{Q_Y}\left[\frac{\partial}{\partial\theta}\left(f^\theta_{Y|X}\log f^\theta_{Y|X}\right)\right]. \tag{60}$$

In addition, we always assume the technical condition that $\int\left[\left|\log\frac{\mathrm{d}P^\theta_Y}{\mathrm{d}Q_Y}\frac{\partial}{\partial\theta}\left(\frac{\mathrm{d}P^\theta_{Y|X}}{\mathrm{d}Q_Y}\right)\right|\mathrm{d}P_X\mathrm{d}Q_Y\right]<\infty$ and $\mathbb{E}[|X_j\log(\mathbb{E}[\phi_iX+\lambda_i|Y])|]<\infty$, $\forall i,j$.

APPENDIX B
PROOF OF THEOREM 1

We first establish the following Lemma which relates to the results in [10].

Lemma 1. *Consider random variables $X\in\mathbb{R}^n$ and $Y\in\mathbb{R}^m$. Let $f^\theta_{Y|X}$ be the Radon-Nikodym derivative of the probability measure $P^\theta_{Y|X}$ with respect to arbitrary measures Q_Y provided that $P^\theta_{Y|X}\ll Q_Y$. $\theta\in\mathbb{R}$ is a parameter. f^θ_Y is the Radon-Nikodym derivative of probability measure P^θ_Y with respect to Q_Y provided that $P^\theta_Y\ll Q_Y$. Assume the regularity conditions RC1 – RC4, we have*

$$\frac{\partial}{\partial\theta}I(X;Y)=\mathbb{E}\left[\frac{\partial\log f^\theta_{Y|X}}{\partial\theta}\log\frac{f^\theta_{Y|X}}{f^\theta_Y}\right]. \tag{61}$$

Proof of Lemma 1. Choose an arbitrary measure Q_Y such that $P^\theta_{Y|X}\ll Q_Y$ and $P^\theta_Y\ll Q_Y$.

$$\begin{aligned}\frac{\partial}{\partial\theta}I(X;Y)&=\frac{\partial}{\partial\theta}D(P^\theta_{Y|X}\|Q_Y)-D(P^\theta_Y\|Q_Y) &(62)\\ &=\frac{\partial}{\partial\theta}\left[\int\log\frac{\mathrm{d}P^\theta_{Y|X}}{\mathrm{d}Q_Y}\frac{\mathrm{d}P^\theta_{Y|X}}{\mathrm{d}Q_Y}\mathrm{d}Q_Y\mathrm{d}P_X-\int\log\frac{\mathrm{d}P^\theta_Y}{\mathrm{d}Q_Y}\frac{\mathrm{d}P^\theta_Y}{\mathrm{d}Q_Y}\mathrm{d}Q_Y\right] &(63)\\ &=\frac{\partial}{\partial\theta}\left[\int\log\frac{\mathrm{d}P^\theta_{Y|X}}{\mathrm{d}Q_Y}\mathrm{d}P^\theta_{Y|X}\mathrm{d}P_X-\int\log\frac{\mathrm{d}P^\theta_Y}{\mathrm{d}Q_Y}\frac{\mathrm{d}P^\theta_Y}{\mathrm{d}Q_Y}\mathrm{d}Q_Y\right]. &(64)\end{aligned}$$

We will calculate the two terms in (64) separately.

$$\begin{aligned}\frac{\partial}{\partial\theta}\left[\int\log\frac{\mathrm{d}P^\theta_{Y|X}}{\mathrm{d}Q_Y}\mathrm{d}P^\theta_{Y|X}\mathrm{d}P_X\right]&=\int\left[\frac{\partial}{\partial\theta}\left(\log\frac{\mathrm{d}P^\theta_{Y|X}}{\mathrm{d}Q_Y}\right)\mathrm{d}P^\theta_{Y|X}\mathrm{d}P_X\right]\\ &\quad+\int\left[\log\frac{\mathrm{d}P^\theta_{Y|X}}{\mathrm{d}Q_Y}\frac{\partial}{\partial\theta}\left(\frac{\mathrm{d}P^\theta_{Y|X}}{\mathrm{d}Q_Y}\right)\mathrm{d}Q_Y\mathrm{d}P_X\right], &(65)\end{aligned}$$

where the equality essentially follows from RC3. By Lemma 1 in [10], we have

$$\int\left[\frac{\partial}{\partial\theta}\left(\log\frac{\mathrm{d}P^\theta_{Y|X}}{\mathrm{d}Q_Y}\right)\mathrm{d}P^\theta_{Y|X}\mathrm{d}P_X\right]=0. \tag{66}$$

Hence,

$$\frac{\partial}{\partial\theta}\left[\int \log\frac{\mathrm{d}P^{\theta}_{Y|X}}{\mathrm{d}Q_Y}\mathrm{d}P^{\theta}_{Y|X}\mathrm{d}P_X\right] = \int\left[\log\frac{\mathrm{d}P^{\theta}_{Y|X}}{\mathrm{d}Q_Y}\frac{\partial}{\partial\theta}\left(\frac{\mathrm{d}P^{\theta}_{Y|X}}{\mathrm{d}Q_Y}\right)\mathrm{d}Q_Y\mathrm{d}P_X\right] \tag{67}$$

$$= \int\left[\log\frac{\mathrm{d}P^{\theta}_{Y|X}}{\mathrm{d}Q_Y}\frac{\partial}{\partial\theta}\left(\log\frac{\mathrm{d}P^{\theta}_{Y|X}}{\mathrm{d}Q_Y}\right)\mathrm{d}P^{\theta}_{Y|X}\mathrm{d}P_X\right]. \tag{68}$$

The second term in (64) can be calculated as follow.

$$\frac{\partial}{\partial\theta}\left[\int \log\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\mathrm{d}Q_Y\right] \overset{RC4}{=} \int\left[\frac{\partial}{\partial\theta}\left(\log\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\right)\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\mathrm{d}Q_Y\right] + \int\left[\log\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\frac{\partial}{\partial\theta}\left(\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\right)\mathrm{d}Q_Y\right] \tag{69}$$

$$= \int\left[\frac{\partial}{\partial\theta}\left(\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\right)\mathrm{d}Q_Y\right] + \int\left[\log\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\frac{\partial}{\partial\theta}\left(\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\right)\mathrm{d}Q_Y\right] \tag{70}$$

$$\overset{RC1}{=} \frac{\partial}{\partial\theta}\int \mathrm{d}P^{\theta}_{Y} + \int\left[\log\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\frac{\partial}{\partial\theta}\left(\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\right)\mathrm{d}Q_Y\right] \tag{71}$$

$$= 0 + \int\left[\log\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\frac{\partial}{\partial\theta}\left(\int\frac{\mathrm{d}P^{\theta}_{Y|X}}{\mathrm{d}Q_Y}\mathrm{d}P_X\right)\mathrm{d}Q_Y\right] \tag{72}$$

$$\overset{RC2}{=} \int\left[\log\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\left(\int\frac{\partial}{\partial\theta}\left(\frac{\mathrm{d}P^{\theta}_{Y|X}}{\mathrm{d}Q_Y}\right)\mathrm{d}P_X\right)\mathrm{d}Q_Y\right] \tag{73}$$

$$= \int\left[\log\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\frac{\partial}{\partial\theta}\left(\frac{\mathrm{d}P^{\theta}_{Y|X}}{\mathrm{d}Q_Y}\right)\mathrm{d}P_X\mathrm{d}Q_Y\right] \tag{74}$$

$$= \int\left[\log\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\frac{\partial}{\partial\theta}\left(\log\frac{\mathrm{d}P^{\theta}_{Y|X}}{\mathrm{d}Q_Y}\right)\mathrm{d}P^{\theta}_{Y|X}\mathrm{d}P_X\right], \tag{75}$$

where the second to the last equality follows from the assumption together with the Fubini's theorem. We denote the specific regularity condition used on top of the corresponding equality symbol. Plugging (68) and (75) back to (64), we have

$$\frac{\partial}{\partial\theta}I(X;Y) = \int\left[\log\frac{\mathrm{d}P^{\theta}_{Y|X}}{\mathrm{d}Q_Y}\frac{\partial}{\partial\theta}\left(\log\frac{\mathrm{d}P^{\theta}_{Y|X}}{\mathrm{d}Q_Y}\right)\mathrm{d}P^{\theta}_{Y|X}\mathrm{d}P_X\right] - \int\left[\log\frac{\mathrm{d}P^{\theta}_{Y}}{\mathrm{d}Q_Y}\frac{\partial}{\partial\theta}\left(\log\frac{\mathrm{d}P^{\theta}_{Y|X}}{\mathrm{d}Q_Y}\right)\mathrm{d}P^{\theta}_{Y|X}\mathrm{d}P_X\right] \tag{76}$$

$$= \int\left[\frac{\partial}{\partial\theta}\left(\log\frac{\mathrm{d}P^{\theta}_{Y|X}}{\mathrm{d}Q_Y}\right)\log\frac{\mathrm{d}P^{\theta}_{Y|X}/\mathrm{d}Q_Y}{\mathrm{d}P^{\theta}_{Y}/\mathrm{d}Q_Y}\mathrm{d}P^{\theta}_{Y|X}\mathrm{d}P_X\right] \tag{77}$$

$$= \mathbb{E}\left[\frac{\partial\log f^{\theta}_{Y|X}}{\partial\theta}\log\frac{f^{\theta}_{Y|X}}{f^{\theta}_{Y}}\right], \tag{78}$$

where the last equality follows from the definition of Radon-Nikodym derivatives $f^{\theta}_{Y|X}$ and f^{θ}_{Y}. $\square$

Proof of Theorem 1. Let the parameter $\theta = \Phi_{ij}$. We first choose a measure Q_Y such that $P^{\Phi_{ij}}_{Y|X} \ll Q_Y$ and $P^{\Phi_{ij}}_{Y} \ll Q_Y$. Let $f^{\Phi_{ij}}_{Y|X}$ and $f^{\Phi_{ij}}_{Y}$ be the Radon-Nikodym derivatives of $P^{\Phi_{ij}}_{Y|X}$ and $P^{\Phi_{ij}}_{Y}$, respectively. By Lemma 1, we have

$$\frac{\partial I(X;Y)}{\partial \Phi_{ij}} = \mathbb{E}\left(\frac{\partial}{\partial \Phi_{ij}} \log f^{\Phi_{ij}}_{Y|X}(Y|X) \times \log \frac{f^{\Phi_{ij}}_{Y|X}(Y|X)}{f^{\Phi_{ij}}_{Y}(Y)} \right) \tag{79}$$

$$= \mathbb{E}\left(\frac{\frac{\partial}{\partial \Phi_{ij}} f^{\Phi_{ij}}_{Y|X}(Y|X)}{f^{\Phi_{ij}}_{Y|X}(Y|X)} \log \frac{f^{\Phi_{ij}}_{Y|X}(Y|X)}{f^{\Phi_{ij}}_{Y}(Y)} \right). \tag{80}$$

Notice that by the Poisson channel assumption, Y is supported on $\mathbb{Z}^m_+$. If we choose the measure Q_Y to be the counting measure, then we have $f^{\Phi_{ij}}_{Y|X} = P^{\Phi_{ij}}_{Y|X}$ and $f^{\Phi_{ij}}_{Y} = P^{\Phi_{ij}}_{Y}$. Therefore, we have

$$\frac{\partial}{\partial \Phi_{ij}} f^{\Phi_{ij}}_{Y|X}(y|x) = \frac{\partial}{\partial \Phi_{ij}} \text{Pois}(y; \Phi x + \lambda) \tag{81}$$

$$= \left(\frac{1}{y_i!} y_i x_j (\phi_i x + \lambda_i)^{y_i - 1} e^{-(\phi_i x + \lambda_i)} + \frac{1}{y_i!} (\phi_i x + \lambda_i)^{y_i} (-x_j) e^{-(\phi_i x + \lambda_i)} \right)$$

$$\times \prod_{k \neq i} \frac{1}{y_k!} (\phi_k x + \lambda_k)^{y_k} e^{-(\phi_k x + \lambda_k)} \tag{82}$$

$$= \frac{1}{y_i!} x_j (\phi_i x + \lambda_i)^{y_i} e^{-(\phi_i x + \lambda_i)} \left(\frac{y_i}{\phi_i x + \lambda_i} - 1 \right) \prod_{k \neq i} \frac{1}{y_k!} (\phi_k x + \lambda_k)^{y_k} e^{-(\phi_k x + \lambda_k)} \tag{83}$$

$$= x_j \left(\frac{y_i}{\phi_i x + \lambda_i} - 1 \right) P^{\Phi_{ij}}_{Y|X}(y|x), \tag{84}$$

where ϕ_i is the i-th row of Φ.

Therefore, we have

$$\frac{\partial I(X;Y)}{\partial \Phi_{ij}} = \mathbb{E}\left(X_j \left(\frac{Y_i}{\phi_i X + \lambda_i} - 1 \right) \log \frac{P^{\Phi_{ij}}_{Y|X}(Y|X)}{P^{\Phi_{ij}}_{Y}(Y)} \right) \tag{85}$$

$$= \mathbb{E}\left(X_j \left(\frac{Y_i}{\phi_i X + \lambda_i} - 1 \right) \log P^{\Phi_{ij}}_{Y|X}(Y|X) \right) \tag{86}$$

$$- \mathbb{E}\left(X_j \left(\frac{Y_i}{\phi_i X + \lambda_i} - 1 \right) \log P^{\Phi_{ij}}_{Y}(Y) \right). \tag{87}$$

We will calculate (86) and (87) separately. In the following derivations, we will omit the superscript Φ_{ij} in $P^{\Phi_{ij}}_{Y|X}(Y|X)$ and $P^{\Phi_{ij}}_{Y}(Y)$ for simplicity.

Term (86) may be expressed as

$$\mathbb{E}\left[X_j\left(\frac{Y_i}{\phi_i X+\lambda_i}-1\right)\sum_k \log\left(\frac{1}{Y_k!}(\phi_i X+\lambda_i)^{Y_k}e^{-(\phi_i X+\lambda_i)}\right)\right] \tag{88}$$

$$=\sum_k \mathbb{E}\left[X_j\left(\frac{Y_i}{\phi_i X+\lambda_i}-1\right)\log\frac{1}{Y_k!}\right] \tag{89}$$

$$+\sum_k \mathbb{E}\left[X_j\left(\frac{Y_i}{\phi_i X+\lambda_i}-1\right)Y_k\log(\phi_i X+\lambda_i)\right] \tag{90}$$

$$-\sum_k \mathbb{E}\left[X_j\left(\frac{Y_i}{\phi_i X+\lambda_i}-1\right)(\phi_i X+\lambda_i)\right]. \tag{91}$$

We claim that (91) equals zero; this term may be expressed as

$$\sum_k \mathbb{E}\left[X_j\left(\frac{Y_i}{\phi_i X+\lambda_i}-1\right)(\phi_i X+\lambda_i)\right]$$

$$=\sum_k \mathbb{E}\left[\mathbb{E}\left[X_j\left(\frac{Y_i}{\phi_i X+\lambda_i}-1\right)(\phi_i X+\lambda_i)\middle| X\right]\right] \tag{92}$$

$$=\sum_k \mathbb{E}\left[X_j\left(\frac{\mathbb{E}[Y_i|X]}{\phi_i X+\lambda_i}-1\right)(\phi_i X+\lambda_i)\right] \tag{93}$$

$$=\sum_k \mathbb{E}\left[X_j\left(\frac{\phi_i X+\lambda_i}{\phi_i X+\lambda_i}-1\right)(\phi_i X+\lambda_i)\right] \tag{94}$$

$$=0, \tag{95}$$

where we use the fact that $\mathbb{E}[Y_i|X]=\mathbb{E}[\text{Pois}(Y_i;\phi_i X+\lambda_i)|X]=\phi_i X+\lambda_i$.

In turn, (86) may be expressed as

$$\sum_k \mathbb{E}\left[X_j\left(\frac{Y_i}{\phi_i X+\lambda_i}-1\right)\log\frac{1}{Y_k!}\right]+\sum_k \mathbb{E}\left[X_j\left(\frac{Y_i}{\phi_i X+\lambda_i}-1\right)Y_k\log(\phi_i X+\lambda_i)\right]. \tag{96}$$

Combining the fact that $\mathbb{E}[Y_i|X]=\phi_i X+\lambda_i$, $\mathbb{E}[Y_i^2|X]=(\phi_i X+\lambda_i)+(\phi_i X+\lambda_i)^2$ and $P_{Y|X}(y|x)=\prod_k P_{Y_k|X}(y_k|x)$, the latter term can be calculated as follow.

$$\sum_k \mathbb{E}\left[X_j\left(\frac{Y_i}{\phi_i X+\lambda_i}-1\right)Y_k\log(\phi_i X+\lambda_i)\right]$$

$$=\sum_k \int x_j\left(\frac{y_i}{\phi_i x+\lambda_i}-1\right)y_k\log(\phi_i x+\lambda_i)dP_X dP_{Y|X} \tag{97}$$

$$=\int x_j\left(\frac{y_i^2}{\phi_i x+\lambda_i}-y_i\right)\log(\phi_i x+\lambda_i)dP_X dP_{Y_i|X}$$

$$+\sum_{k\neq i}\int x_j\left(\frac{y_i}{\phi_i x+\lambda_i}-1\right)y_k\log(\phi_i x+\lambda_i)dP_X dP_{Y_i|X}dP_{Y_k|X} \tag{98}$$

$$=\int x_j\left(\frac{(\phi_i x+\lambda_i)^2+(\phi_i x+\lambda_i)}{\phi_i x+\lambda_i}-\phi_i x+\lambda_i\right)\log(\phi_i x+\lambda_i)dP_X$$

$$+\sum_{k\neq i}\int x_j\left(\frac{\phi_i x+\lambda_i}{\phi_i x+\lambda_i}-1\right)y_k\log(\phi_i x+\lambda_i)dP_X dP_{Y_k|X} \tag{99}$$

$$=\mathbb{E}[X_j\log(\phi_i X+\lambda_i)]+0. \tag{100}$$

We now establish the following technical Lemmas that will be used later. We note that the following Lemmas generalize the results in [10].

Lemma 2.

$$\mathbb{E}\left[\frac{X_j}{\phi_i X+\lambda_i}\middle| Y=y\right]=\frac{1}{y_i}\frac{P_Y(y_i-1,y_i^c)}{P_Y(y)}. \tag{101}$$

Proof of Lemma 2. First observe that by the Poisson channel assumption, we have

$$\frac{1}{\phi_i x+\lambda_i}=\frac{1}{y_i}\frac{P_{Y_i|X}(y_i-1|x)}{P_{Y_i|X}(y_i|x)} \tag{102}$$

$$\mathbb{E}\left[\frac{X_j}{\phi_i X+\lambda_i}\middle| Y=y\right]=\mathbb{E}\left[\frac{1}{y_i}\frac{P_{Y_i|X}(y_i-1|X)}{P_{Y_i|X}(y_i|X)}X_j\middle| Y=y\right] \tag{103}$$

$$=\frac{1}{y_i}\int\frac{P_{Y_i|X}(y_i-1|x)}{P_{Y_i|X}(y_i|x)}x_j dP_{X|Y=y} \tag{104}$$

$$=\frac{1}{y_i}\int\frac{P_{Y_i|X}(y_i-1|x)}{P_{Y_i|X}(y_i|x)}x_j\frac{P_{Y|X}(y|x)}{P_Y(y)}dP_X \tag{105}$$

$$=\frac{1}{y_i}\frac{P_Y(y_i-1,y_i^c)}{P_Y(y)}\int\frac{P_{Y_i|X}(y_i-1|x)}{P_{Y_i|X}(y_i|x)P_Y(y_i-1,y_i^c)}\prod_k P_{Y_k|X}(y_k|x)x_j dP_X \tag{106}$$

$$=\frac{1}{y_i}\frac{P_Y(y_i-1,y_i^c)}{P_Y(y)}\int_x\frac{P_{Y_i|X}(y_i-1|x)}{P_Y(y_i-1,y_i^c)}\prod_{k\neq i}P_{Y_k|X}(y_k|x)x_j dP_X \tag{107}$$

$$=\frac{1}{y_i}\frac{P_Y(y_i-1,y_i^c)}{P_Y(y)}\mathbb{E}[X_j|Y=(y_i-1,y_i^c)]. \tag{108}$$

□

Lemma 3.

$$\mathbb{E}(\phi_i X+\lambda_i|Y=y)=(y_i+1)\frac{P_Y(y_i+1,y_i^c)}{P_Y(y)}. \tag{109}$$

Proof of Lemma 3. First observe that

$$\phi_i x + \lambda_i = (y_i + 1)\frac{P_{Y_i|X}(y_i + 1|x)}{P_{Y_i|X}(y_i|x)}. \tag{110}$$

We have

$$\mathbb{E}(\phi_i X + \lambda_i | Y = y) = (y_i + 1)\mathbb{E}\left[\frac{P_{Y_i|X}(y_i + 1|X)}{P_{Y_i|X}(y_i|X)}\middle| Y = y\right] \tag{111}$$

$$= (y_i + 1)\int \frac{P_{Y_i|X}(y_i + 1|x)}{P_{Y_i|X}(y_i|x)} dP_{X|Y=y} \tag{112}$$

$$= \frac{y_i + 1}{P_Y(y)}\int_x \frac{P_{Y_i|X}(y_i + 1|x)}{P_{Y_i|X}(y_i|x)} P_{Y|X}(y|x) dP_X \tag{113}$$

$$= \frac{y_i + 1}{P_Y(y)}\int_x P_{Y_i|X}(y_i + 1|x)\prod_{k\neq i} P_{Y_k|X}(y_k|x) dP_X \tag{114}$$

$$= (y_i + 1)\frac{P_Y(y_i + 1, y_i^c)}{P_Y(y)}. \tag{115}$$

□

Lemma 4.

$$\mathbb{E}\left[\frac{1}{\phi_i X + \lambda_i}\middle| Y = y\right] = \frac{1}{y_i}\frac{P_Y(y_i - 1, y_i^c)}{P_Y(y)}. \tag{116}$$

Proof of Lemma 4. From the same observation in the proof of Lemma 2, we have

$$\mathbb{E}\left[\frac{1}{\phi_i X + \lambda_i}\middle| Y = y\right] = \mathbb{E}\left[\frac{1}{y_i}\frac{P_{Y_i|X}(y_i - 1|X)}{P_{Y_i|X}(y_i|X)}\middle| Y = y\right] \tag{117}$$

$$= \frac{1}{y_i}\int \frac{P_{Y_i|X}(y_i - 1|x)}{P_{Y_i|X}(y_i|x)} dP_{X|Y=y} \tag{118}$$

$$= \frac{1}{y_i}\int \frac{P_{Y_i|X}(y_i - 1|x)}{P_{Y_i|X}(y_i|x)}\frac{P_{Y|X}(y|x)}{P_Y(y)} dP_X \tag{119}$$

$$= \frac{1}{y_i}\frac{1}{P_Y(y)}\int \frac{P_{Y_i|X}(y_i - 1|x)}{P_{Y_i|X}(y_i|x)}\prod_k P_{Y_k|X}(y_k|x) dP_X \tag{120}$$

$$= \frac{1}{y_i}\frac{1}{P_Y(y)}\int P_{Y_i|X}(y_i - 1|x)\prod_{k\neq i} P_{Y_k|X}(y_k|x) dP_X \tag{121}$$

$$= \frac{1}{y_i}\frac{P_Y(y_i - 1, y_i^c)}{P_Y(y)}. \tag{122}$$

□

Combing previous derivations, we get

$$\frac{\partial I(X;Y)}{\partial \Phi_{ij}} = \mathbb{E}\left(X_j \log(\phi_i X + \lambda_i)\right) - \mathbb{E}\left\{X_j \left(\frac{Y_i}{\phi_i X + \lambda_i} - 1\right) \log\left((\prod_k Y_k!) P_Y(Y)\right)\right\}$$
$$= \mathbb{E}\left(X_j \log(\phi_i X + \lambda_i)\right) - \mathbb{E}\left\{\left(\mathbb{E}\left(\frac{X_j}{\phi_i X + \lambda_i}|Y\right) Y_i - \mathbb{E}(X_j|Y)\right) \log\left((\prod_k Y_k!) P_Y(Y)\right)\right\} \tag{123}$$

$$= \mathbb{E}\left(X_j \log(\phi_i X + \lambda_i)\right)$$
$$- \mathbb{E}\left\{\frac{P_Y(Y_i - 1, Y_i^c)}{P_Y(Y)} \mathbb{E}\left[X_j | Y = (Y_i - 1, Y_i^c)\right] \log\left((\prod_k Y_k!) P_Y(Y)\right)\right\}$$
$$+ \mathbb{E}\left\{\left(\mathbb{E}(X_j|Y)\right) \log\left((\prod_k Y_k!) P_Y(Y)\right)\right\} \tag{124}$$

$$= \mathbb{E}\left(X_j \log(\phi_i X + \lambda_i)\right)$$
$$- \int \left\{\frac{P_Y(y_i - 1, y_i^c)}{P_Y(y)} \mathbb{E}\left[X_j | Y = (y_i - 1, y_i^c)\right] \log\left((\prod_k y_k!) P_Y(y)\right) \frac{dP_Y}{dQ_Y} dQ_Y\right\}$$
$$+ \mathbb{E}\left\{\left(\mathbb{E}(X_j|Y)\right) \log\left((\prod_k Y_k!) P_Y(Y)\right)\right\} \tag{125}$$

$$= \mathbb{E}\left(X_j \log(\phi_i X + \lambda_i)\right)$$
$$- \int \left\{P_Y(y) \mathbb{E}\left[X_j | Y = y\right] \log\left(((y_i + 1)! \prod_{k \neq i} y_k!) P_Y(y_i + 1, y_i^c)\right) dQ_Y\right\}$$
$$+ \mathbb{E}\left\{\left(\mathbb{E}(X_j|Y)\right) \log\left((\prod_k Y_k!) P_Y(Y)\right)\right\} \tag{126}$$

$$= \mathbb{E}\left(X_j \log(\phi_i X + \lambda_i)\right)$$
$$- \int \left\{\mathbb{E}\left[X_j | Y = y\right] \log\left(((y_i + 1)! \prod_{k \neq i} y_k!) P_Y(y_i + 1, y_i^c)\right) \frac{dP_Y}{dQ_Y} dQ_Y\right\}$$
$$+ \mathbb{E}\left\{\left(\mathbb{E}(X_j|Y)\right) \log\left((\prod_k Y_k!) p_Y(y)\right)\right\} \tag{127}$$

$$= \mathbb{E}\left(X_j \log(\phi_i X + \lambda_i)\right)$$

$$
\begin{aligned}
&- \mathbb{E}\left\{\mathbb{E}\left[X_j|Y\right]\log\left(((Y_i+1)!\prod_{k\neq i}Y_k!)P_Y(Y_i+1,Y_i^c)\right)\right\}\\
&+ \mathbb{E}\left\{(\mathbb{E}(X_j|Y))\log\left((\prod_k Y_k!)P_Y(Y)\right)\right\} &(128)\\
&= \mathbb{E}\left(X_j\log(\phi_i X+\lambda_i)\right) - \mathbb{E}\left\{\mathbb{E}(X_j|Y)\log(Y_i+1)\frac{P_Y(Y_i+1,Y_i^c)}{P_Y(Y)}\right\} &(129)\\
&= \mathbb{E}\left(X_j\log(\phi_i X+\lambda_i)\right) - \mathbb{E}[\mathbb{E}[X_j|Y]\log(\mathbb{E}[\phi_i X+\lambda_i|Y])] &(130)\\
&= \mathbb{E}\left(X_j\log(\phi_i X+\lambda_i)\right) - \mathbb{E}[\mathbb{E}[X_j]\log(\mathbb{E}[\phi_i X+\lambda_i|Y])] &(131)\\
&= \mathbb{E}\left(X_j\log\frac{\phi_i X+\lambda_i}{\mathbb{E}[\phi_i X+\lambda_i|Y]}\right), &(132)
\end{aligned}
$$

where (124) follows from Lemma 2. (126) is obtained by a change of variable on y_i, together with the fact that $\frac{dP_Y}{dQ_Y} = P_Y$ for the counting measure Q_Y. (130) follows from Lemma 3 and (131) follows from Fubini's Theorem [28].

Hence, we have

$$
(\nabla_\Phi I(X;Y))_{ij} = \mathbb{E}\left[X_j\log((\Phi X)_i+\lambda_i)\right] - \mathbb{E}\left[\mathbb{E}[X_j|Y]\log\mathbb{E}[(\Phi X)_i+\lambda_i|Y]\right]. \tag{133}
$$

Now we present the proof for the gradient of mutual information with respect to the dark current.

$$
\begin{aligned}
\frac{\partial I(X;Y)}{\partial\lambda_i} &= \mathbb{E}\left(\frac{\partial}{\partial\lambda_i}\log f_{Y|X}^{\lambda_i}(y|x)\log\frac{f_{Y|X}^{\lambda_i}}{f_Y^{\lambda_i}}\right) &(134)\\
&= \mathbb{E}\left(\frac{\frac{\partial}{\partial\lambda_i}f_{Y|X}^{\lambda_i}(y|x)}{f_{Y|X}^{\lambda_i}(y|x)}\log\frac{f_{Y|X}^{\lambda_i}}{f_Y^{\lambda_i}}\right). &(135)
\end{aligned}
$$

Given the Poisson channel assumption, we can get that

$$
\begin{aligned}
\frac{\partial}{\partial\lambda_i}f_{Y|X}^{\lambda_i}(y|x) &= \frac{\partial}{\partial\lambda_i}\text{Pois}(y;\Phi x+\lambda) &(136)\\
&= \left(\frac{1}{y_i!}y_i(\phi_i x+\lambda_i)^{y_i-1}e^{-(\phi_i x+\lambda_i)} + \frac{1}{y_i!}(\phi_i x+\lambda_i)^{y_i}(-e^{-(\phi_i x+\lambda_i)})\right)\\
&\times\prod_{k\neq i}\frac{1}{y_k!}(\phi_k x+\lambda_k)^{y_k}e^{-(\phi_k x+\lambda_k)} &(137)\\
&= \frac{1}{y_i!}(\phi_i x+\lambda_i)^{y_i}e^{-(\phi_i x+\lambda_i)}\left(\frac{y_i}{\phi_i x+\lambda_i}-1\right)\prod_{k\neq i}\frac{1}{y_k!}(\phi_k x+\lambda_k)^{y_k}e^{-(\phi_k x+\lambda_k)} &(138)\\
&= \left(\frac{y_i}{\phi_i x+\lambda_i}-1\right)P_{Y|X}^{\lambda_i}(y|x). &(139)
\end{aligned}
$$

Followed by similar steps from (85) to (100), we obtain

$$\frac{\partial I(X;Y)}{\partial \lambda_i} = \mathbb{E}\left(\log(\phi_i X + \lambda_i)\right) - \mathbb{E}\left\{\left(\frac{Y_i}{\phi_i X + \lambda_i} - 1\right)\log\left((\prod_k Y_k!)P_Y(Y)\right)\right\}$$

$$= \mathbb{E}\left(\log(\phi_i X + \lambda_i)\right) - \mathbb{E}\left\{\left(\mathbb{E}\left(\frac{1}{\phi_i X + \lambda_i}|Y\right)Y_i - 1\right)\log\left((\prod_k Y_k!)P_Y(Y)\right)\right\} \quad (140)$$

$$= \mathbb{E}\left(\log(\phi_i X + \lambda_i)\right)$$
$$- \mathbb{E}\left\{\frac{P_Y(Y_i - 1, Y_i^c)}{P_Y(Y)}\log\left((\prod_k Y_k!)P_Y(Y)\right)\right\} + \mathbb{E}\left\{\log\left((\prod_k Y_k!)P_Y(Y)\right)\right\} \quad (141)$$

$$= \mathbb{E}\left(\log(\phi_i X + \lambda_i)\right)$$
$$- \int\left\{\log\left(((y_i+1)!\prod_{k\neq i} y_k!)P_Y(y_i+1, y_i^c)\right)dP_Y\right\} + \mathbb{E}\left\{\log\left((\prod_k Y_k!)P_Y(Y)\right)\right\} \quad (142)$$

$$= \mathbb{E}\left(\log(\phi_i X + \lambda_i)\right)$$
$$- \mathbb{E}\left\{\log\left(((Y_i+1)!\prod_{k\neq i} Y_k!)P_Y(Y_i+1, Y_i^c)\right)\right\} + \mathbb{E}\left\{\log\left((\prod_k Y_k!)P_Y(Y)\right)\right\} \quad (143)$$

$$= \mathbb{E}\left(\log(\phi_i X + \lambda_i)\right) - \mathbb{E}\left\{\log(Y_i+1)\frac{P_Y(Y_i+1, Y_i^c)}{p_Y(Y)}\right\} \quad (144)$$

$$= \mathbb{E}\left(\log(\phi_i X + \lambda_i)\right) - \mathbb{E}\left[\log(\mathbb{E}[\phi_i X + \lambda_i|Y])\right], \quad (145)$$

where (141) and (145) follow from Lemma 4 and Lemma 3. (142) is obtained by a change of variable on y_i, together with the fact that $\frac{dP_Y}{dQ_Y} = P_Y$ for the counting measure Q_Y. Hence, we have

$$(\nabla_\lambda I(X;Y))_i = \mathbb{E}\left[\log((\Phi X)_i + \lambda_i)\right] - \mathbb{E}\left[\log \mathbb{E}[(\Phi X)_i + \lambda_i|Y]\right]. \quad (146)$$

□

APPENDIX C
PROOF OF THEOREM 2

Proof. First we notice that

$$I(C;Y) = H(Y) - H(Y|C) \quad (147)$$

$$= H(Y) - H(Y|X) + H(Y|X,C) - H(Y|C) \tag{148}$$

$$= I(X;Y) - I(X;Y|C), \tag{149}$$

where the second equality is due to the fact that $C \to X \to Y$ forms a Markov chain and $P_{Y|X,C} = P_{Y|X}$. Following similar steps as in the proof of Theorem 1, we have

$$[\nabla_\Phi I(X;Y|C)]_{ij} = \left[\mathbb{E}\left[X_j \log((\Phi X)_i + \lambda_i)\right] - \mathbb{E}\left[\mathbb{E}[X_j|Y,C]\log\mathbb{E}[(\Phi X)_i + \lambda_i|Y,C]\right]\right].$$

Hence,

$$[\nabla_\Phi I(C;Y)]_{ij} = -\mathbb{E}\left[\mathbb{E}[X_j|Y]\log\mathbb{E}[(\Phi X)_i + \lambda_i|Y]\right] + \mathbb{E}\left[\mathbb{E}[X_j|Y,C]\log\mathbb{E}[(\Phi X)_i + \lambda_i|Y,C]\right] \tag{150}$$

$$= -\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}[X_j|Y,C]\middle|Y\right]\log\mathbb{E}[(\Phi X)_i + \lambda_i|Y]\right] + \mathbb{E}\left[\mathbb{E}[X_j|Y,C]\log\mathbb{E}[(\Phi X)_i + \lambda_i|Y,C]\right] \tag{151}$$

$$= -\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}[X_j|Y,C]\right]\log\mathbb{E}[(\Phi X)_i + \lambda_i|Y]\middle|Y\right] + \mathbb{E}\left[\mathbb{E}[X_j|Y,C]\log\mathbb{E}[(\Phi X)_i + \lambda_i|Y,C]\right] \tag{152}$$

$$= -\mathbb{E}\left[\mathbb{E}[X_j|Y,C]\log\mathbb{E}[(\Phi X)_i + \lambda_i|Y]\right] + \mathbb{E}\left[\mathbb{E}[X_j|Y,C]\log\mathbb{E}[(\Phi X)_i + \lambda_i|Y,C]\right] \tag{153}$$

$$= \mathbb{E}\left[\mathbb{E}[X_j|Y,C]\log\frac{\mathbb{E}[(\Phi X)_i + \lambda_i|Y,C]}{\mathbb{E}[(\Phi X)_i + \lambda_i|Y]}\right] \tag{154}$$

Similarly, we have

$$(\nabla_\lambda I(X;Y|C))_i = \mathbb{E}\left[\log((\Phi X)_i + \lambda_i)\right] - \mathbb{E}\left[\log\mathbb{E}[(\Phi X)_i + \lambda_i|Y,C]\right]. \tag{155}$$

Therefore the gradient with respect to the dark current can be represented as

$$(\nabla_\lambda I(C;Y))_i = \mathbb{E}\left[\log\frac{\mathbb{E}[(\Phi X)_i + \lambda_i|Y,C]}{\mathbb{E}[(\Phi X)_i + \lambda_i|Y]}\right]. \tag{156}$$

□

APPENDIX D
PROOFS OF THEOREM 3, THEOREM 4 AND 5

Proof of Theorem 3. We first show the non-negativity. Since F is strictly K-convex and Fréchet differentiable, by the first order derivative characterization of K-convexity for Banach space [33], we

have $F(x) \succeq_K F(y) + DF(y)(x-y)$. Hence,

$$D_F(x,y) \succeq_K \mathbf{0}. \tag{157}$$

Now we show the linearity. Let $c_1 > 0$ and $c_2 > 0$ be two arbitrary positive constants. We have

$$D_{c_1F+c_2G}(x,y) = c_1F(x) + c_2G(x) - c_1F(y) - c_2G(y) - D(c_1F + c_2G)(y)(x-y) \tag{158}$$

$$= c_1(F(x) - F(y)) - c_1DF(y)(x-y) + c_2(G(x) - G(y)) - c_2DG(y)(x-y) \tag{159}$$

$$= c_1D_F(x,y) + c_2D_G(x,y). \tag{160}$$

Last, we show the K-convexity. For $0 \leq \theta \leq 1$ and $x, y, z \in \Omega$, we have

$$D_F(\theta x + (1-\theta)z, y) = F(\theta x + (1-\theta)z) - F(y) - DF(y)(\theta x + (1-\theta)z - y) \tag{161}$$

$$\preceq_K \theta F(x) + (1-\theta)F(z) - \theta F(y) - (1-\theta)F(y)$$

$$- \theta DF(y)(x) - (1-\theta)DF(y)(z) + \theta DF(y)(y) + (1-\theta)DF(y)(y) \tag{162}$$

$$= \theta(F(x) - F(y) - DF(y)(x-y)) + (1-\theta)(F(z) - F(y) - DF(y)(z-y)) \tag{163}$$

$$= \theta D_F(x,y) + (1-\theta)D_F(z,y). \tag{164}$$

□

Proof of Theorem 4. By the assumption that K is the space of the first quadrant, we have that the $\preceq_K$ means the entry-wise convexity. Recall from [51] that the Legendre transform $(F^\star, y^\star)$ on a convex set Ω for the pair (F, y) is such that as

$$F(y) = -F^\star(y^\star) + [y^T y^\star] \tag{165}$$

$$DF(y)(x) = [(y^\star)^T x], \tag{166}$$

where $[a]$ denotes the $m \times n$ matrix with all identical entries a. The dual point of y is the vector $y^\star$ such that the following equality holds for all vector $x \in \mathbb{R}^k$

$$DF^\star(y^\star)(x) = [y^T x]. \tag{167}$$

We also have the following properties

$$F^\star(y^\star) = -F(y) + [(y^\star)^T y] \tag{168}$$

$$DF^\star(y^\star)(x) = [y^T x]. \tag{169}$$

Plugging the above equations in $D_F(x, y)$, we have

$$D_F(x,y) = F(x) + F^\star(y^\star) - [y^T y^\star] - [(y^\star)^T(x-y)] \tag{170}$$

$$= F(x) + F^\star(y^\star) - [(y^\star)^T x] \tag{171}$$

$$= F(x) + F^\star(y^\star) - [x^T x^\star] - [x^T(y^\star - x^\star)] \tag{172}$$

$$= D_{F^\star}(y^\star, x^\star). \tag{173}$$

□

Proof of Theorem 5. Let $Y' := \mathbb{E}[X|s_1]$. We have

$$\mathbb{E}_{X,Y}[D_F(X,Y)] - \mathbb{E}_{X,Y'}[D_F(X,Y')]$$

$$= \mathbb{E}_{X,Y,Y'}[F(Y') - F(Y) - D_F(Y)(X-Y) + D_F(Y')(X-Y')] \tag{174}$$

$$= \mathbb{E}_{Y,Y'}[F(Y') - F(Y)] - \mathbb{E}[\mathbb{E}[D_F(Y)(X-Y)]|s_1] + \mathbb{E}[\mathbb{E}[D_F(y')(X-Y')|s_1]] \tag{175}$$

$$= \mathbb{E}_{Y,Y'}[F(Y') - F(Y)] - \mathbb{E}[D_F(Y)(Y'-Y)] + \mathbb{E}[D_F(Y')(Y'-Y')] \tag{176}$$

$$= \mathbb{E}_{Y,Y'}[F(Y') - F(Y) - \mathbb{E}[D_F(Y)(Y'-Y)]] \tag{177}$$

$$= \mathbb{E}_{Y,Y'}[D_F(Y)(Y',Y)] \succeq_K \mathbf{0}. \tag{178}$$

The last inequality follows from the non-negativity property and (176) follows from linearity of $D_F(Y)$. On the other hand, in case that $\mathbb{E}_{Y,Y'}[D_F(Y)(Y',Y)] = \mathbf{0}$, we must have $Y' = Y$, which follows from the property that F is strictly K-convex if and only if $F(y) \succ_K F(x) + DF(x)(y-x)$ for $x \neq y$ [33].

□

APPENDIX E
PROOF OF THEOREM 6, 7, 8 AND 9

Proof of Theorem 6. We first show the Poisson case. Notice that $DF(\mathbb{E}[X|Y])(\cdot)$ is a linear operator. Thus,

$$\mathbb{E}[DF(\mathbb{E}[X|Y])(X-\mathbb{E}[X|Y])] = \mathbb{E}_Y\left[\mathbb{E}[DF(\mathbb{E}[X|Y])(X)|Y]\right] - \mathbb{E}_Y[DF(\mathbb{E}[X|Y])(\mathbb{E}[X|Y])] \quad (179)$$

$$= \mathbb{E}_Y\left[DF(\mathbb{E}[X|Y])(\mathbb{E}[X|Y])\right] - \mathbb{E}_Y[DF(\mathbb{E}[X|Y])(\mathbb{E}[X|Y])] \quad (180)$$

$$= \mathbf{0}. \quad (181)$$

Hence,

$$\mathbb{E}\left[D_F(X,\mathbb{E}[X|Y])\right]$$

$$= \mathbb{E}\left[X(\log(\Phi X+\lambda))^T - [X,\ldots,X] - \mathbb{E}[X|Y](\log(\Phi\mathbb{E}[X|Y])+\lambda)^T + [\mathbb{E}[X|Y],\ldots,\mathbb{E}[X|Y]]\right] \quad (182)$$

$$= \mathbb{E}\left[X(\log(\Phi X+\lambda))^T - \mathbb{E}[X|Y](\log(\Phi\mathbb{E}[X|Y]+\lambda))^T\right] \quad (183)$$

$$= \nabla_\Phi I(X;Y). \quad (184)$$

For the Gaussian case, the differential of the function $DF(x)$ can be represented under the standard basis [52] as

$$DF(x) = (I_n \otimes \Lambda\Phi x) + (x \otimes I_m)\Lambda\Phi \quad (185)$$

$$= \begin{pmatrix} \Lambda\Phi x & & \\ & \ddots & \\ & & \Lambda\Phi x \end{pmatrix} + \begin{pmatrix} x_1\Lambda\Phi \\ \vdots \\ x_n\Lambda\Phi \end{pmatrix} \quad (186)$$

$$= \begin{pmatrix} \Lambda\Phi & & \\ & \ddots & \\ & & \Lambda\Phi \end{pmatrix}\begin{pmatrix} x & & \\ & \ddots & \\ & & x \end{pmatrix} + \begin{pmatrix} \Lambda\Phi & & \\ & \ddots & \\ & & \Lambda\Phi \end{pmatrix}\begin{pmatrix} x_1 I_n \\ \vdots \\ x_n I_n \end{pmatrix} \quad (187)$$

$$= (I_n \otimes \Lambda\Phi)(I_n \otimes x) + (I_n \otimes \Lambda\Phi)(x \otimes I_n), \quad (188)$$

where I_n and I_m are the $n \times n$ and $m \times m$ identity matrices. Let $y \in \mathbb{R}^n$, we have

$$DF(x)(y-x) = (I_n \otimes \Lambda\Phi)(I_n \otimes x + x \otimes I_n)(y-x) \tag{189}$$

$$= (I_n \otimes \Lambda\Phi) \begin{pmatrix} 2x_1 & 0 & \dots & 0 \\ x_2 & x_1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ x_n & 0 & \vdots & x_1 \\ \vdots & \vdots & \vdots & \vdots \\ x_n & 0 & \dots & x_1 \\ 0 & \ddots & \vdots & x_2 \\ \vdots & \dots & x_n & \vdots \\ 0 & \dots & 0 & 2x_n \end{pmatrix} (y-x). \tag{190}$$

Note that $DF(x)(y-x)$ is a vector of size $mn \times 1$, which is obtained by vectorizing the matrix form $DF(x)(y-x)$. Since those two forms are just two different representations of the same differential, we abuse the notation $DF(x)(y-x)$ without discrimination. By re-vectorizing $DF(x)(y-x)$ to the matrix form, we have

$$DF(x)(y-x) = \Lambda\Phi \begin{pmatrix} 2x_1y_1 - 2x_1^2 & x_1y_2 + x_2y_1 - 2x_1x_2 & \cdots & x_1y_n + x_ny_1 - 2x_1x_n \\ x_2y_1 + x_1y_2 - 2x_2x_1 & 2x_2y_2 - 2x_2^2 & \cdots & x_2y_n + x_ny_2 - 2x_2x_n \\ \vdots & \vdots & \ddots & \vdots \\ x_ny_1 + x_1y_n - 2x_nx_1 & x_ny_2 + x_2y_n - 2x_nx_2 & \dots & 2x_ny_n - 2x_n^2 \end{pmatrix}. \tag{191}$$

Hence it is straightforward to verify that

$$\nabla_\Phi I(X;Y)$$

$$= \Lambda\Phi\mathbb{E}[(X - \mathbb{E}[X|Y])(X - \mathbb{E}[X|Y])^T] \tag{192}$$

$$= \Lambda\Phi\mathbb{E}[XX^T - X\mathbb{E}[X|Y]^T - \mathbb{E}[X|Y]X^T + \mathbb{E}[X|Y]\mathbb{E}[X|Y]^T] \tag{193}$$

$$= \Lambda\Phi\mathbb{E}\left[XX^T - \mathbb{E}[X|Y]\mathbb{E}[X|Y]^T - DF(\mathbb{E}[X|Y])(X - \mathbb{E}[X|Y])\right] \tag{194}$$

$$= \mathbb{E}[D_F(X, \mathbb{E}[X|Y])]. \tag{195}$$

The first equality follows from the result in [3].

Finally, we need to show $F(x)$ is strictly K-convex for both Poisson and Gaussian cases, which depends on specific choice of the cone K. Thus, here we only show that there exists a K such that $F(x)$ is strictly K-convex. This is straightforward to check, if we choose $K = \{M \in \mathbb{R}^{m\times n} | M_{11} > 0 \text{ and } M_{ij} = 0, \forall (i,j) \neq (1,1)\}$. □

Proof of Theorem 7. We first show the Poisson case. Notice that $DF(\mathbb{E}[X|Y])(\cdot)$ is a linear operator. Thus,

$$\begin{aligned}
&\mathbb{E}[DF(\mathbb{E}[X|Y])(\mathbb{E}[X|Y,C] - \mathbb{E}[X|Y])] \\
&= \mathbb{E}_Y\left[\mathbb{E}_{C|Y}[DF(\mathbb{E}[X|Y])(\mathbb{E}[X|Y,C])|Y]\right] - \mathbb{E}_Y[DF(\mathbb{E}[X|Y])(\mathbb{E}[X|Y])] && (196)\\
&= \mathbb{E}_Y\left[DF(\mathbb{E}[X|Y])(\mathbb{E}[X|Y])\right] - \mathbb{E}_Y[DF(\mathbb{E}[X|Y])(\mathbb{E}[X|Y])] && (197)\\
&= \mathbf{0}. && (198)
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}\left[D_F(\mathbb{E}[X|Y,C], \mathbb{E}[X|Y])\right] \\
&= \mathbb{E}[\mathbb{E}[X|Y,C](\log(\Phi\mathbb{E}[X|Y,C] + \lambda))^T - [\mathbb{E}[X|Y,C], \ldots, \mathbb{E}[X|Y,C]] \\
&- \mathbb{E}[X|Y](\log(\Phi\mathbb{E}[X|Y]) + \lambda)^T + [\mathbb{E}[X|Y], \ldots, \mathbb{E}[X|Y]]] && (199)\\
&= \mathbb{E}\left[\mathbb{E}[X|Y,C](\log(\Phi\mathbb{E}[X|Y,C] + \lambda))^T - \mathbb{E}[X|Y,C](\log(\Phi\mathbb{E}[X|Y] + \lambda))^T\right] && (200)\\
&= \nabla_\Phi I(X;Y) && (201)
\end{aligned}$$

For the Gaussian case, we notice that (191) is also vacuously valid for arbitrary x and y and in particularly, by the similar arguments in Gaussian case proof of Theorem 6, we can obtain

$$\begin{aligned}
&\nabla_\Phi I(C;Y) \\
&= \Lambda\Phi\mathbb{E}\left[(\mathbb{E}[X|Y] - \mathbb{E}[X|Y,C])(\mathbb{E}[X|Y] - \mathbb{E}[X|Y,C])^T\right] && (202)\\
&= \Lambda\Phi\mathbb{E}\left[\mathbb{E}[X|Y]\mathbb{E}[X|Y]^T - \mathbb{E}[X|Y]\mathbb{E}[X|Y,C]^T - \mathbb{E}[X|Y,C]\mathbb{E}[X|Y]^T + \mathbb{E}[X|Y,C]\mathbb{E}[X|Y,C]^T\right] && (203)\\
&= \Lambda\Phi\mathbb{E}\left[\mathbb{E}[X|Y,C]\mathbb{E}[X|Y,C]^T - \mathbb{E}[X|Y]\mathbb{E}[X|Y]^T - DF(\mathbb{E}[X|Y])(\mathbb{E}[X|Y,C] - \mathbb{E}[X|Y])\right] && (204)
\end{aligned}$$

$$= \mathbb{E}[D_F(\mathbb{E}[X|Y,C]), \mathbb{E}[X|Y])]. \tag{205}$$

The first equality follows from the result in [21].

Finally, we need to show $F(x)$ is strictly K-convex for both Poisson and Gaussian cases, which depends on specific choice of the cone K. Thus, here we only show that there exists a K such that $F(x)$ is strictly K-convex. This is straightforward to check, if we choose $K = \{M \in \mathbb{R}^{m\times n} | M_{11} > 0 \text{ and } M_{ij} = 0, \forall (i,j) \neq (1,1)\}$. □

Proof of Theorem 8. By using (191), we have that

$$\begin{aligned} &D_F(\mathbb{E}_P[X|\sqrt{\gamma}X+N], \mathbb{E}_Q[X|\sqrt{\gamma}X+N]) \\ &= \left(\mathbb{E}_P[X|\sqrt{\gamma}X+N] - \mathbb{E}_Q[X|\sqrt{\gamma}X+N]\right)\left(\mathbb{E}_P[X|\sqrt{\gamma}X+N] - \mathbb{E}_Q[X|\sqrt{\gamma}X+N]\right)^T. \end{aligned} \tag{206}$$

Hence,

$$\begin{aligned} &\frac{1}{2}\mathbb{E}_P\left[\int_0^\infty \mathrm{Tr}\left[D_F(\mathbb{E}_P[X|\sqrt{\gamma}X+N], \mathbb{E}_Q[X|\sqrt{\gamma}X+N])\right] \mathrm{d}\gamma\right] \\ &= \frac{1}{2}\mathbb{E}_P\left[\int_0^\infty \| \left(\mathbb{E}_P[X|\sqrt{\gamma}X+N] - \mathbb{E}_Q[X|\sqrt{\gamma}X+N]\right) \|_2^2 \mathrm{d}\gamma\right] &(207)\\ &= D(P\|Q). &(208)\end{aligned}$$

The last equality follows from Theorem 1 in [4]. □

Proof of Theorem 9. It is straightforward to see that the natural map $\mathcal{I} : M_\Phi \to M$ defined as $\mathcal{I} : \Phi \mapsto (\Phi, I_\Phi(X;Y))$ is a diffeomorphism. It is enough to show that the length of any C^1 curve $c(t) : [0,1] \to M_\Phi$ connecting two points Φ_1 and Φ_2 under the metric $\mathfrak{g}$ is the same as the length of $\mathcal{I}(c(t))$ connecting $\mathcal{I}(\Phi_1)$ and $\mathcal{I}(\Phi_2)$ under the metric $\mathfrak{I}$. The length of $c(t)$ in $(M_\Phi, \mathfrak{g})$ can be calculated as

$$\begin{aligned} \|c(t)\|_{\mathfrak{g}} &= \int_0^1 \sqrt{\mathfrak{g}(c'(t), c'(t))} dt &(209)\\ &= \int_0^1 \sqrt{(c'(t))^T \mathfrak{g} c'(t)} dt &(210)\\ &= \int_0^1 \sqrt{\sum_{i=1}^{mn} [\mathbb{E}[D_F(X, \mathbb{E}[X|Y])]]_i^2 (c_i'(t))^2} dt &(211)\end{aligned}$$

$$= \int_0^1 \sqrt{\sum_{i=1}^{mn} [\mathbb{E}[D_F(X, \mathbb{E}[X|Y])]_i c_i'(t)]^2} dt \tag{212}$$

$$= \int_0^1 \sqrt{\sum_{i=1}^{mn} ((\nabla_{c(t)} I(X;Y))_i)^2} dt \tag{213}$$

$$= \int_0^1 \sqrt{((\mathcal{I}(c(t)))')^T \mathfrak{I} (\mathcal{I}(c(t)))'} dt \tag{214}$$

$$= \int_0^1 \sqrt{\mathfrak{I}(\mathcal{I}(c(t))', \mathcal{I}(c(t))')} dt \tag{215}$$

$$= \|\mathcal{I}(c(t))\|_{\mathfrak{I}}, \tag{216}$$

where (213) follows from Theorem 6 and chain rule. □

REFERENCES

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

[2] D. Guo, S. Shamai, and S. Verdú, "Mutual information and conditional mean estimation in Poisson channels," *IEEE Transactions on Information Theory*, vol. 54, no. 5, pp. 1837–1849, May 2008.

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

[4] S. Verdú, "Mismatched estimation and relative entropy," *IEEE Transactions on Information Theory*, vol. 56, no. 8, pp. 3712–3720, Aug. 2010.

[5] T. Weissman, "The relationship between causal and noncausal mismatched estimation in continuous-time AWGN channels," *IEEE Transactions on Information Theory*, vol. 56, no. 9, pp. 4256–4273, Sept. 2010.

[6] C.G. Taborda and F. Perez-Cruz, "Mutual information and relative entropy over the binomial and negative binomial channels," in *IEEE International Symposium on Information Theory Proceedings (ISIT)*. IEEE, 2012, pp. 696–700.

[7] D. Guo, "Information and estimation over binomial and negative binomial models," *arXiv preprint arXiv:1207.7144*, 2012.

[8] Y.M. Kabanov, "The capacity of a channel of the Poisson type," *Theory of Probability & Its Applications*, vol. 23, no. 1, pp. 143–147, 1978.

[9] R.S. Liptser and A.N. Shiryaev, *Statistics of Random Processes: II. Applications*, vol. 2, Springer, 2000.

[10] D.P. Palomar and S. Verdú, "Representation of mutual information via input estimates," *IEEE Transactions on Information Theory*, vol. 53, no. 2, pp. 453–470, Feb. 2007.

[11] S. Verdú, "Poisson communication theory," *Invited talk in the International Technion Communication Day in honor of Israel Bar-David*, May 1999.

[12] I. Bar-David, "Communication under the poisson regime," *IEEE Transactions on Information Theory*, vol. 15, no. 1, pp. 31–37, Jan. 1969.

[13] D. Snyder, "Filtering and detection for doubly stochastic poisson processes," *IEEE Transactions on Information Theory*, vol. 18, no. 1, pp. 91–102, Jan. 1972.

[14] M. Davis, "Capacity and cutoff rate for poisson-type channels," *IEEE Transactions on Information Theory*, vol. 26, no. 6, pp. 710–715, June 1980.

[15] A.D. Wyner, "Capacity and error exponent for the direct detection photon channel. ii," *IEEE Transactions on Information Theory*, vol. 34, no. 6, pp. 1462–1471, June 1988.

[16] S. Shamai and A. Lapidoth, "Bounds on the capacity of a spectrally constrained poisson channel," *IEEE Transactions on Information Theory*, vol. 39, no. 1, pp. 19–29, Jan. 1993.

[17] A. Lapidoth and S. Shamai, "The poisson multiple-access channel," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 488–501, Feb. 1998.

[18] S.M. Haas and J.H. Shapiro, "Capacity of wireless optical communications," *IEEE Journal on Selected Areas in Communications*, vol. 21, no. 8, pp. 1346–1357, Aug. 2003.

[19] I.A. Elbakri and J.A. Fessler, "Statistical image reconstruction for polyenergetic X-ray computed tomography," *IEEE Transactions on Medical Imaging*, vol. 21, no. 2, pp. 89–99, Feb. 2002.

[20] M. Zhou, L. Hannah, D. Dunson, and L. Carin, "Beta-negative binomial process and Poisson factor analysis," *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[21] M. Chen, W. Carson, M. Rodrigues, R. Calderbank, and L. Carin, "Communications inspired linear discriminant analysis," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2012.

[22] R. Atar and T. Weissman, "Mutual information, relative entropy, and estimation in the Poisson channel," *IEEE Transactions on Information Theory*, vol. 58, no. 3, pp. 1302–1318, March 2012.

[23] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2010.

[24] Z.T. Harmany, R.F. Marcia, and R.M. Willett, "This is SPIRAL-TAP: sparse Poisson intensity reconstruction algorithms-theory and practice," *IEEE Transactions on Image Processing*, vol. 21, no. 3, pp. 1084–1096, March 2012.

[25] M. Hellman and J. Raviv, "Probability of error, equivocation, and the Chernoff bound," *IEEE Transactions on Information Theory*, vol. 16, no. 4, pp. 368–372, April 1970.

[26] R.M. Fano, *Transmission of Information: A Statistical Theory of Communication*, Wiley, New York, 1961.

[27] Z. Nenadic, "Information discriminant analysis: Feature extraction with an information-theoretic objective," *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, vol. 29, no. 8, pp. 1394–1407, Aug. 2007.

[28] G.B. Folland, *Real Analysis: Modern Techniques and Their Applications*, Wiley New York, 1999.

[29] L.M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR computational mathematics and mathematical physics*, vol. 7, no. 3, pp. 200–217, March 1967.

[30] A. Banerjee, S. Merugu, I. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[31] B.A. Frigyik, S. Srivastava, and M.R. Gupta, "Functional Bregman divergence and Bayesian estimation of distributions," *IEEE Transactions on Information Theory*, vol. 54, no. 11, pp. 5130–5139, Nov. 2008.

[32] R. Iyer and J. Bilmes, "Submodular-Bregman and the Lovász-Bregman divergences with applications," in *Advances in Neural Information Processing Systems*, 2012, pp. 2942–2950.

[33] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, 2004.

[34] B. Kulis, M.A. Sustik, and I.S. Dhillon, "Low-rank kernel learning with Bregman matrix divergences," *Journal of Machine Learning Research*, vol. 10, pp. 341–376, 2009.

[35] R. Bishop, *Tensor Analysis on Manifolds*, Dover Publications, 1968.

[36] S. Lang, *Differential and Riemannian manifolds*, vol. 160, Springer Verlag, 1995.

[37] S-I. Amari and A. Cichocki, "Information geometry of divergence functions," *Bulletin of the Polish Academy of Sciences: Technical Sciences*, vol. 58, no. 1, pp. 183–195, 2010.

[38] D. Guo, "'relative entropy and score function: New information-estimation relationships through arbitrary additive perturbation," in *IEEE Int. Symp. Information Theory*, 2009.

[39] A. Ben-Tal, T. Margalit, and A. Nemirovski, "The ordered subsets mirror descent optimization method with applications to tomography," *SIAM Journal on Optimization*, vol. 12, no. 1, pp. 79–108, Jan. 2001.

[40] A. Agarwal, P.L. Bartlett, P. Ravikumar, and M.J. Wainwright, "Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization," *IEEE Transactions on Information Theory*, vol. 58, no. 5, pp. 3235–3249, May 2012.

[41] A.S. Nemirovsky and D.B. Yudin, *Problem Complexity and Method Efficiency in Optimization.*, Wiley, 1983.

[42] A. Banerjee, X. Guo, and H. Wang, "On the optimality of conditional expectation as a Bregman predictor," *IEEE Transactions on Information Theory*, vol. 51, no. 7, pp. 2664 –2669, July 2005.

[43] F.W. Warner, *Foundations of Differentiable Manifolds and Lie Groups*, vol. 94, Springer, 1971.

[44] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer New York, 2006.

[45] M. Zhou, L. Hannah, D. Dunson, and L. Carin, "Beta-negative binomial process and Poisson factor analysis," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

[46] L. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in *Proceedings of Association of Computational Linguistics*, 2004.

[47] M.W. Berry, M. Browne, A.N. Langville, V.P. Pauca, and R.J. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Computational Statistics & Data Analysis*, vol. 52, no. 1, pp. 155–173, 2007.

[48] N. Gillis, "Sparse and unique nonnegative matrix factorization through data preprocessing," *Journal of Machine Learning Research*, vol. 13, no. 1, pp. 3349–3386, 2012.

[49] D. M. Blei, A. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[50] Y. W. Teh, M. I. Jordan, Matthew J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[51] I.M. Gelfand and S.V. Fomin, *Calculus of Variations*, Dover publications, 2000.

[52] H. Neudecker and J. Magnus, *Matrix Differential Calculus with Applications in Statistics and Econometrics*, John Wiley, 1999.