ISIT'2013 1569719699

Generalized Bregman Divergence and Gradient of Mutual Information for Vector Poisson Channels

Liming Wang[†], Miguel Rodrigues[⋆], Lawrence Carin[†]
[†]Dept. of Electrical & Computer Engineering, Duke University, Durham, NC 27708, USA
Email: {liming.w, lcarin}@duke.edu
[⋆]Dept. of Electronic & Electrical Engineering, University College London, London, U.K.
Email: m.rodrigues@ucl.ac.uk

***Abstract*—We investigate connections between information-theoretic and estimation-theoretic quantities in vector Poisson channel models. In particular, we generalize the gradient of mutual information with respect to key system parameters from the scalar to the vector Poisson channel model. We also propose, as another contribution, a generalization of the classical Bregman divergence that offers a means to encapsulate under a unifying framework the gradient of mutual information results for scalar and vector Poisson and Gaussian channel models. The so-called generalized Bregman divergence is also shown to exhibit various properties akin to the properties of the classical version. The vector Poisson channel model is drawing considerable attention in view of its application in various domains: as an example, the availability of the gradient of mutual information can be used in conjunction with gradient descent methods to effect compressive-sensing projection designs in emerging X-ray and document classification applications.**

I. Introduction

There has been a recent emergence of intimate connections between various quantities in information theory and estimation theory. The perhaps most prominent connections reveal the interplay between two notions with operational relevance in each of the domains: *mutual information* and *conditional mean estimation*.

In particular, Guo, Shamai and Verdú [1] have expressed the derivative of mutual information in a scalar Gaussian channel via the (non-linear) *minimum mean-squared error* (MMSE), and Palomar and Verdú [2] have expressed the gradient of mutual information in a vector Gaussian channel in terms of the MMSE matrix. The connections have also been extended from the scalar Gaussian to the scalar Poisson channel model, which has been ubiquitously used to model optical communications [3], [4]. Recently, parallel results for scalar binomial and negative binomial channels have been established [5], [6]. Inspired by the Lipster-Shiryaev formula [7], it has been demonstrated that it is often easier to investigate the gradient of mutual information rather than mutual information itself [3]. Further, it has also been shown that the derivative of mutual information with respect to key system parameters also relates to the conditional mean estimator [3].

This paper also pursues this overarching theme. One of the goals is to generalize the gradient of mutual information from scalar to vector Poisson channel models. This generalization is relevant not only from the theoretical but also from the practical perspective, in view of the numerous emerging applications of the vector Poisson channel model in X-ray systems [8] and document classification systems (based on word counts) [9]. The availability of the gradient then provides the means to optimize the mutual information with respect to specific system parameters via gradient descent methods.

The other goal is to encapsulate under a unified framework the gradient of mutual information results for scalar Gaussian channels, scalar Poisson channels and their vector counterparts.

This encapsulation, which 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 loss function between the input and the conditional mean estimate of the input [10], is possible by constructing a generalization of the classical Bregman divergence from the scalar to the vector case. This generalization of Bregman divergence appears to be new to the best of our knowledge. The gradients of mutual information of the vector Poisson model and the vector Gaussian model, as well as the scalar counterparts, are then also expressed - and akin to [10] - in terms of the average value of the so called generalized Bregman divergence associated with particular (vector) loss function between the input vector and the conditional mean estimate of the input vector.

We also study in detail various properties of the generalized Bregman divergence: the properties of the proposed divergence are shown to mimic closely those of the classical Bregman divergence.

The generalized Bregman divergence framework is of interest not only from the theoretical but also the practical standpoint: for example, it has been shown that re-expressing results via a Bregman divergence can often lead to enhancements to the speed of various optimization algorithms [11].

This paper is organized as follows: Section II introduces the channel model. Section III derives the gradient of mutual information with respect to key system parameters for vector Poisson channel models. Section IV introduces the notion of a generalized Bregman divergence and its properties. Section V re-derives the gradient of mutual information of vector Poisson and Gaussian channel models under the light of the proposed Bregman divergence. A possible application of the theoretical results in an emerging domain is succintly described in Section VI. Section VII concludes the paper.

II. The Vector Poisson Channel

We define the vector Poisson channel model via the random transformation:

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

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. $\text{Pois}\left(z\right)$ denotes a standard Poisson distribution with parameter z.

This vector Poisson channel model associated with arbitrary m and n is a generalization of the standard scalar Poisson model associated with $m=n=1$ given by [3], [10]:

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

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 associated with the dark current.[1]

The generalization of the scalar Poisson model in (2) to the vector one in (1) offers the means to address relevant problems in various emerging applications, most notably in X-ray and document classification applications as discussed in the sequel [9], [12].

The goal is to define the gradient of mutual information between the input and the output of the vector Poisson channel with respect to the scaling matrix, i.e.

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

where $\nabla_\Phi I(X;Y)_{ij}$ represents the (i,j)-th entry of the matrix $\nabla_\Phi I(X;Y)$, and with respect to the dark current, i.e.

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

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

We will also be concerned with drawing connections between the gradient result for the vector Poisson channel and the gradient result for the Gaussian counterpart in the sequel. In particular, we will consider the vector Gaussian channel model given by:

$$Y=\Phi X+N \tag{5}$$

where $X\in\mathbb{R}^n$ 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 $N\sim\mathcal{N}\left(0,I\right)\in\mathbb{R}^m$ represents white Gaussian noise.

It has been established that the gradient of mutual information between the input and the output of the vector Gaussian channel model in (5) with respect to the channel matrix obeys the simple relationship [2]:

$$\nabla_\Phi I(X;Y)=\Phi E, \tag{6}$$

where

$$E=\mathbb{E}\left[(X-\mathbb{E}(X|Y))(X-\mathbb{E}(X|Y))^T\right] \tag{7}$$

denotes the MMSE matrix.

III. Gradient of Mutual Information for Vector Poisson Channels

We now introduce the gradient of mutual information with respect to the scaling matrix and with respect to the dark current for vector Poisson channel models. In particular, we assume that the regularity conditions necessary to interchange freely the order of integration and differentiation hold in the sequel, i.e., order of the differential operators $\frac{\partial}{\partial\Phi_{ij}}$, $\frac{\partial}{\partial\lambda_i}$ and the expectation operator $\mathbb{E}(\cdot)$. [2]

Theorem 1. *Consider the vector Poisson channel model in* (1)*. Then, the gradient of mutual information between the input and output of the channel with respect to the scaling matrix is given by:*

$$\begin{aligned}[\nabla_\Phi I(X;Y)_{ij}]=&\big[\mathbb{E}\left[X_j\log((\Phi X)_i+\lambda_i)\right]\\&-\mathbb{E}\left[\mathbb{E}[X_j|Y]\log\mathbb{E}[(\Phi X)_i+\lambda_i|Y]\right]\big],\end{aligned} \tag{8}$$

and with respect to the dark current is given by:

$$\begin{aligned}[\nabla_\lambda I(X;Y)_i]=&\big[\mathbb{E}[\log((\Phi X)_i+\lambda_i)]\\&-\mathbb{E}[\log\mathbb{E}[(\Phi X)_i+\lambda_i|Y]]\big].\end{aligned} \tag{9}$$

irrespective of the input distribution provided that the regularity conditions hold.

It is clear that Theorem 1 represents a multi-dimensional generalization of Theorems 1 and 2 in [3]. The scalar result follows immediately from the vector counterpart by taking $m=n=1$.

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

$$\begin{aligned}\frac{\partial}{\partial\phi}I(X;Y)=&\mathbb{E}\left[X\log((\phi X)+\lambda)\right]\\&-\mathbb{E}\left[\mathbb{E}[X|Y]\log\mathbb{E}[\phi X+\lambda|Y]\right],\end{aligned} \tag{10}$$

and with respect to the dark current is given by:

$$\begin{aligned}\frac{\partial}{\partial\lambda}I(X;Y)=&\mathbb{E}[\log(\phi X+\lambda)]\\&-\mathbb{E}[\log\mathbb{E}[\phi X+\lambda|Y]].\end{aligned} \tag{11}$$

irrespective of the input distribution provided that the regularity conditions hold.

[1] 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.

[2] We consider for convenience natural logarithms throughout the paper.

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 (6) and (7) (see also [2]): 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.

IV. Generalized Bregman Divergences: Definitions and Properties

The classical Bregman divergence was originally constructed to determine common points of convex sets [13]. It has been discovered later the Bregman divergence induces numerous well-known metrics and has a bijection to the exponential family [14].

Definition 1 (Classical Bregman Divergence [13])**.** *Let $F : \Omega \to \mathbb{R}_+$ be a continuously-differentiable real-valued and strictly convex function defined on a closed convex set Ω. The Bregman divergence between $x, y \in \Omega$ is defined as follows:*

$$D_F(x, y) := F(x) - F(y) - \langle \nabla F(y), x - y \rangle . \tag{12}$$

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 [14].

There exist several generalizations of the classical Bregman divergence, including the extension to functional spaces [15] and the sub-modular extension [16]. 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 (7).

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 start by reviewing several notions that are useful for the definition of the generalized Bregman divergence.

Definition 2 (Generalized Inequality [17])**.** *Let $F : \Omega \to \mathbb{R}^{m \times n}$ be a continuously-differentiable function, where $\Omega \in \mathbb{R}^l$ 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:*

$$x \preceq_K y \iff y - x \in K, \tag{13}$$

$$x \prec_K y \iff y - x \in int(K), \tag{14}$$

where $int(\cdot)$ 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:

$$F(\theta x + (1-\theta)y) \preceq_K \theta F(x) + (1-\theta)F(y) \tag{15}$$

for $\theta \in [0, 1]$.

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

$$F(\theta x + (1-\theta)y) \prec_K \theta F(x) + (1-\theta)F(y) \tag{16}$$

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

Definition 3 (Fréchet Derivative [18])**.** *Let V and Z be Banach spaces with norms $\|\cdot\|_V$ and $\|\cdot\|_Z$, respectively, and $U \subset V$ 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*

$$\lim_{\|h\|_V \to 0} \frac{\|F(x+h) - F(x) - DF(x)(h)\|_Z}{\|h\|_V} = 0. \tag{17}$$

$DF(x)$ 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.

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

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

$$D_F(x, y) := F(x) - F(y) - DF(y)(x - y), \tag{18}$$

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

This notion of a generalized Bregman divergence 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 coordinators are non-negative), we have the entry-wise convexity extension. 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 [15].

The generalized Bregman divergence 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 [11], [12].

Theorem 2. *Let $K \subset \mathbb{R}^{m \times n}$ be a proper cone and Ω be a convex subset in a Banach space W. $F, G : \Omega \to \mathbb{R}^{m \times n}$ are Fréchet-differentiable strictly K-convex functions. Then the generalized Bregman divergence D_F associated with the function F exhibits the properties:*

1) $D_F(x, y) \succeq_K \mathbf{0}$.
2) $D_{c_1 F + c_2 G}(x, y) = c_1 D_F(x, y) + c_2 D_G(x, y)$ *for constants c_1, $c_2 > 0$.*
3) $D_F(\cdot, y)$ *is K-convex for any $y \in \Omega$.*

The generalized Bregman divergence also exhibits a duality property similar to the duality property of the classical Bregman divergence, that may be useful for many optimization problems [12], [19].

Theorem 3. *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 space of first quadrant $\mathbb{R}_+^{m\times n}$ (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:*

$$D_F(x, y) = D_{F^\star}(y^\star, x^\star). \tag{19}$$

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 [12], [20], leverage this idea.

The generalized Bregman divergence 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 then the optimal error relates to the conditional mean estimator [21]. Similarly, it can also be shown that for a metric that can be expressed in terms of a generalized Bregman divergence the optimal error also relates to the conditional mean estimator. However, this generalization from the scalar to the vector case requires the partial order interpretation of the minimization.

Theorem 4. *Consider a probability space $(\mathcal{S}, s, \mu)$. 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 $\mathbb{E}[\|X\|] < \infty$ and $\mathbb{E}[\|F(X)\|] < \infty$. Let $s_1 \subset s$ be a sub σ-algebra. Then, for any s_1-measurable random variable y, we have that:*

$$\arg\min_y \mathbb{E}\left[D_F(X, Y)\right] = \mathbb{E}\left[X|s_1\right], \tag{20}$$

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

V. Gradient of Mutual Information: A Generalized Bregman Divergences Perspective

We now re-visit the gradient of mutual information for vector Poisson channel models and for vector Gaussian channel models with respect to the scaling/channel matrix, under the light of the generalized Bregman divergence.

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 generalized Bregman divergences pave the way to the unification of the various Theorems. In particular, we offer two Theorems that reveal that the gradient of mutual information for vector Poisson and vector Gaussian channels admit a representation that involves the average of the generalized Bregman divergence between the channel input X and the conditional mean estimate of the channel input $\mathbb{E}[X|Y]$ under appropriate choices of the vector-valued loss functions.

Theorem 5. *The gradient of mutual information with respect to the scaling matrix for the vector Poisson channel model in* (1) *can be represented as follows:*

$$\nabla_\Phi I(X; Y) = \mathbb{E}\left[D_F(X, \mathbb{E}[X|Y])\right], \tag{21}$$

where $D_F(\cdot, \cdot)$ is a generalized Bregman divergence associated with the function

$$F(x) = x(\log(\Phi x + \lambda))^T - [x, \ldots, x] + [\mathbf{1}, \ldots, \mathbf{1}]^T, \tag{22}$$

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

Theorem 6. *The gradient of mutual information with respect to the channel matrix for the vector Gaussian channel model in* (5) *can be represented as follows:*

$$\nabla_\Phi I(X; Y) = \mathbb{E}\left[D_F(X, \mathbb{E}[X|Y])\right], \tag{23}$$

where $D_F(\cdot, \cdot)$ is a generalized Bregman divergence associated with the function

$$F(x) = \Phi x x^T. \tag{24}$$

Atar and Weissman [10] 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 Theorems 5 and 6, respectively, in view of the fact that the classical Bregman divergence is a specialization of the generalized one.

Corollary 2. *The derivative of mutual information with respect to the scaling factor for the scalar Poisson channel model is given by:*

$$\frac{\partial}{\partial \phi} I(X; Y) = \mathbb{E}\left[D_F(X, \mathbb{E}[X|Y])\right], \tag{25}$$

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

Proof. By Theorem 5, we have $F(x) = x\log(\phi x) - x + 1$. It is straightforward to verify that $F(x)$ induces the scalar gradient result. □

Corollary 3. *The derivative of mutual information with respect to the scaling factor for the scalar Gaussian channel model is given by:*

$$\frac{\partial}{\partial \phi} I(X; Y) = \mathbb{E}\left[D_F(X, \mathbb{E}[X|Y])\right], \tag{26}$$

where $F(x) = \phi x^2$.

Proof. By Theorem 6, $F(x) = \phi x^2$. (26) follows from a simple calculation and the result from [2] that $\frac{\partial}{\partial \phi} I(X; Y) = \phi\mathbb{E}[(X - \mathbb{E}(X|Y))^2]$ □

A. Algorithmic Advantages

Theorem 5 and 6 suggest a deep connection between the gradient and the generalized Bregman divergence. Besides, if a gradient is given in terms of a generalized Bregman divergence, it is possible to simplify optimization algorithms based on gradient-descent. Rather than calculating the gradient itself, one may work directly on its dual form provided that it is easier to calculate the dual function. This idea is behind the essence of the mirror descent methods which have been shown to be very computationally efficient [12], [20].

VI. Applications: Document Classification

The practical relevance of the vector Poisson channel model relates to its numerous applications in various domains. We now briefly shed some light on how our results link to one emerging application that involves classification of documents

Let the random vector $X \in \mathbb{R}_+^n$ model the Poisson rates of n count measurements, e.g. the Poisson rates of the counts of words in a documents for a vocabulary/dictionary of n words.

It turns out that – in view of its compressive nature – it may be preferable to use the model $Y \sim \text{Pois}(\Phi X)$, where $\Phi \in \{0,1\}^{m\times n}$ with $m \ll n$, rather than the conventional model $Y \sim \text{Pois}(X)$ [9], as the basis for document classification. In particular, each row of Φ defines a *set* of words (those with row elements equal to one) that characterize a certain topic. The corresponding count relates to the number of times words in that set are manifested in a document.

The problem then relates to the determination of the "most informative" set of topics, i.e. the matrix Φ. The availability of the gradient of mutual information with respect to the scaling matrix, which has been unveiled in this work, then offers a means to tackle this problem via gradient descent methods.

VII. Conclusion

The focus has been on the generalization of connections between information-theoretic and estimation-theoretic quantities from the scalar to the vector Poisson channel model. In particular, in doing so, we have revealed that the connection between the gradient of mutual information with respect to key system parameters and conditional mean estimation is an overarching theme that transverses not only the scalar but also the vector counterparts of the Gaussian and Poisson channel.

By constructing a generalized version of the classical Bregman divergence, we have also established further intimate links between the gradient of mutual information in vector Poisson channel models and the gradient of mutual information in vector Gaussian channels. This generalized notion, which aims to extend the range of the conventional Bregman divergence from scalar to vector domains, has been shown to exhibit various properties akin to the properties of the classical notion, including non-negativity, linearity, convexity and duality.

By revealing the gradient of mutual information with respect to key system parameters of the vector Poisson model, including the scaling matrix and the dark current, it will be possible to use gradient-descent methods to address several problems, including generalizations of compressive-sensing projection designs from the Gaussian [22] to the Poisson model, that are known to be relevant in emerging applications (e.g. in X-ray and document classification).

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.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.

[3] 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.

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

[5] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] 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.

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

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

[19] 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.

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

[21] 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.

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